A coherent computational Approach to model the bottom-up visual attention.

Olivier Le Meur, Patrick Le Callet, Dominique Barba, Dominique Thoreau

A Coherent Computational Approach to Model Bottom-Up Visual Attention

Olivier Le Meur, Patrick Le Callet, *Member*, *IEEE*,
Dominique Barba, *Senior Member*, *IEEE*, and Dominique Thoreau

Abstract—Visual attention is a mechanism which filters out redundant visual information and detects the most relevant parts of our visual field. Automatic determination of the most visually relevant areas would be useful in many applications such as image and video coding, watermarking, video browsing, and quality assessment. Many research groups are currently investigating computational modeling of the visual attention system. The first published computational models have been based on some basic and well-understood Human Visual System (HVS) properties. These models feature a single perceptual layer that simulates only one aspect of the visual system. More recent models integrate complex features of the HVS and simulate hierarchical perceptual representation of the visual input. The bottom-up mechanism is the most occurring feature found in modern models. This mechanism refers to involuntary attention (i.e., salient spatial visual features that effortlessly or involuntary attract our attention). This paper presents a coherent computational approach to the modeling of the bottom-up visual attention. This model is mainly based on the current understanding of the HVS behavior. Contrast sensitivity functions, perceptual decomposition, visual masking, and center-surround interactions are some of the features implemented in this model. The performances of this algorithm are assessed by using natural images and experimental measurements from an eye-tracking system. Two adequate well-known metrics (correlation coefficient and Kullbacl-Leibler divergence) are used to validate this model. A further metric is also defined. The results from this model are finally compared to those from a reference bottom-up model.

Index Terms—Computationally modeled human vision, bottom-up visual attention, coherent modeling, eye tracking experiments.

1 Introduction

Visual attention is one of the most important features of the human visual system. Rather than speaking about the usefulness of visual attention, which seems obvious, it is worth lingering about its description. The first trial dates back to 1890 when James [1] suggested that *everyone knows what attentions is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought*. In others words, visual attention serves as a mediating mechanism involving competition between different aspects of the visual scene and selecting the most relevant areas to the detriment of others.

Nevertheless, our environment presents far more perceptual information than can be effectively processed. In order to keep the essential visual information, humans have developed a particular strategy, first outlined by James. This strategy, confirmed during the last two decades, involves two mechanisms. The first refers to the sensory attention driven by environmental events, commonly called bottom-up or stimulus-driven. The second one is the voluntational attention to both external and internal stimuli, commonly called top-down or goal-driven.

- *O. Le Meur and D. Thoreau are with the Video Compression Laboratory, Thomson, 1 avenue Belle Fontaine-CS 17616, 35576 Cesson-Sévigné Cedex, France. E-mail: {olivier.le-meur, dominique.thoreau}thomson.net.*
- *P. Le Callet and D. Barba are with the Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN) Laboratory, Ecole Polytechnique de l'Université de Nantes, Rue Christian Pauc-BP 50609, 44306 Nantes Cedex 3, France. E-mail: {patrick.lecallet, dominique.barba}@polytech.univ-nantes.fr.*

Manuscript received 20 July 2004; revised 24 Aug. 2005; accepted 12 Sept. 2005; published online 13 Mar. 2006.
Recommended for acceptance by M. Srinivasan.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0367-0704.

Most recent computational models of visual attention can be placed in two categories. A recent trend concerns a statistical signal-based approach [2] which consists of automatically predicting salient regions of the visual scene by directly using image statistics at the point of gaze. In fact, several studies have recently reported [3], [4], [5] that the human fixation regions present higher spatial contrast and spatial entropy than random fixation regions. These studies show that human eyes movements are not necessarily random but rather driven by particular features. The second category consists of models [6], [7], [8], [9], [10], [11] built around two important concepts: the Feature Integration Theory (FIT) from Treisman and Gelade [12] and a neurally plausible architecture proposed by Koch and Ullman [13]. The FIT suggests that visual information is analyzed in parallel from different maps. These maps are retinotopically organized according to locations in our visual field. There is a map for each early visual feature. From this theory, several frameworks for simulating human visual attention have been designed. The most interesting one has been proposed by Koch and Ullman [13]. Their framework is based on the concept of saliency map which is a two-dimensional topographic representation of conspicuity for every pixels in the image. Fig. 1 illustrates the general synoptic of their model. It mainly consists of early visual features extraction, feature maps building, and feature map fusion.

In this paper, a new bottom-up model based on the FIT and the plausible architecture proposed by Koch and Ullman [13] is described. Its purpose is to automatically detect the most relevant parts of a color picture displayed on a television screen. The general philosophy of this approach is to design a biologically-inspired algorithm that performs better than

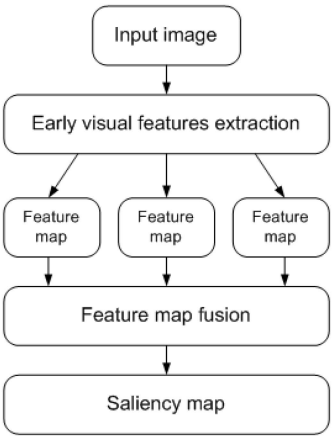

Fig. 1. Framework proposed by Koch and Ullman. Early visual features are extracted from the visual input into several separate parallel channels. After this extraction and a particular treatment, a feature map is obtained for each channel. Next, the saliency map is built by fusing all these maps.

conventional approaches. The proposed model is based on a coherent psychovisual space from which a saliency map is deduced. This space, well justified with psychophysic experiments, is used to combine the visual features (intensity, color, orientation, spatial frequencies...) of the image, that are normalized to their individual visibility threshold. Accurate nonlinear models simulating visual cells behaviors are used to calculate the visibility threshold associated to each value of each component. From this coherent psychovisual space, a new way of calculating a saliency map is proposed.

The paper is organized as follows: Section 2 gives insight into the natural mechanisms that allow us to reduce the amount of visual information. Experiments are conducted to record and track real observer's eye movements with an eye tracking apparatus. These experiments aim to build the ground truth required to achieve a performance assessment of the bottom-up model described here. This experiment is presented in Section 3. The proposed coherent computational approach to model the bottom-up visual attention is described in Section 4. In Section 5, the performances of this model are evaluated, both qualitatively and quantitatively, using relevant metrics. A particular saliency-based application is then briefly described. Finally, the results are summarized and some conclusions are drawn in Section 6.

2 The Natural Selection of the Visual Information

2.1 A Passive Selection

HVS acts as a passive selector, acknowledging some stimuli but rejecting others. The first information reduction appears in the retina in which the photoreceptors only process the wavelengths of the visible light. The neural signal is then treated by ganglion cells which are insensitive to uniform illumination. This particular property is due to the spatial organization of their receptive fields (RF). This fundamental notion was first emphasized in the work of Hartline [33]: The RF is defined as a particular region of the retina within which an appropriate stimulation gives a relevant response. The RF presents an antagonistic center-surround organization. The center is roughly circular surrounded by an annulus. These two regions provide an opposite response for the same stimulation. This center-surround organization is responsible for our great sensibility to the contrast and to the spatial frequency leading to the definition of Contrast Sensitivity Function (CSF).

The responses stemming from the retina neurons are then transmitted to the primary visual cortex. Hubel and Wiesel who received the Nobel prize for medicine and physiology in 1981 discovered that the RF's structure of the cortical cells is considerably different to the structure of the RF of retinal and lateral geniculate nucleus (LGN) cells. The RFs of retinal and LGN cells have a circular structure with a center-surround organization whereas the cortical cells present an elongated RF and respond best to a particular orientation and to a particular spatial frequency. In addition, recent studies [15], [16], [17], [18], [19], [20] have shown that the cortical cell's response can be influenced by stimuli outside their classical RF. These contextual influences are mediated by long-range connections linking cells with nonoverlapping receptive fields. Studies by Kapadia et al. [19], [20] show that the cell's response can be greatly enhanced by the presentation of coaligned, coorientated stimuli in the neighborhood and increases with the number of appropriate stimuli placed outside the CRF. Generally speaking, the contour, feature linking [21], [23], [43], and texture segmentation [22] are assumed to be in close relation with the long-range connections.

2.2 An Active Selection

Human beings have a collection of passive mechanisms lessening the amount of incoming visual information. For instance, the signal stemming from the photoreceptors is assumed to be compressed by a factor of about 130:1, before it is transmitted to the visual cortex. Nevertheless, the visual system is still faced with too much information. To deal with the still overwhelming amount of input, an active selection, involving eye movement, is required to allocate processing resources to some parts of our visual field. Oculomotor mechanisms involve different types of eye movements. A saccade is a rapid eye movement allowing jump from one location to another. The purpose of this type of eye movement, occurring up to three times per second, is to direct a small part of our visual field into the fovea in order to achieve a closer inspection. This last step corresponds to a fixation.

Saccades are therefore a major instrument of the selective visual attention. This active selection is assumed to be controlled by two major mechanisms called bottom-up and top-down control. The former, the bottom-up attentional selection, is linked to involuntary attention. This mechanism is fast, involuntary, and stimulus-driven. Our attention is effortlessly drawn to salient parts in our visual field. These salient parts consist of an abrupt onsets [25] or a local singularity [12]. An image containing one green circle (called target) located among a number of red circles (distractors) is a classic example. The target is easily seen against the red circles due to its local singularity (its local hue), no matter how many distractors are present. The appearance of new perceptual object consistent or not with the context of the scene could also attract our attention [24], [26]. Several studies have shown that observers tend to make longer and more frequent fixations on such object [24].

The second control, top-down attentional selection, refers to voluntary attention closely linked to the experience

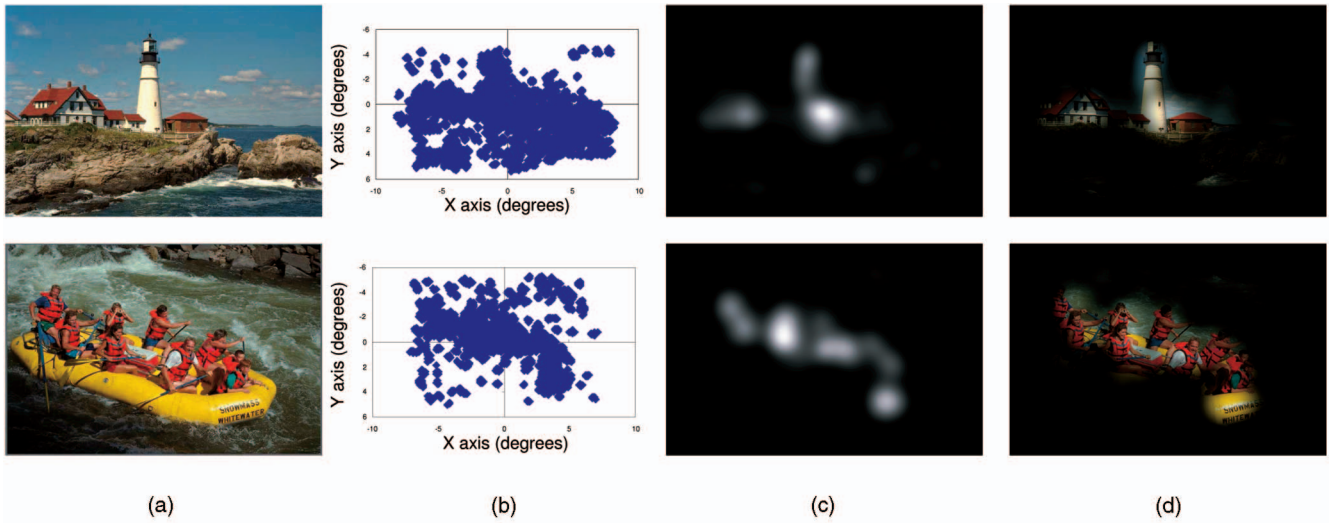

Fig. 2. (a) The original picture, (b) the spatial distribution of human fixations for 14s of viewing time, (c) fixation density map obtained by convolved the spatial distribution with a 2D Gaussian filter, and (d) highlighted human RoI (Regions of Interest) obtained by redrawing the original picture by leaving in the darkness in the nonfixated areas.

of the observers and to the task they have in mind. Compared to the bottom-up attentional selection, the top-down mechanism, voluntary and task-driven, is slower.

3 Eye Tracking Experiments

3.1 Apparatus and Procedure

In order to track and record real observers eye movements, experiments have been conducted using an eye tracker from Cambridge Research Corporation. This apparatus is mounted on a rigid headrest for greater measurement accuracy (less than 0.5 degree on the fixation point). Experiments were conducted in normalized conditions (ITU-R BT 500-10) at a viewing distance of four times the TV monitor height. Ten natural color images with various contents have been selected. The quality of these pictures was then degraded using different techniques (spatial filtering, JPEG, JPEG200 coding...). Forty-six pictures were finally obtained. Every image was seen in random order by up to 40 observers for 15 seconds each in a task-free viewing mode. The collected data corresponds to the regular time sampling (20 ms) of eye gaze on the monitor.

3.2 Human Fixation Density Map Computation

A fixation map, which encodes the conspicuous locations, is computed from the collected data. For a particular picture and for each observer, the samples corresponding to saccades are filtered out. A data point is removed if the number of data included in a squared window is below a given threshold. The size of the window and the threshold are functions of the viewing distance, the accuracy of the eye tracker (0.25 degrees of visual angle) and the resolution of the display (800×600 pixels). In practice, the size of the window and the threshold are, respectively, 9×9 (corresponding to 0.25 degrees of visual angle) and 5 (corresponding to the number of data required in the previous defined window).

All fixation patterns for a given picture are added together providing a spatial distribution of human fixation (see examples in Fig. 2). The resulting map is then smoothed using a two-dimensional Gaussian filter. Its standard deviation is determined according to the accuracy of the eye-tracking apparatus. The result is a fixation density map [34] which represents the observer's regions of interest (RoI). This is often compared to a landscape map [35] consisting of peaks and valleys (see examples in Fig. 2).

3.3 Conclusions from Empirical Data

3.3.1 Coverage

Coverage has been previously defined by Wooding [34] in the following terms: *the coverage is a measure of the amount of the original stimulus covered by the fixations*. The coverage value is therefore given by the ratio between the number of fixated pixels and the number of inspected pixels. A threshold, called T, is required in order to decide whether a pixel is fixated or not.

The coverage value is assessed on the human fixation density maps for three threshold values $0.25, 0.5, 0.75$ and for three viewing times ($2s$, $8s$, and $14s$). Table 1 gives the results for three pictures (*Kayak* (see Fig. 3), *Rapids* (second row of the Fig. 2), and *ChurchandCapitol*).

As expected, the coverage value increases with increasing viewing time and decreasing threshold T. Moreover, the coverage value is highly dependent on the picture content: for picture *Kayak*, the coverage is equal to 21 percent for a viewing time of 14s and for a threshold of 0.75 whereas, in the same condition, the coverage is about 40 percent for picture *ChurchandCapitol*. It is worth noticing that only a small area of the pictures (on average 36 percent for the three thresholds T for 14s of viewing time) has been fixated. In fact, humans pursue to fixate areas of interest rather than to scan the whole scene.

3.3.2 Bias toward the Central Part of Pictures

Fig. 2 shows the spatial distribution and the density of human fixations. These results are coherent with a well-known property of the human visual strategy. Observers have a general tendency to stare at the central locations of the screen. This tendency is not reduced with the viewing time: It can be

TABLE 1
Coverage Evolution in Function of Viewing Time and of Picture Content

Viewing time	t=2s	t=8s	t=14s
coverage (%), $T = 0.25$			
Kayak	7.4	32.7	46.3
Rapids	13.3	30.8	39.6
ChurchandCapitol	17.8	46	57.8
coverage (%), $T = 0.5$			
Kayak	5.7	20.1	30.2
Rapids	9.2	20.9	28.7
ChurchandCapitol	11.3	35.6	46.7
coverage (%), $T = 0.75$			
Kayak	4.6	13	21.8
Rapids	6.6	17.9	23.1
ChurchandCapitol	7.2	29.7	40.2

shown that observers continue to focus on these areas rather than to scan the whole picture. There are at least two plausible explanations: The nonuniform distribution of photoreceptors is a biological candidate. However, it seems more logical to tackle this question by introducing a top-down or higher-level explanation as proposed by Parkhurst et al. [14]. The great majority of visually important information is traditionally located in the central part of the picture frame. Consequently, observers unconsciously tend to select central locations in order to catch the potentially most important visual information.

4 The Proposed Computational Model

The model proposed in this paper is based on the architecture of Koch and Ullman. The model designed by Itti et al. [7] was one of the first to take advantage of such architecture. It has been chosen as a benchmark for the model presented here and is therefore briefly described hereafter.

The first step of Itti et al.'s model consists of the extraction of early visual features. The visual input is broken down into three separate feature channels (color, intensity, and orientation). Each channel is obtained from Gaussian pyramids as in [32]. This allows the computation of different spatial scales by progressively applying a low-pass filter and subsampling the visual features. In order to take into account the organization of the visual cells, a center-surround mechanism based on a Difference of Gaussian (DoG) is applied on each scale. The resulting maps are then linearly summed across feature channels to form the saliency map.

Although this model provides good results on several types of picture, it contains arbitrary steps that are difficult to justify with respect to the HVS:

- several normalization steps are applied before and after the fusion step,
- each channel is normalized independently to a common scale in order to be independent of the feature extraction mechanisms, and
- there are strong links between the visual sensitivity and the viewing distance. However, this has been overlooked.

The proposed computational bottom-up model has been developed bearing numerous properties of human visual cells in mind. Three aspects of the vision process are sequentially tackled, namely, the visibility, the perception, and the perceptual grouping. The complete synoptic is shown in Fig. 3 and described in the following sections.

4.1 Visibility Process

The visibility process simulates the limited sensitivity of the HVS. Despite the seemingly complex mechanisms underlying the human vision, the visual system is not able to perceive all information present in the visual field with the same accuracy. A coherent normalization is first used to scale all the visual data. A value of 1 represents a feature which is just noticeable. All the normalized data is grouped into a psychovisual space. This space is built from the following set of basic mechanisms entirely identified and validated from psychophysic experiments.

4.1.1 *Transformation of the RGB Luminance into the Krauskopf's Color Space*

There are two different types of photoreceptors in the retina: cones and rods. As TV displays luminance levels not corresponding to scotopic conditions (low light levels), rods can be neglected. Cones form the basis of color perception and work at photopic conditions. Cones are of three types: L-cones, M-cones, and S-cones which are sensitive to long, medium, and short wavelengths, respectively. They are mainly located in the central part of the retina, called fovea, which is 2 degrees in diameter. Both psychological and physiological experiments give evidences to the theory of early transformation in the HVS of the L, M, and S signals issued from cones absorption. This transformation provides an opponent-color space in which the signals are less correlated. The principal components of opponent colors space are black-white (B-W), red-green (R-G), and blue-yellow (B-Y). There is a variety of opponent color spaces which differ in the way they combine the different cone responses. The color space proposed by Krauskopf was validated from psychophysic experiments. These experiments are based on the interaction between a color masking signal and a color stimulus signal in term of differential visibility threshold[1] (DVT) of the stimulus. The color orientations of the masking and stimulus signal, respectively, for which the DVT value is minimum are determined. These experiments have been made with still and time varying stimulus. The color space is given by the relation (1):

$$\begin{pmatrix} A \\ Cr_1 \\ Cr_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ -0.5 & -0.5 & 1 \end{pmatrix} \begin{pmatrix} L \\ M \\ S \end{pmatrix}. \quad (1)$$

1. The differential visibility threshold of a stimulus superimposed to a background (masking signal) is defined as the magnitude required by the stimulus to be just noticeable.

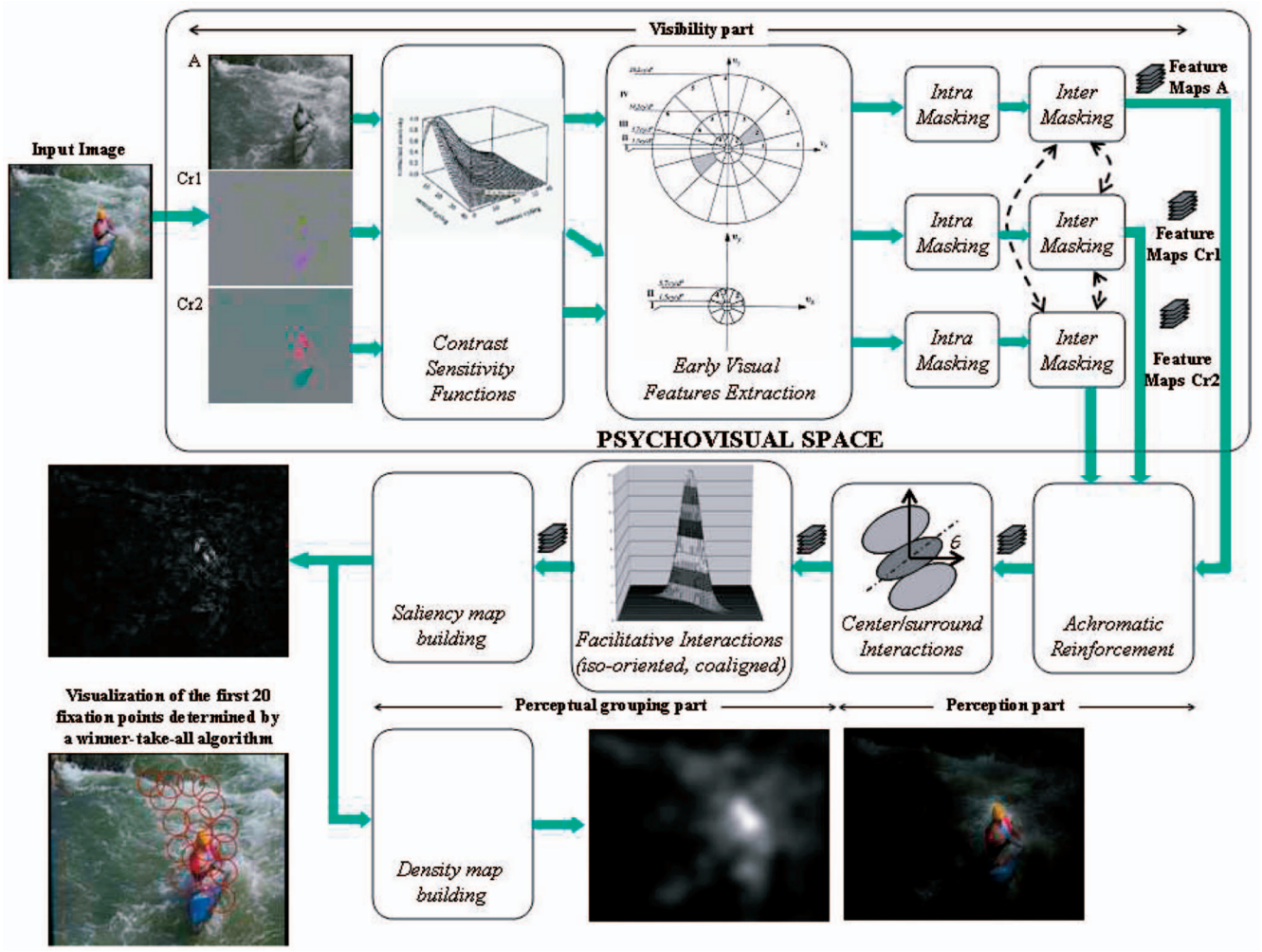

Fig. 3. Flow chart of the proposed computational model of bottom-up visual selective attention. It presents three aspects of the vision: visibility, perception, and perceptual grouping. The visibility part, also called the psychovisual space, simulates the limited sensitivity of the human eyes and takes into account the major properties of the retinal cells. The perception is used to suppress the redundant visual information by simulating the behavior of cortical cells. Finally, the non-CRF and the saliency map building are achieved by the perceptual grouping.

A is a pure achromatic perceptual signal whereas Cr_1 and Cr_2 are pure chromatic perceptual signals.

During these experiments, the adaptation effects through a mechanism of "desensibilization" [16] were taken into account. While Krauskopf used only a temporal "desensibilization" mechanism, a spatial "desensibilization" mechanism was used here. Both methods produced the same result.

4.1.2 Early Visual Features Extraction

It was previously mentioned that visual cells can be characterized by a radial spatial frequency and by orientation. It could therefore be interesting to group visual cells sharing similar properties. The early visual features extraction performed by a perceptual channel decomposition consists of splitting the 2D spatial frequency domain both in spatial radial frequency and in orientation. This decomposition is applied to each of the three perceptual components. Psychophysic experiments [17] show that psychovisual spatial frequency partitioning for the achromatic component leads to 17 psychovisual channels in standard TV viewing conditions while only five channels are obtained for chromatic component (see Fig. 3). Each resulting subband or channel may be regarded as the neural image corresponding to a population of visual cells tuned to a range of spatial frequency and to a particular orientation.

The achromatic subbands are distributed over four crowns noted I, II, III, and IV (see Fig. 3). Chromatic subbands are distributed over two crowns noted I, II. The main properties of these decompositions and the main differences from a similar transform, called the cortex transform [27], are a nondyadic radial selectivity and an orientation selectivity that increases with radial frequency (except for the chromatic components).

4.1.3 Contrast Sensitivity Functions

Contrast sensitivity functions (CSF) have been widely used to measure the visibility of natural images components. In fact, these components can be described by a set of Fourier function and their amplitude. The visibility of a specific component can be assessed by applying a CSF in the frequency domain. When the amplitude of a frequency component is greater than a threshold CT_0, the frequency component is perceptible. This threshold is called the visibility threshold, and its inverse defines the value of the CSF at this spatial frequency. In the

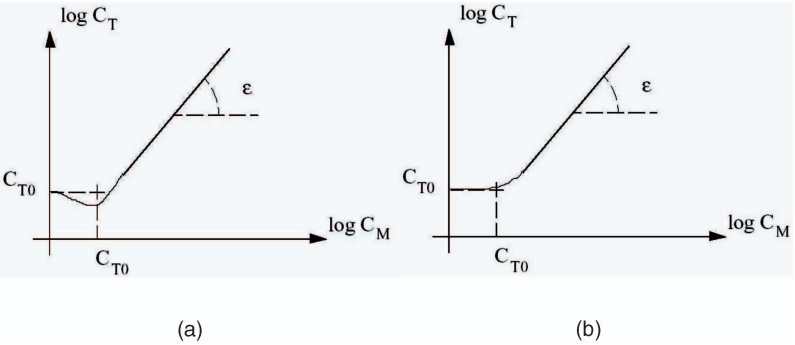

Fig. 4. Nonlinear transducer model of masking effect. When CM varies, three regions can be defined: 1) At low values of CM, the detection threshold remains constant. The visibility of the target is not modified by the masker. 2) When CM tends toward CT_0, the masker eases the detection of the target by decreasing the visibility threshold. This phenomenon is called facilitative or pedestal effect. 3) When CM increases, the target is masked by the masker. Its contrast threshold increases. (a) The target and the masker have similar properties. (b) They have different properties.

approach presented here, CSFs are applied to each components (A, Cr_1, Cr_2). A 2D anisotropic CSF designed by Daly is applied on the achromatic component [31]. The CSFs of the two color visual components Cr_1 and Cr_2 are modeled using sinusoidal color gratings. Two 2D anisotropic functions are obtained [28], [29], [30]. They are two low pass filters with a cut-off frequency of about 5.5 cpd (cycle per degree) and 4.1 cpd for Cr_1 and Cr_2 component, respectively, and are given in relations (2) and (3), respectively. They are functions of the radial pulsation w (expressed in cpd) and the orientation θ (expressed in degrees).

$$S_{Cr_1}(\omega,\theta) = \frac{33}{1+(\frac{w}{5.52})^{1.72}}(1-0.27sin(2\theta)), \quad (2)$$

$$S_{Cr_2}(\omega,\theta) = \frac{5}{1+(\frac{w}{4.12})^{1.64}}(1-0.24sin(2\theta)). \quad (3)$$

These typical CSF show that the human eye is more sensitive to chromatic components with frequencies up to 4-5 cpd. Sensitivity rolls off at both higher and lower frequencies.

4.1.4 Visual Masking

Masking effect refers to the modification of the differential visibility threshold CT_0 of a stimulus due to the influences of the context, called the masking signal [18]. The value CT_0 of the DVT (without any masking effects) is modified into CT by the masking effect. This modification is simply given by the relation $CT = CT_0 \times T$. When $T > 1$, the threshold increases meaning that there is a masking effect. When $0 < T < 1$, the threshold decrease corresponding to a pedestal effect. The visibility of the stimulus is increased. An illustration of the masking effect is shown in Fig. 4: CT and CM are the magnitude of the target in the presence of the masker and the contrast of the masker, respectively. CT_0 is the contrast threshold measured using a CSF without masking effects. Fig. 4a refers to a masker and a target having similar properties (orientation and spatial frequencies). Fig. 4b refers to different cues. When the contrast of the masker varies, three regions can be defined (see Fig. 4):

- At low values of CM, the DVT remains constant. The visibility of the target is not modified by the masker.
- When CM is close to CT_0, the masker eases the detection of the target by decreasing the contrast threshold. This phenomenon is called facilitative or pedestal effect.
- When CM increases, the target is masked by the masker. The contrast threshold increases.

Most of the time, psychophysics experiments based on the detection of simple signals (such as sinusoidal patterns) are used to determine an analytic expression for the visual masking. It is obvious that this is a strong simplification from the intrinsic complexity of natural pictures. Nevertheless, numerous applications (watermarking and video quality assessment) are built around such principles with interesting results. In the context of subband decomposition, three types of masking (intrachannel masking, interchannel masking, and intercomponent masking) can be defined.

Intrachannel masking occurs between signals having the same features (frequency and orientation) and consequently belonging to the same subband. It is the most important masking effect. Intracomponent interchannel masking corresponds to interaction between signals coming from the same component but having different visual features (orientation and spatial frequency). This masking effect is weaker than the intracomponent intrachannel masking. So, it will be neglected here. Finally, intercomponent masking involves two different components, one for the masker, another one for the stimulus.

- Intramasking: The function designed by Daly [31] is used to model the intramasking effect for the achromatic component. The strength of this model comes from its optimization which uses vast amount of experimental results, even if the pedestal effect has been overlooked. The variation of the visibility threshold is given by:

$$T^{intra}_{i,j,A}(x,y) = \left(1+\left(k_1\times\left(k_2\times\left|R^{(0)}_{i,j,A}(x,y)\right|\right)^s\right)^b\right)^{\frac{1}{b}}, \quad (4)$$

where $R^{(0)}_{i,j,c}$ is a subband coming from the perceptual decomposition. (i,j,c) represent, respectively, the spatial frequencies range, the orientation index, and the considered component $(A, Cr1, Cr2)$. The superscript of the letter R (in parenthesis) is used to number each processing step of the model. (x,y) is the considered spatial location, $k_1 = 0.0153$, $k_2 = 392.5$, s, b are constant per subband and given in [31].

The function designed by Le Callet [28] was used to model the intramasking effect for the chromatic components. The analytic form of the model given in relation (5) takes into account the pedestal effect:

$$T^{intra}_{i,j,Cr}(x,y) = \frac{1 + a\|R^{(0)}_{i,j,Cr}(x,y)\| + b\|R^{(0)}_{i,j,Cr}(x,y)\|^2}{1 + c\|R^{(0)}_{i,j,Cr}(x,y)\|}. \quad (5)$$

The parameters a, b, and c are a function of (i,j,Cr). For example, the masking parameters $\{a,b,c\}$ for the channel I of the component Cr_1 (respectively, Cr_2) are equal to $\{0.45, 0.06, 1.22\}$ ($\{0.72, 0.22, 2.78\}$, respectively).

- Interchannel masking: Experimental data clearly shows that there are two different masking effect behaviors. They are functions of the type of the component of the masker and of the type of the stimulus. In some cases, depending on the subbands involved, a facilitation effect lowering the DVT, appears before the masking effect takes place. It corresponds to a first model (Model A). In the other cases, only a pure masking effect is observed. It corresponds to a second model (Model B). The analytic form of theses models:

  Model A:

$$T^{inter}_{i,j,C}(x,y) = \frac{1 + a\|R^{(0)}_{i\prime,j\prime,C\prime}(x,y)\| + b\|R^{(0)}_{i\prime,j\prime,C\prime}(x,y)\|^2}{1 + c\|R_{i\prime,j\prime,C\prime}(x,y)^{(0)}\|}. \quad (6)$$

  Model B:

$$T^{inter}_{i,j,C}(x,y) = a - bexp\Big(-c\|R^{(0)}_{i\prime,j\prime,C\prime}(x,y)\|\Big). \quad (7)$$

  As before, the parameters a, b, and c are function of (i,j,Cr) and depend on the model type (A or B).

The final DVT is given by $CT = CT_0 \times T$, where T is defined in (8) for a particular channel (i,j) and for a particular component C:

$$T_{i,j,C}(x,y) = T^{intra}_{i,j,C}(x,y) \prod_{i\prime} \prod_{j\prime} \prod_{C\prime} T^{inter}_{i\prime,j\prime,C\prime \to i,j,C}(x,y). \quad (8)$$

The term $T^{inter}_{i\prime,j\prime,C\prime \to i,j,C}(x,y)$ expresses a particular interaction produced by a particular location (x,y) of the channel (i',j') of the component C' on the channel (i,j) of the component c. Table 2 gives all the masking interactions that have been integrated in the proposed model. Finally, the modification of the DVT is the product of all the variations of the visibility threshold stemming from both the intrachannel intracomponent masking and from the interchannel intercomponent masking. All the subbands are then weighted by the appropriate modulation value of the DVT:

$$R^{(1)}_{i,j,C}(x,y) = \frac{R^{(0)}_{i,j,C}(x,y)}{T_{i,j,C}(x,y)} \quad (9)$$

with $C = \{A, Cr_1, Cr_2\}$. These mechanisms transform the image into a fully psychovisual space. This space consists of all the visual features normalized to their own differential visibility threshold. It is thus possible to manage visual features stemming from different modalities. For instance, chromatic information could be directly compared, in term of visibility, to achromatic ones.

TABLE 2
Intra and Intermasking Considered in This Bottom-Up Model (a Couple $[Y,Z]$ Means that Channel Y Decreases or Increases the Differential Visibility Threshold of Signal Containing in Channel Z)

	Stimulus		
Masker	A	$Cr1$	$Cr2$
A	intra-masking	$[I,I]$ $[I,IIn]$ $[IIn,I]$ $[IIn,IIn]$	no significant
$Cr1$	$[I,I]$ $[I,IIn]$ $[I,IIIn]$ $[IIn,I]$ $[IIn,IIn]$	intra-masking	$[I,I]$ $[I,IIn]$
$Cr2$	no significant	$[I,I]$ $[I,IIn]$	intra-masking

Example of the bold couple: channels II,n of the $Cr1$ component (crown II with orientation n) can be masked by channel I of the A component.

4.2 Perception

The second part of the model described here deals with perception. The goal is to determine the achromatic components necessary for the calculation of the saliency map. Two mechanisms are involved to detect these components.

4.2.1 Achromatic Reinforcement by Chromatic Context

Color is one of the major visual feature attractors (see [40] for a recent review of early visual features) and can efficiently guide the attention to the most salient areas of our visual field. Any computational models of the visual attention should take advantage of this visual dimension. An original and a plausible way to use the color information is proposed. It consists of increasing the magnitude R of each site of the achromatic channels by accounting for the locally oriented smooth gradient of the chromatic low frequencies. In others words, the saliency of an achromatic structure will be enhanced if this structure is surrounded by a high color contrast (the spatial coordinates (x,y) have been omitted):

$$R^{(2)}_{i,j,A} = R^{(1)}_{i,j,A}(1 + \eta_{Cr_1}\Delta_{Cr_1} + \eta_{Cr_2}\Delta_{Cr_2}), \quad (10)$$

where η_{Cr_1} and η_{Cr_2} control the strength of the contribution of the Cr_1 and Cr_2 component (respectively). These strengths are set to 1 by default.

Δ_{Cr_1} and Δ_{Cr_2} are the locally oriented smooth gradient computed over a small area around the current position. This area is elongated in accordance to the prefered orientation of the subband. Fig. 5 highlights the interest of this process. The relevancy of the first fixation points is greatly enhanced by using the two psychovisual chrominance signals.

4.2.2 Center/Surround Suppressive Interaction

To deal with a large amount of visual information, the visual system uses attentional mechanisms to select relevant areas and to reduce the redundancy of the incoming visual

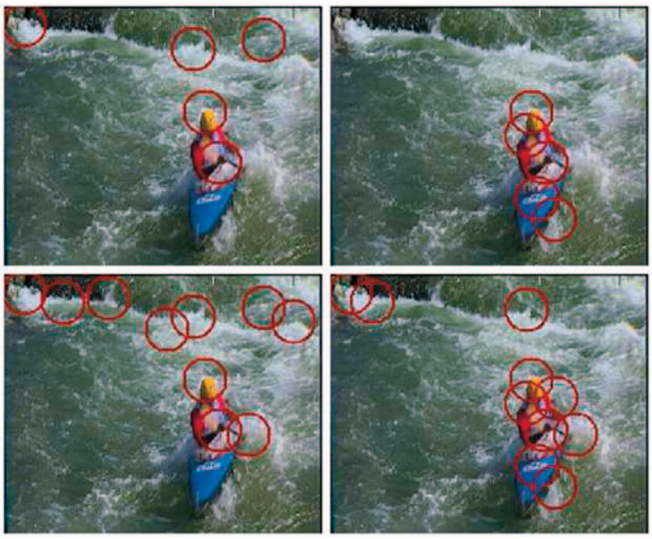

Fig. 5. The first fixation points (five for the first row and 10 for the second) obtained by using only the component A (left) and the three components (A, $Cr1$, $Cr2$) (right) on the picture Kayak. When the number of conspicuous points is decreased, the importance of the color component on the saliency ordering is important. The scan path is more relevant when the color component is used.

information. The former mechanism concerns the bottom-up and top-down behaviors, previously mentioned. The latter mechanism is probably the most straightforward common feature shared by all the visual cells. In order to form an economical representation of the visual world, the particular oriented center/surround organization of the cortical cells is really important. For instance, center/surround organizations imply that visual cells are insensitive to uniform illumination. The responses of such cells are efficiently simulated by a difference-of-Gaussian function.

This mechanism is thus simulated by subtracting an inhibition contribution to the current subband location as proposed in (1). The inhibition contribution is obtained from the convolution of a normalized weighting function called $\omega_{\sigma_x,\sigma_y}$ with the current signal in the subband (i,j).

$$R^{(3)}_{i,j,A} = H\Big(R^{(2)}_{i,j,A} - R^{(2)}_{i,j,A} * \omega_{\sigma_x,\sigma_y}\Big) \tag{11}$$

with

$$\omega_{\sigma_x,\sigma_y}(x,y) = \frac{1}{||H(DoG_{\sigma_x,\sigma_y}(x,y))||_1} H(DoG_{\sigma_x,\sigma_y}(x',y')), \tag{12}$$

$$H(z) = \begin{cases} z & \text{if } z \geq 0 \\ 0 & \text{if } z < 0, \end{cases} \tag{13}$$

and $||.||_1$ denotes the L_1 norm. $(x',y')^T$ is obtained by translating the original coordinate system by $(x_0,y_0)^T$ and rotating it by $\theta_{i,j}$:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} cos\theta_{i,j} & sin\theta_{i,j} \\ -sin\theta_{i,j} & cos\theta_{i,j} \end{pmatrix} \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix}. \tag{14}$$

4.3 Perceptual Grouping

Perceptual grouping refers to the human visual ability to group and bind visual features to organize a meaningful higher-level structure. There are numerous mechanisms involved in the perceptual grouping. One of the most common is the facilitative interactions that have been reported in numerous studies. In most cases, these interactions appear outside the CRF along the preferred orientation axis and are maximal when center and surround stimuli are iso-oriented and coaligned [15], [19] (due to the long-range horizontal connections). In other words, the activity of cells is enhanced when the stimuli within the CRF and a stimuli within the surrounding area are bound to form a contour. This facilitative interaction is usually termed contour enhancement or contour grouping. Most of the recent computational models are based on the Gestalt principles of colinearity and proximity [42], [43]. Fig. 6 gives an illustration of contour grouping due to the long-range grouping interactions. Contour grouping is simulated using two half butterfly filters $B^0_{\theta_{i,j,A}}$ and $B^1_{\theta_{i,j,A}}$. Butterfly filters are obtained by a directional term $D_{i,j}(x,y)$ and a proximity term generated by a circle C_r blurred by a Gaussian filter $G(x,y)$. They are given by:

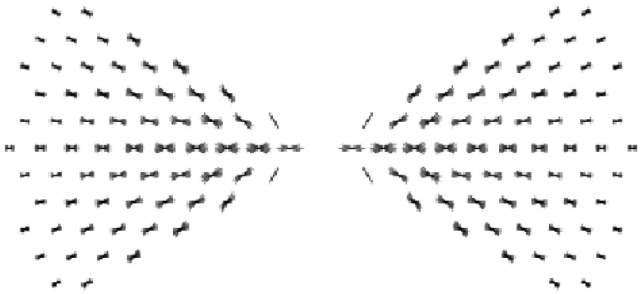

Fig. 6. The long-range grouping interactions for a preferred horizontal orientation are often represented by the above picture [21]. The filtering kernel of a Butterfly filter favorably simulates this spatial distribution.

$$B_{i,j,A}(x,y) = D_{i,j}(x,y) \cdot C_r(x,y) * G(x,y) \tag{15}$$

with

$$D_{i,j}(x,y) = \begin{cases} cos(\frac{\pi/2}{\alpha}\varphi) & \text{if } -\alpha < \varphi < \alpha \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

and $\varphi = arctan(\frac{y'}{x'})$, where $(x',y')^T$ is obtained by (14). The parameter α defines the opening angle of the Butterfly filters. It depends on the angular selectivity of the considered subband.

The Butterfly filter $B_{i,j,A}$ is then decomposed into the half butterfly filters $B^0_{i,j,A}$ and $B^1_{i,j,A}$. For every oriented subbands (i,j) and location (x,y), we compute the facilitative factor:

$$f^{iso}_{i,j,A}(x,y) = \frac{L^1_{i,j}(x,y) + L^0_{i,j}(x,y)}{max\Big(\beta, |L^1_{i,j}(x,y) - L^0_{i,j}(x,y)|\Big)} \tag{17}$$

with β a constant (saturation), $L^0_{i,j}(x,y) = R^{(3)}_{i,j,A}(x,y) * B^0_{i,j,A}(x,y)$, and $L^1_{i,j}(x,y) = R^{(3)}_{i,j,A}(x,y) * B^1_{i,j,A}(x,y)$.

Finally, the subband stemming from this facilitation step, noted $R^{(4)}_{i,j,A}$, is obtained by weighting the subband $R^{(3)}_{i,j,A}$ by a factor $\kappa^{iso}(x,y)$ depending on the ratio of the local maximum of the facilitative factor $f^{iso}_{i,j}(x,y)$ and on the global maximum of this factor computed on all subbands

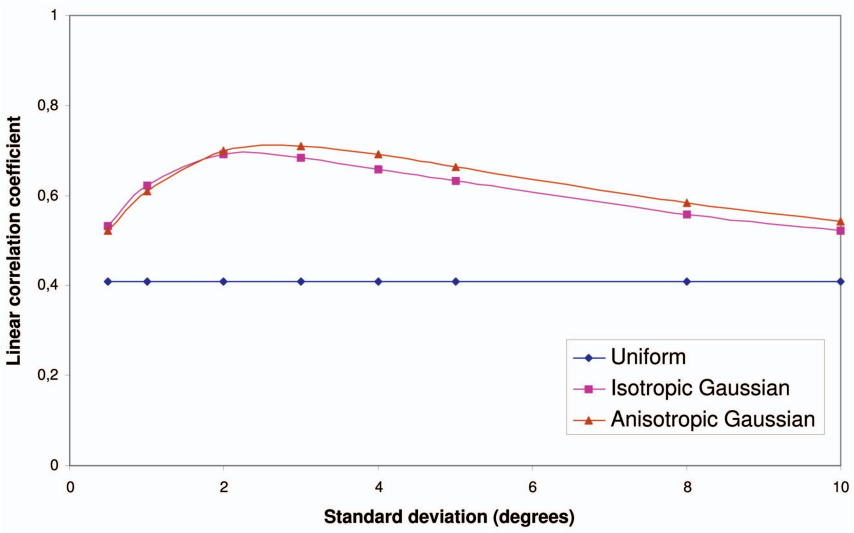

Fig. 7. Linear correlation coefficient as a function of standard deviations of the Gaussian function used to weight the predicted saliency map computed over 18 pictures. The best performances are achieved by anisotropic Gaussian weighting. For $\sigma_x^e = 2.5$ degrees, the correlation coefficient is equal to 0.71

having the same radial spatial frequency range. The resulting subband is thus given by

$$R_{i,j,A}^{(4)}(x,y) = R_{i,j,A}^{(3)}(x,y)\left(1 + \kappa^{iso}(x,y) f_{i,j,A}^{iso}(x,y)\right) \quad (18)$$

with

$$\kappa^{iso}(x,y) = \frac{\max_{(x,y)}\left(f_{i,j,A}^{iso}(x,y)\right)}{\max_j\left(\max_{(x,y)}\left(f_{i,j,A}^{iso}(x,y)\right)\right)}. \quad (19)$$

4.4 Saliency/Density Map Building

A two-dimensional saliency map S is computed by summing directly the output of the different achromatic channels. During eye tracking experiments, participants have to stare at the center of the screen prior to stimulus onset. In order to deal with this constraint, the saliency map can be favorably weighted by a anisotropic Gaussian with standard deviations (σ_x^e, σ_y^e) centered on the location where the participant was fixating at the beginning of the experiments ((x_0, y_0) represent the picture's center coordinates in (20)). The resulting saliency map, called S', is then given by (20):

$$S'(x,y) = S(x,y) exp\left(-\frac{(x-x_0)^2}{2\sigma_x^{e\,2}} - \frac{(y-y_0)^2}{2\sigma_y^{e\,2}}\right). \quad (20)$$

The standard deviation values (σ_x^e, σ_y^e) have been obtained from an optimization routine conducted over 18 pictures. The average correlation coefficient between human fixation density maps and the predictions was maximized during the optimization. The average correlation coefficient evolution is plotted in Fig. 7. Three types of weighting function are considered: uniform weighting, isotropic Gaussian and anisotropic Gaussian. As expected, the uniform weighing function gives the worst results. Gaussian functions improve results with an advantage for the anisotropic Gaussian function. The best results are obtained for a standard deviation related to the x-axis equal to $\sigma_x^e = 2.5$ degrees (correlation coefficient close to 0.71). The standard deviation related to the y-axis σ_y^e is obtained by calculating a ratio depending on the picture's size on the standard deviation σ_x^e:

$$\sigma_y^e = \sigma_x^e \times \left(\frac{R_x}{R_y} Ind(R_x < R_y) + \frac{R_y}{R_x} Ind(R_x > R_y)\right), \quad (21)$$

where R_x and R_y are the picture's size (width and height, respectively) expressed in degree of visual angle and $Ind()$ is the indicatric function. Several remarks can be made regarding the weighting function. First, the weighting function is centered on the screen in order to deal with one of the constraints encountered during the previous eye tracking tests. Second, the context of this study concerns the detection of the most visually important regions of a picture displayed on a TV screen. The center of the screen is then a natural point that attracts the attention. This weighting is not applied to all the fixations points because the goal is to simulate the behavior of an average observer. Moreover, the scan path is idiosyncratic, meaning that the sequence of fixation points is different for each of the participants. These precisions were necessary to unravel the role of the anisotropic Gaussian weighting function. It is not a bottom-up mechanism. It is rather an experimental constraint encountered during the eye tracking experiments.

5 Performance Evaluation

Computational bottom-up model performance can not be readily assessed and there is no real consensus on any assessment method. Nevertheless, several objective methods have already been proposed. They depend on both the type of assessed images (synthetic or natural) and the knowledge of the ground truth coming from eye tracking experiments. In addition, there are two major ways to conduct the objective assessment. The first one consists in comparing the first fixations of the scan paths whereas the second compares two fixation density maps [14], [34], [36], [37], [38], [39].

Fig. 8. From left to right: the original picture, the highlighted human RoI and the highlighted prediction RoI. Predicted and human RoI are highlighted while the nonfixated areas remain in darkness.

In this paper, the performance of the proposed model is evaluated both qualitatively and quantitatively. The former method refers to qualitative assessments involving human appreciations with all underlying drawbacks. The latter deals with objective methods. As the saliency ordering of the fixation points is not of interest here, the objective methods have to yield a unique scalar value representing the effectiveness of the proposed model. Two objective metrics (the linear correlation coefficient and the Kullback-Leibler divergence in the continuity of the work initiated in [41]) have been used. Further analysis is also performed in order to determine the contribution of the visual masking and the achromatic reinforcement by chromatic context.

5.1 Qualitative Evaluation

Qualitative or subjective evaluation provides an insight into the effectiveness of the proposed model. First of all, it is important to point out that the way to display the results can affect judgment. For example, a logarithmic function lessening the high values is applied on the ground truth in order to favor the display. Fig. 8 shows the ground truth and the results for three pictures. The similarity between the predictions and the experimental results is good. The most relevant areas are well detected.

5.2 Quantitative Evaluations

Quantitative or objective evaluations are conducted following two different methods (the linear correlation coefficient and the Kullback-Leibler divergence). A comparison with the model of Itti is also conducted. The degraded pictures used during the psychovisual tests do not produce significant modifications both in the human and in the predicted saliency map. Therefore, only the 10 original color pictures are used to perform the assessments.

5.2.1 *The Linear Correlation Coefficient*

The linear correlation coefficient, noted cc and given by (22), is widely used to compare two images for applications such as image registration, object recognition, and disparity measurement.

The linear correlation coefficient measures the strength of a linear relationship between two variables. It has some interesting advantages. The first one is its capacity to compare two variables by providing a single scalar value. The correlation coefficient has a value between -1 and $+1$. When the correlation is close to $+/-1$, there is an almost perfectly linear relationship between the two variables.

$$cc(p,h) = \frac{cov(p,h)}{\sigma_p \sigma_h}. \tag{22}$$

p and h represent, respectively, the human fixation density map and the predicted fixation density map. $cov(p,h)$ is the covariance value between p and h. σ_p and σ_h are the standard deviation for the human and the predicted density map, respectively.

The proposed approach outperforms the reference model in all tested configurations. The first assessment concerns the computation and the comparison of the correlation

TABLE 3
Correlation Values for Different Pictures

Pictures	Weighting	Proposed model without weighting	Itti's model	Proposed model + weighting	Itti's model + weighting
Kayak	0.43	0.33	0.38	0.60	0.45
Manfishing	0.77	0.46	0.26	0.87	0.79
Churchandcapitol	0.55	0.55	0.35	0.57	0.54
Vautour538	0.68	0.49	0.28	0.75	0.66
Zebres797	0.70	0.10	0.38	0.69	0.70
patin	0.82	0.46	0.29	0.81	0.83
Lighthouse2	0.61	0.60	0.57	0.71	0.64
Rapids	0.53	0.51	0.45	0.58	0.58
Dancers2	0.70	0.66	0.28	0.78	0.77
Sailing1	0.64	0.20	0.45	0.66	0.60
Average	0.64	0.44	0.37	0.70	0.66
t-test		$p > 0.3$		$p > 0.02$	

The degree of similarity is given for different predictions: The weighting function, the proposed model, and the reference model (Itti's model) with/without eccentricity function for a viewing time of $14s$.

TABLE 4
Average Correlation Coefficients and Kullback-Leibler Values for Different Pictures and Different Viewing Times

Pictures	Weighting	Proposed model + weighting	Itti's model + weighting	Absolute gain	Gain in percentage
Correlation coefficients				$cc_{Proposed} - cc_{Itti}$	
$4s$	0.64	0.68	0.65	+0.03	$+4.62\%$
$10s$	0.65	0.70	0.66	+0.04	$+6.06\%$
$14s$	0.64	0.70	0.66	+0.04	$+6.06\%$
KL coefficients				$KL_{Itti} - KL_{Proposed}$	
$4s$	0.78	0.65	0.66	+0.01	-1.52%
$10s$	0.56	0.49	0.56	+0.07	-12.5%
$14s$	0.57	0.47	0.54	+0.07	-12.96%

Several predictions are compared: The weighting function, the proposed model, and Itti's model when the weighting is enabled. The gain yielding by the proposed model is given when Itti's model is taken as a reference.

coefficients (listed in Table 3) for both models. The average correlation coefficient is improved by 0.07 $(0.44 - 0.37)$ when the weighting function is disabled. This means that the proposed mechanisms are of interest. Even though a simple t-test does not reveal a significant improvement (the probability that a significant difference exist between the means of two sets is only about 0.7), the Kullback-Leiblur metric will give the same tendency.

The performances of the two purely bottom-up models are less relevant than a basic Gaussian function centered on the center of the image (called weighting in Tables 3 and 4). To deal with both the experimental constraint encountered during the eye tracking experiments and the natural attraction toward the center of the picture, the predicted saliency maps are weighted by this Gaussian function. When the weighting function is enabled, the improvement is less important than the previous one but more significant regarding the t-test value ($p > 0.02$). The gain is about 0.04 on average (see Table 3) regardless of the viewing time (see Table 4 in which the gain is shown in the last column). As Parkhurst et al. emphasized in [14] (they applied a weighting function on the prediction stemming from Itti's model), the efficiency of the two approaches is improved by the application of a weighting function: The average correlation of the proposed model (Itti's model) is multiplied by a factor 1.59 (respectively, 1.78). The benefit yielded by the weighting function is due to two reasons: The first one concerns the aforementioned constraint of the experiments whereas the second deals with a top-down property.

Nevertheless, the linear correlation computed between the predicted and the human saliency map for the particular picture *Parrots*[2] is equal to 0.59. As the regions of interest of the picture *Parrots* are not centered, the

2. This picture is not used in the different simulations because predictions coming from Itti's model are not available.

TABLE 5
Comparisons between the KL_{avg} Values and the KL Values Stemming from the Proposed Model and from the Reference Model

Pictures	Proposed model	Itti's model	KL_{avg}	Proposed model	Itti's model	KL_{avg}
Viewing time	$14s$			$4s$		
kayak	0.20	0.65	0.61	1.27	1.01	0.75
manfishing	0.40	0.53	0.54	0.38	0.61	0.70
churchandcapitol	0.49	0.62	0.54	0.44	0.60	1.01
vautour538	0.49	0.60	0.51	0.80	0.82	0.74
zebres797	0.94	0.66	0.72	0.40	0.39	0.90
patin	0.26	0.29	0.60	0.35	0.41	0.82
lighthouse2	0.46	0.48	0.58	0.50	0.46	0.84
rapids	0.70	0.83	0.42	1.26	1.27	0.74
dancers2	0.29	0.44	0.49	0.49	0.50	0.70
sailing1	0.50	0.46	0.57	0.66	0.54	0.89
Average	0.47	0.56	0.56	0.65	0.66	0.81
t-test	$p > 0.18$			$p > 0.89$		

Results are given for two viewing times.

weighting function only yields a correlation value of 0.28. It means that the influence of the weighting function has to be interpreted with care.

Compared to predictions obtained by the weighting function solely, the average correlation is increased by 3 percent (relative to Itti's model). The gain of the proposed model is about 6 percent (Table 3).

Finally, Itti's model seems to be more relevant on pictures that contain a relatively small and unique region of interest. For instance, it is more relevant for pictures *Kayak* and *Sailing1*. For pictures *Patin* and *Lighthouse2*, both models yield the same results. This relevancy on such pictures is probably due to the feature normalization operator used by Itti. This iterative operator, as described in [44], discards all feature maps that presents over extended regions of the input image and enhances the feature maps containing isolated salient locations. This operator, directly inspired by physiological and psychological studies of long-range cortico-cortical connections, is likely to be the most interesting element of Itti's model. The behavior of the reference model is therefore better when there are fewer salient locations in the image. The reference model is less relevant when the picture contains numerous salient locations.

5.2.2 The Kullback-Leibler Divergence

The Kullback-Leibler divergence is used to compute the degree of dissimilarity between two probability density functions. Two probability density functions are deduced from the human saliency maps and the predicted saliency maps. The Kullback-Leibler divergence, noted KL, is given by (23):

$$KL(p|h) = \sum_x p(x) Log\left(\frac{p(x)}{h(x)}\right) \quad (23)$$

with h the probability density from human results, and p the predicted probability density function. When the two probability densities are strictly equal, the KL value is zero. The performance evaluation mainly consists in comparing the two approaches when the weighting function is enabled. The proposed model exhibits better performances than the reference model (see Table 4). On average, the gain is greater than 10 percent regardless of the viewing time. As the Kullback-Leibler metric is very sensitive to dissimilarities, these results show that the proposed approach yields less erroneous data than the reference model. These measurements confirm the results obtained with the linear correlation coefficient. Notwithstanding the average performance difference, the reference model outperforms the proposed approach for the picture *Kayak*. For a viewing time of $14s$, the performance difference is about 0.152. The t-test value (see Table 5) shows a difference ($p > 0.18$) for a viewing time of $14s$. For a viewing time of $4s$, there are no significant differences between the two models.

Another way to objectively evaluate the performances of this model consists in computing the average dissimilarity over all the observers. This could be obtained by computing the Kullback-Leibler divergence between the probability density function for one observer and the probability density function obtained for all participants. This computation is iterated over the set of observers. The average of the Kullback-Leibler values, called KL_{avg} is given in (24). The behavior of an average observer can then be identified: a high KL_{avg} value means that the visual strategy of all observers is different. In others words, the dispersion inter-observers is high. A weak value means that the visual strategy of all observers is similar. The minimum value is zero and will be obtained only if all observers stare at the same locations during the same amount of time.

$$KL_{avg} = \frac{1}{N} \sum_i KL(h_i|h) \quad (24)$$

TABLE 6
Contribution of the Visual Masking (VM), the Achromatic Reinforcement (AR) as Regard to the CC Value

Pictures	Proposed model	no VM	no AR	Proposed model	no VM	no AR
Viewing time	$14s$			$4s$		
Kayak	0.33	0.34	0.28	0.29	0.31	0.25
Manfishing	0.46	0.32	0.44	0.38	0.26	0.37
Churchandcapitol2	0.55	0.46	0.54	0.55	0.52	0.54
Vautour538	0.49	0.51	0.48	0.36	0.37	0.36
Patin	0.46	0.46	0.46	0.45	0.40	0.44
Lighthouse2	0.60	0.58	0.57	0.55	0.56	0.52
Rapids	0.51	0.54	0.48	0.48	0.51	0.52
Dancers2	0.66	0.59	0.63	0.55	0.49	0.52
Sailing1	0.20	0.09	0.19	0.12	0.01	0.11
Parrots	0.60	0.64	0.59	0.59	0.64	0.58
Plane	0.61	0.56	0.58	0.49	0.45	0.47
Bike	0.52	0.42	0.45	0.54	0.45	0.48
Average	0.50	0.46	0.47	0.45	0.41	0.43

with h the global probability density function from the data for all the participants, h_i the probability density function for the i observer, N the number of observers.

Table 5 gives the KL_{avg} values and the results coming from the proposed model for different pictures. The temporal evolution of the KL_{avg} value is noticeable. When the viewing time increases, the KL_{avg} value decreases. This means that the visual strategies of the observers are the closest. A possible explanation could refer to a property of human visual strategy previously observed: Rather than scanning the whole scene, humans concentrate on areas of interest. Therefore, when the viewing time increases, the contribution of spatial locations that are visited with a very low frequency decreases.

It is also interesting to compare the divergence value, noted $KL(p|h)$ with the KL_{avg} value. $KL(p|h)$ is computed from the predicted probability density and the global probability density. Three cases can be considered:

- $KL(p|h) \approx KL_{average}$: When the two values are similar, there is a good pairing between the predicted density functions and the set of density functions obtained for each observer. For examples, for a viewing time of $14s$, the pictures *Churchandcapitol* and *Vautour538* fall in this category.
- $KL(p|h) < KL_{average}$: When the value associated to the prediction is smaller than the KL_{avg} value, the most important part of the predicted density is well paired with the set of density functions obtained for each observer. In others words, the most conspicuous areas of the picture are well predicted. The predicted saliency map is almost fully included in the saliency map produced by the observers.
- $KL(p|h) > KL_{average}$: When the value is greater than the KL_{avg} value, there is a weak pairing between the predicted density and the set of density functions obtained for each observer. Differences stem from the spatial locations of the most important areas in the two density functions. There are major dissimilarities between the two sets.

Therefore, in the light of the results summarized in Table 5, the proposed model succeeds in predicting the spatial locations of the most important areas. There is no major dissimilarity in average. It is in accordance with the previous results listed in Table 3.

5.3 Contribution of Major Computational Steps of the Proposed Model

In this section, the contributions of the most important biologically plausible mechanisms proposed in this model are evaluated. The weighting function is disabled. Therefore, the contributions yielded from each step are not biased by a higher level mechanism. Table 6 gives both the contributions of the visual masking (called VM) and the contributions of the achromatic reinforcement (called AR) regarding the linear coefficient correlation. For the two viewing durations, the best performances are obtained by the model including the two aforementioned mechanisms. The evaluation of the perceptual grouping is not assessed because its contribution is only relevant to particular pictures having high contrasted straight lines.

The visual masking contribution is considered first. The visual masking is a bottom-up mechanism and the influences of such mechanism are the strongest just after the stimulus onset, prevailing against the higher-level mechanisms. Once information from the visual input has been acquired through the bottom-up mechanism, top-down influences are exerted, involving goal-oriented mechanisms. It is likely that the gain yielding from the visual masking decreases with the viewing time. Nevertheless, it is noticeable that the gain obtained by this mechanism is the same for the two viewing times. On average, the gain is about 0.04 (the improvement is about 9 percent). These results indicate that the allocation of attention depends on the visual features throughout the

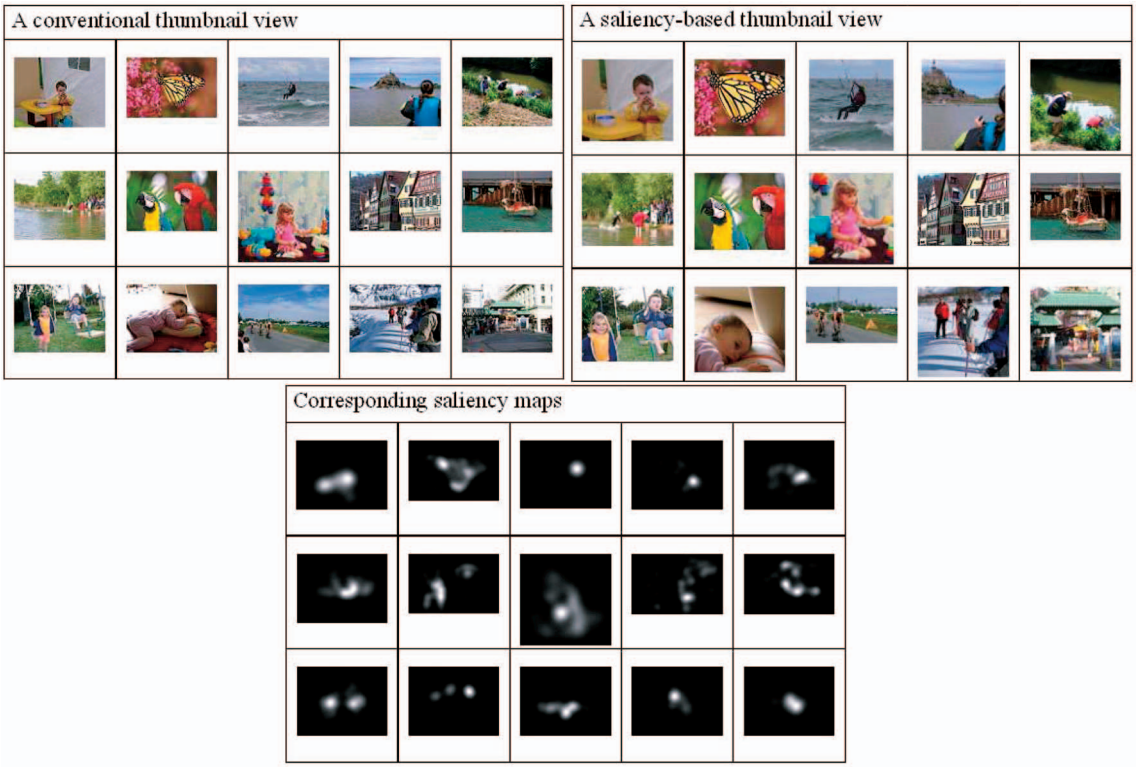

Fig. 9. Example of a saliency-based image browsing. On the left, the results coming from a conventional approach are given. On the right, the results coming from an approach using a saliency map are shown. The last row gives the saliency maps corresponding to the different pictures.

experiments. In other words, eye movements are influenced by stimulus properties not only after its onset but even for a longer viewing time. It is consistent with the work of Parkhurst et al. [14].

The improvement is similar ($+0.03$), as far as the achromatic reinforcement is concerned, for the two proposed viewing times. The gain of the achromatic reinforcement is weak on average. Nevertheless, this plausible biologically mechanism is still interesting for several reasons.

First, the predicted saliency maps for the pictures *Kayak* and *Bikes* are really improved. These two pictures contain color regions of interest which become more conspicuous as the proposed mechanism is applied. In addition, the correlation coefficient is either improved or unmodified for all other pictures. For example, the linear correlation coefficients of the pictures *Vautour538*, *Patin* (this picture is shown in the last row of the Fig. 8) and *Sailing1* are not improved. In fact, these pictures contain little color information. It is clear that the efficiency of the achromatic reinforcement can only be assessed on pictures featuring relevant color information.

The second point concerns the metric used here. The linear correlation coefficient is a global metric computed over the complete picture, meaning that the impact of a local improvement is lessened by the unmodified remaining values. The relevance of this reinforcement is therefore difficult to measure. Nevertheless, as it was previously mentioned, the achromatic reinforcement is interesting, especially on the first fixation points as illustrated in Fig. 5. This reinforcement allows to concentrate the first fixation points on the most interesting parts of a color picture.

5.4 An Example of Saliency-Based Application

A specific application of this model is given. In order to facilitate the image viewing on devices with limited display sizes, the saliency map can be a very useful tool. In the proposed approach, the most important salient parts of the picture are cropped to fit the limited display size. This proposed solution eases the browsing, as illustrated by Fig. 9.

Two thumbnail views are depicted: The first one is obtained by a conventional approach consisting of downsampling the original picture. The second approach is the proposed saliency-based application. The saliency maps, corresponding to the different pictures and obtained by the proposed model, are shown on the last row of Fig. 9.

6 Conclusion

The automatic determination of the most visually relevant areas in a picture is important for many applications, such as image and video browsing, watermarking, image, and video coding and quality assessment.

A coherent computational model of visual selective attention is described in this paper. It aims at building a saliency map for a still color picture, indicating the most relevant spatial locations. The architecture of the proposed model is similar in spirit to the Koch and Ullman architecture. The fundamental difference concerns the normalization of all the early visual features. They are all normalized by their own visibility threshold such that a value of one refers to a just noticeable data. The visibility threshold can be modified by the context, and this is incorporated by the modeling of visual masking. This coherent normalization allows for the expression of all the visual features in term of visibility. The saliency

values are obtained from this psychovisual space. A particular strategy is proposed based on the achromatic structure reinforced by color information.

The proposed model is compared both qualitatively and quantitatively to a reference saliency map. This reference, also called the "ground truth," is built from the data collected by an eye-tracking apparatus. Two well-known metrics, the linear correlation coefficient and the Kullback-Leibler divergence, are used to conduct the qualitative comparison. These coefficients are 0.71 and 0.46, respectively. The proposed model outperforms the model of Itti in all the tested configurations.

This model only considers the determination of the most important achromatic structure. It would be therefore possible to improve its performances by including more combinations of the early visual features. For example, the definition of a chromatic saliency map may enhance the prediction of the most relevant areas of the picture. As all the early visual features have been coherently normalized, it would be straightforward to implement other combination strategies. In addition, psychophysic experiments could be performed in order to establish several parameters that have been arbitrarily set. Finally, the framework presented here is limited to still color images. This model could be further improved by including the time dimension in order to process complex dynamic sequences.

REFERENCES

[1] W. James, *The Principles of Psychology.* New York: Holt, 1890.

[2] U. Rajashekar, L.K. Cormack, and A.C. Bovik, "Point of Gaze Analysis Reveals Visual Search Strategies," *Proc. SPIE Human Vision and Electronic Imaging IX,* 2004.

[3] P. Reinagel and A.M. Zador, "Natural Scene Statistics at the Centre of Gaze," *Network: Computational Neural Systems, 10,* pp. 1-10, 1999.

[4] D.J. Parkhurst and E. Niebur, "Scene Contant Selected by Active Vision," *Spatial Vision,* vol. 16, pp. 125-154, 2003.

[5] M. Mack, M.S. Castelhano, J.M. Henderson, and A. Oliva, "What the Visual System Sees: The Relationship between Fixation Positions and Image Properties During a Search Task in Real-World Scenes," *Proc. Ann. Object, Perception, Attention, and Memory Conf.,* 2003.

[6] J. Tsotsos and S.M. Culhane, "Modeling Visual Attention via Selective Tuning," *Artificial Intelligence 78,* pp. 507-545, 1995.

[7] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[8] L. Itti and C. Koch, "A Comparison of Feature Combination Strategies for Saliency-Based Visual Attention Systems," *Proc. SPIE Human Vision and Electronic Imaging IV,* vol. 3644, pp. 373-382, 1999.

[9] L. Itti and C. Koch, "A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention," *Vision Research,* vol. 40, nos. 10-12, pp. 1489-1506, 2000.

[10] B. Bruce and E. Jernigan, "Evolutionary Design of Context-Free Attentional Operators," *Proc. Int'l Conf. Image Processing '03,* 2003.

[11] R.L. Canosa, "High-Level Aspects of Oculomotor Control During Viewing of Natural-Task Images," *Proc. SPIE Human Vision and Electronic Imaging VIII,* vol. 5007, 2003.

[12] A.M. Treisman and G. Gelade, "A Feature-Integration Theory of Attention," *Cognitive Psychology,* vol. 12, no. 1, pp. 97-136, 1980.

[13] C. Koch and S. Ullman, "Shifts in Selection in Visual Attention: Toward the Underlying Neural Circuitry," *Human Neurobiology,* vol. 4, no. 4, pp. 219-27, 1985.

[14] D. Parkhurst, K. Law, and E. Niebur, "Modeling the Role of Salience in the Allocation of Overt Visual Attention," *Vision Research,* vol. 42, pp. 107-123, 2002.

[15] J.I. Nelson and B.J. Frost, "Intracortical Facilitation among Co-Oriented, Co-Axially Aligned Simple Cells in Cat Striate Cortex," *Experimental Brain Research,* vol. 61, no. 1, pp. 54-61, 1985.

[16] L. Bedat, A. Saadane, and D. Barba, "Masking Effects of Perceptual Color Components on Achromatic Grating," *Proc. European Conf. Visual Perception,* 1997.

[17] H. Senane, A. Saadane, and D. Barba, "Visual Bandwiths Estimated by Masking," *Proc. Eighth IEEE Workshop Image and Multidimensional Signal Processing,* 1993.

[18] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "Masking Effect in Visual Attention Modeling," *Proc. Workshop Image Analysis for Multimedia Interactive Services,* Apr. 2004.

[19] M.K Kapadia, M. Ito, C.D. Gilbert, and G. Westheimer, "Improvement in Visual Sensitivity by Changes in Local Context: Parallel Studies in Human Observers and in V1 of Alert Monkeys," *Neuron,* vol. 15, no. 4, pp. 843-856, 1995.

[20] M.K Kapadia, G. Westheimer, and C.D. Gilbert, "Spatial Distribution of Contextual Interactions in Primary Visual Cortex and in Visual Perception," *J. Neurophysiology,* vol. 84, no. 4, pp. 2048-2062, 2000.

[21] Z. Li, "A Neural Model of Contour Integration in the Primary Visual Cortex," *Neural Computation,* vol. 10, no. 4, pp. 903-940, 1998.

[22] Z. Li, "Pre-Attentive Segmentation in the Primary Visual Cortex," *Spatial Vision,* vol. 13, pp. 25-50, 1999.

[23] S. Grossberg and E. Mingolla, "Neural Dynamics of Perceptual Grouping: Textures, Boundaries, and Emergent Segmentation," *Perception and Psychophysics,* vol. 38, pp. 141-171, 1985.

[24] J.M. Henderson, P.A. Weeks, and A. Hollingworth, "Effects of Semantic Consistency on Eye Movements During Scene Viewing," *J. Experimental Psychology: Human Perception and Performance,* vol. 25, no. 210, 1999.

[25] S. Yantis and J. Jonidas, "Attentional Capture by Abrupt Onsets and Selective Attention: Evidence from Visual Search," *J. Experimental Psychology: Human Perception Performance,* vol. 20, pp. 1505-1513, 1996.

[26] A.P. Hillstrom and S. Yantis, "Visual Motion and Attentional Capture," *Perception Psychophysic,* vol. 55, pp. 399-411, 1994.

[27] A.B. Watson, "The Cortex Transform: Rapid Computation of Simulated Neural Images," *Computer Vision, Graphics, and Image Processing,* vol. 39, pp. 311-327, 1987.

[28] P. Le Callet, A. Saadane, and D. Barba, "Interactions of Chromatic Components on the Perceptual Quantization of the Achromatic Component," *SPIE Human Vision and Electronic Imaging,* vol. 3644, 1999.

[29] P. Le Callet, A. Saadane, and D. Barba, "Frequency and Spatial Pooling of Visual Differences for Still Image Quality Assessment," *SPIE Human Vision and Electronic Imaging,* vol. 3959, 2000.

[30] P. Le Callet and D. Barba, "Image Quality Assessment: From Sites Errors to a Global Appreciation of Quality," *PCS,* 2001.

[31] S. Daly, "A Visual Model for Optimizing the Design of Image Processing Algorithms," *Proc. IEEE Int'l Conf. Image Processing,* pp. 16-20, 1994.

[32] P.J. Burt and E.H. Adelson, "The Laplacian Pyramid as a Compact Image Code," *IEEE Trans. Comm.* , vol. 31, pp. 532-540, 1983.

[33] H.K. Hartline, "The Response of Single Optic Nerve Fibers of the Vertebrate Eye to Illumination of the Retina," *Am. J. Physiology,* vol. 121, pp. 400-415, 1938.

[34] D.S. Wooding, "Eye Movements of Large Population: II. Deriving Regions of Interest, Coverage, and Similarity Using Fixation Maps," *Behavior Research Methods, Instruments and Computers,* vol. 34, no. 4, pp. 509-517, 2002.

[35] B. Velichkovsky, M. Pomplum, and J. Rieser, "Attention and Communication: Eye-Movement-Based Research Paradigms," *Visual Attention and Cognition,* pp. 125-154, 1996.

[36] S.A. Brandt and L.W Stark, "Spontaneous Eye Movements During Visual Imagery Reflect the Content of the Visual Scene," *J. Cognitive Neuroscience,* vol. 9, pp. 27-38, 1997.

[37] C.M. Privitera and L.W. Stark, "Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, pp. 970-982, 2000.

[38] S. Mannan, K.H. Ruddock, and D.S. Wooding, "Fixation Sequences Made during Visual Examination of Briefly Presented 2D Images," *Spatial Vision,* vol. 11, pp. 157-178, 1997.

[39] M. Eigen, R. Winkleroswatitsch, and A. Dress, "Statistical Geometry in Sequence Space: A Method of Quantitative Comparative Sequence-Analysis," *Proc. Nat'l Academy of Sciences,* vol. 85, pp. 5913-5917, 1988.

[40] J.M. Wolfe and T.S. Horowitz, "What Attributes Guide the Deployment of Visual Attention and How Do They Do It?" *Nature Rev. Neuroscience,* vol. 5, pp. 1-7, 2004.

[41] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "Performance Assessment of a Visual Attention System Entirely Based on a Human Vision Modeling," *Proc. IEEE Int'l Conf. Image Processing,* 2004.
[42] J.H. Elder and R.M. Glodberg, "Ecological Statistics for the Gestalt Laws of Perceptual Organization of Contours," *J. Vision,* vol. 2, pp. 323-353, 2002.
[43] T. Hansen and H. Neumann, "A Computational Model of Recurrent, Collinear Long-Range Interaction in V1 for Contour Enhancement and Junction Detection," *Proc. Vision Sciences Soc., Second Ann. Meeting,* p. 42, 2002.
[44] L. Itti, "Models of Bottom-Up and Top-Down Visual Attention," California Inst. of Technology, Jan. 2000.

Olivier Le Meur graduated from Ecole Nationale Superieure des Sciences Appliquees et de Technologie in 1999. From 1999 to 2002, he worked with Nextream, a joint venture of Thomson and Alcatel, on MPEG-2 professional broadcasting applications. He has submitted his PhD thesis at the University of Nantes and is awaiting his oral examination. His doctorate research involves the modeling of the visual attention system. He is currently working on visual attention models for video coding applications.

Patrick Le Callet received the aggregation degree in electronics in 1996. He received the PhD degree in image processing from the University of Nantes, engineer in electronic and informatics, and he was also a student at the Ecole Normale Superieure de Cachan. He is an associate professor at the University of Nantes, where he is engaged in research dealing with the application of human vision modeling in image processing. His current centers of interest are human vision modelization and applications in the fields of image quality assessment, watermarking, and saliency map exploitation in image coding techniques. He is a member of the IEEE.

Dominique Barba received the MS degree from the University of Reims in 1968, the PhD degree with honors in telecommunications in 1972 from the University of Rennes and the Doctorat es Sciences Mathematiques degree (with honors) in computer science in 1981 from the University of Paris VI, in the field of digital image processing. He was an assistant professor at the University of Rennes in 1968, an associate professor at INSA of RENNES in 1973, and, in 1985, he joined the newly formed Engineer School at the University of Nantes, IRESTE (transformed in 2000 into Ecole Polytechnique de l'Universite de Nantes), with the position of full professor. He actively participated in its development and is in charge of its research activities. He launched a new research laboratory in Nantes, in the field of image processing and was the deputy director of IRCCYN (Institut de Recherche en Communications et Cybernetique de Nantes). His research interests include pattern recognition and image analysis, image and video description and compression with a high-quality reconstruction, and human visual system modeling with application to the design of objective image/video quality criterion. He is an author or coauthor of more than 300 papers in scientific journals or international and national conferences and is member of many scientist and technical societies (SPIE and SEE), and he is a senior member of the IEEE and the IEEE Computer Society.

Dominique Thoreau received the PhD degree in image processing and coding from the University of Marseille Saint Jerome, in 1982. From 1982 to 1984, he worked for GERDSM Labs on underwater acoustic signal and image processing. In 1984, he joined the Rennes Electronic Labs of Thomson and worked successively on sonar image processing, detection/tracking on IR picture, and picture coding. Currently, he is involved in the MPEG-4 AVC algorithm at Corporate Research Thomson, Rennes.