Predicting visual fixations on video based on low-level visual features

Olivier Le Meur, Patrick Le Callet, Dominique Barba

▶ **To cite this version:**

Olivier Le Meur, Patrick Le Callet, Dominique Barba. Predicting visual fixations on video based on low-level visual features. Vision Research, 2007, 47 (19), pp.2483-2498. 10.1016/j.visres.2007.06.015 . hal-00287424

HAL Id: hal-00287424
https://hal.science/hal-00287424v1

Submitted on 11 Jun 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A spatio-temporal model to predict visual fixation: description and assessment

Olivier Le Meur [a], Patrick Le Callet [b], Dominique Barba [b]

[a] *THOMSON R&D*
1 Avenue Belle Fontaine
35511 Cesson-Sevigne, France

[b] *IRCCyN UMR n° 6597 CNRS*
Ecole Polytechnique de l'Universite de Nantes
rue Christian Pauc, La Chantrerie, 44306 Nantes, France

Abstract

To what extent can a computational model of the bottom-up visual attention predict what an observer is looking at? What is the contribution of the low-level visual features in the attention deployment? To answer these questions, a new spatio-temporal computational model is proposed. This model incorporates several visual features; therefore, a fusion algorithm is required to combine the different saliency maps (achromatic, chromatic and temporal). To quantitatively assess the model performances, eye movements were recorded while naive observers viewed natural dynamic scenes. Four completing metrics have been used. In addition, predictions from the proposed model are compared to the predictions from a state of the art model (Itti's model (Itti et al., 1998)) and from three non-biologically plausible models (uniform, flicker and centered models). Regardless of the metric used, the proposed model shows significant improvement over the selected benchmarking models (except the centered model). Conclusions are drawn regarding both the influence of low-level visual features over time and the central bias in an eye tracking experiment.

Key words: salience, visual attention, eye movements, bottom-up, top-down

1 Introduction

Our visual environment contains much more information than we are able to perceive at once. In order to optimize the visual treatment of what surrounds us, we have evolved several biological mechanisms. Out of those, the visual attention is probably the most important one. It allows the detection of some interesting parts in visual field. It then guides the movement of the

eyes, allowing an accurate inspection of the chosen area by the fovea. This is where most of the processing resources are concentrated (Wandell, 1995). This implies that visual attention and eye movements are closely linked. This link is however not so obvious. In fact, two disjoint mechanisms for directing spatial attention have been identified. They are called covert and overt shift of attention. The former does not involve eye movements and refers to the act of mentally focusing on a particular area (Hoffman & Subramanian, 1995; Hoffman, 1998). The latter, involving eye movements, is used both to explore complex visual scenes and to direct the gaze towards interesting spatial locations. A number of studies (Findlay, 1997; Maioli et al., 2001) have shown that, in most circumstances, overt shifts of attention are mainly associated with the execution of saccadic eye movements. Saccade targeting is controlled by many factors: the task in mind (behavioral goals, motivational state) and both the local and global spatial properties of the visual scene. The former is also called top-down processing. The latter is called bottom-up or stimulus-driven selection. It occurs when a target item effortlessly attracts the gaze.
The design of computational models simulating the bottom-up human selective visual attention is a difficult issue. Existing approaches can be differentiated by the way they integrate or reproduce the visual system. Categorizing computational models would yield two main subsets. The first one would include models based on a statistical signal approach. A number of studies (Reinagel & Zador, 1999; Parkhurst & Niebur, 2003; Mack et al., 2003; Rajashekar et al., 2003) have indeed shown that fixated areas present higher spatial contrast, variance and local symmetry than non-fixated areas. The objective is then to design local operators able to detect these areas. The second category would include biologically plausible models, which are mainly based on two original concepts, the Feature Integration Theory from Treisman et al. (Treisman & Gelade, 1980) and a neurally plausible architecture proposed by Koch and Ullman (Koch & Ullman, 1985) (see figure 1). Among them, some models (Milanese, 1993; Itti et al., 1998; Le Meur et al., 2006) compute a topographic saliency map (or master map), which quantitatively predicts the salience of each location of the input picture whereas the others tackle the problem more broadly by attempting to mix together top-down and bottom-up interactions (Olshausen et al., 1993; Deco & Schurmann, 2000).

This paper aims at describing a spatio-temporal model of the bottom-up selective visual attention, purely based on the low-level visual features. The proposed model is an extension to the temporal dimension of previously published work (Le Meur et al., 2006), as illustrated in figure 1. An eye tracking experiment was conducted in order to investigate the relevance of the dynamic saliency map stemming from the proposed model. This experiment is presented in section II. The rationale of this evaluation rests on the assumption that eye movements and attention are correlated. This assumption was validated through several publications (Hoffman & Subramanian, 1995; Findlay,

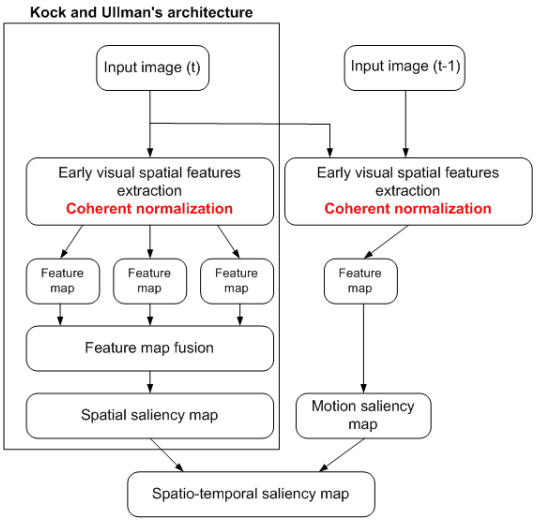


Fig. 1. Proposed framework based on Koch and Ullman (Koch & Ullman, 1985) model. This influential model describes how salient features can be identified in natural scenes. First, early visual features are extracted from the visual input into several separate parallel channels. A feature map is obtained for each channel. A unique saliency map is then built from the combination of those channels. The major novelty proposed here lies in the inclusion of the temporal dimension as well as the addition of a coherent normalization scheme.

1997; Maioli et al., 2001). Section III focuses on the proposed spatio-temporal model. It has been previously described in several papers (Le Meur et al., 2005, 2006), therefore, only its major features will be discussed here. Compared to the original model, three novelties are implemented. The first one deals with the computation of two chromatic saliency maps. The temporal dimension was also added to the static model. The aim is to detect the contrast of motion. This is deemed to be one of the strongest attractor of attention (Wolfe, 1998; Itti, 2005).

The underlying principle of this study rests on the assumption of the existence of a unique topographic saliency map. This assumption is strong because there is no consensus on this point. In a recent paper, Fecteau and Munoz (Fecteau & Munoz, 2006) concluded that the concept of saliency map must be first broadened to include top-down influences, leading to a new map, called priority map. The locus of the priority map is also an open-issue. Fecteau and Munoz (Fecteau & Munoz, 2006) also concluded that this map is more the result of a network involving different brain areas than the result of a particular area of the brain. Keeping in mind both the aforementioned assumption and its limitations, the four different saliency maps have to be combined to form an unique map. What is the best way to combine these maps arising from different visual dimensions? A fusion algorithm is proposed. Section IV examines the similarity degree between experimental and predicted saliency maps. Several models, featuring different levels of complexity, are used (proposed model, uniform, centered...). Some conclusions are drawn in section V.

2 Eye tracking apparatus and experiment procedure

2.1 Eye tracking apparatus

Eye movements of real observers were tracked using a dual-Purkinje eye tracker from *Cambridge Research Corporation*. The eye tracker is mounted on a rigid EyeLock headrest that incorporates an infrared camera, an infrared mirror and two infrared illumination sources. To obtain accurate data regarding the diameter of the subject's pupil, a calibration procedure is mandatory. It requires the subject to view a number of 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. The camera records a close-up image of the eye. The video is processed real-time and spatial location of eye position is extracted. Both Purkinje reflections are used to calculate this location. The guaranteed sampling frequency is 50 Hz. The mean spatial accuracy of the eye tracker is $0.5 \pm 0.25°$.

2.2 Subjects

Unpaid subjects (see Table 1) participated to the experiments. They came from both the University of Nantes and Thomson R&D Rennes. All had normal or corrected to normal vision. All were unexperienced observers (not expert in video processing) and naive to the experiment. Before each trial, the subject's head was positioned such that their chin rested on the chin-rest and their forehead rested against the head-strap. The height of chin-rest was adjusted so that the subject was comfortable and their eyes level with the center of the presentation display.

2.3 Stimuli

Seven video clips ($25Hz$, 352×288, $4.5 - 33.8$ seconds, for a total of 2451 distinct frames, see figure 2) have been selected for several reasons. First and foremost, these video clips contain important spatio-temporal events that are classically used in TV broadcast (zoom in/out, translation motion with different velocities, still video, fade-in/fade-out, scene cuts...). The second aspect concerns the various content of clips (such as faces, sporting events, audiences, landscapes, logos, incrustations, one, two or no actors in the scene, low and high spatio-temporal activities...).

2.4 Experimental procedure

Each video clip was presented to subjects as part of a free-viewing exercise. The subjects were instructed to "look around the image". The free-viewing task is an important aspect of the experiment. The objective is to lessen the top-down influences or to encourage a bottom-up behaviour. Obviously, it is impossible to completely remove all top-down influences.
Each trial begins with a calibration phase: the observers are asked to sequentially fixate a series of nine fixation circles. In order to ensure a high degree of relevance, the calibration of the eye tracker is intermittently repeated between video clips as required. Clips are presented on a CRT display with a resolution of (800×600). The active screen size is 36×27 and the viewing distance is 81cm ($25° \times 19°$ usable field-of-view).
Between the calibration phase and the beginning of the video clip, observers are asked to fixate a small square centered on the screen. This keeps the observer concentrated before the stimulus onset. This approach can strenghten the importance of the central locations and may induce a significant change in the results (Tatler et al., 2005). This general tendency for observers to fixate near the center of scenes, whatever the salience, has a number of reasons (Parkhurst et al., 2002; Tatler et al., 2005). In this study, this choice is not so important as the degree of similarity between prediction and human fixations is computed by taking into account all fixations, not only those occurring just after the clip onset.

2.5 Experimental priority maps

As recommended in (Fecteau & Munoz, 2006), the expression, "priority map", (Serences & Yantis, 2006) is much more suitable to feature the results coming from eye tracking experiments. Indeed saliency map is a conceptual framework in which neither the relevance of an object nor the goals of observers are taken into account. In the following, the term priority map will be used.
The raw eye tracking data was segmented into fixations, saccades and eye blinks. Analysis of the eye movement record was carried out off-line. Fixations were characterized by consecutive eye data having a velocity below a given threshold (the fixation label includes smooth-pursuit periods). This type of algorithm is generally called velocity-threshold fixation identification (Salvucci & Goldberg, 2000). The velocity is the distance between two consecutive points multiplied by the sampling frequency. The distance threshold was set at 1° visual angle. This choice is rather arbitrary even if it is coherent with both previous works (Sen & Megaw, 1984; Itti, 2005) and spatial accuracy of the eye tracking apparatus. Results from the segmentation of the raw eye tracking experiments are described in Table 1. These results are consistent with those

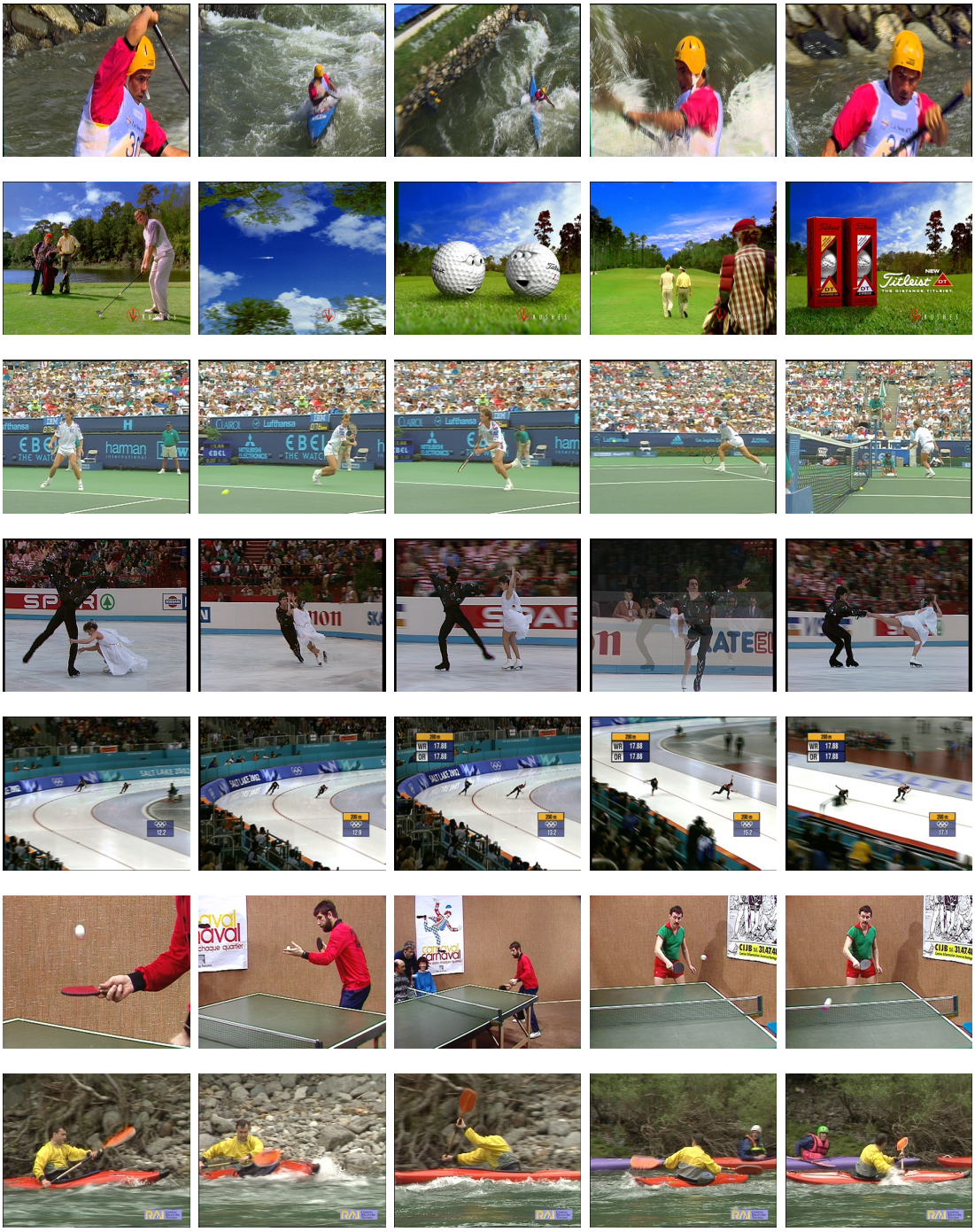


Fig. 2. Representative pictures from the video clips. From top to bottom, pictures respectively concern clips called *Kayak*, *Titleist*, *Stefan*, *Skate*, *Speed Skate*, *Table* and *Canoa*. These clips feature a varied content (one or more regions of interest per picture, centered on the screen or not, spatio-temporal events (cut, fade) or continuous clip, incrustation of text and logo).

Table 1
Summary of the results obtained from the velocity-threshold fixation identification. The average fixation duration, the total number of fixations and the number of fixation per second are shown for each clip and for all observers. SEM pertains for standard error of the mean, obtained by dividing the standard deviation by the square root of the sample size (confidence of 95%).

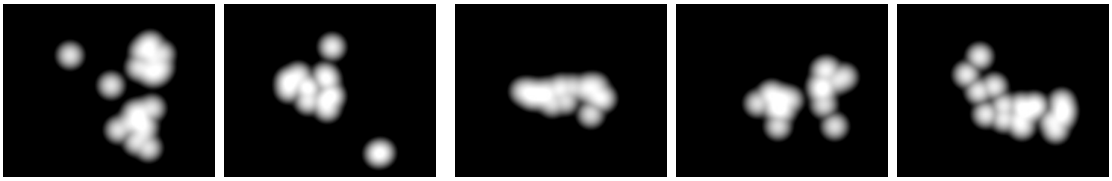
Clip	Number of observers	Fixation duration (ms) $mean \pm sem$	Total number of fixations $mean \pm sem$	Number of fixation per second $mean \pm sem$
Titleist	25	317 ± 32	106 ± 13.4	3.6 ± 0.46
Stefan	27	340 ± 26	36 ± 3	3.09 ± 0.29
Skate	25	368 ± 36	67 ± 9	3.06 ± 0.41
Canoa	23	480 ± 66	23 ± 2.9	2.67 ± 0.33
Kayak	20	389 ± 35	23 ± 2.9	3.05 ± 0.38
SpeedSkate	17	337 ± 24	14 ± 0.7	2.91 ± 0.14
Table	20	273 ± 26	34 ± 3.8	3.8 ± 0.4
Average	22	343	$Sum = 303$	3.16

generally obtained. For instance, there are about 2 to 4 eye fixations per second.

A fixation map per observer is computed from the collected data at a picture level (along the sequence). Fixation maps are then averaged over the observers for each picture. This indicates the most visually important regions for an average observer. Finally, each frame of the average saliency sequence is processed using a 2D Gaussian filter. The standard deviation σ is determined in accordance with the accuracy of the eye-tracking apparatus.

Figure 3 shows several experimental priority maps, extracted from *Titleist* clip. On the same figure, the influences of scene cuts on the visual attention are underlined. An abrupt change in the content of the entire image induces an increase of the temporal masking. This refers to the inability of the human vision to instantaneously adjust to changes in its visual field. Previous studies (Seyler & Budrikis, 1959, 1965; Tam, 1995) demonstrated that the perception is reduced for period of up to 100 ms following the scene change. Figure 4 presents priority maps before and after a scene cut. Before the cut, the salient areas are quite coherent (the boy and the logo are well detected). After the scene cut, the same areas still remains the same during a periods of 200 ms. As shown in (Tatler et al., 2005), the fixation position is dependent on the content that was displayed prior to the change.

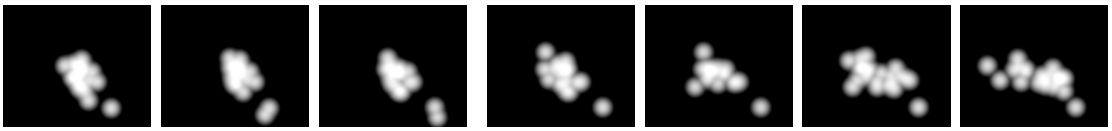
(a)



(b)

Fig. 3. Experimental priority maps. (a) show several pictures, extracted from the *Titleist* clip. These pictures represent different parts of the video clip. (b) show the corresponding experimental priority maps.

(a)



(b)

Fig. 4. Experimental priority maps around a scene cut. (a) show several pictures, extracted from the *Titleist* clip separating by a temporal interval equal to 100 *ms*. (b) show the corresponding experimental priority maps. A scene cut occurs after the fourth picture. It is interesting to notice the non influence of scene cut on priority maps. Immediately after the scene cut, the fixation position is dependent on the content that was displayed prior to the change. It illustrates the temporal masking induced by a scene cut.

3 Dynamic saliency map computation

The proposed biologically plausible computational model implements many of key features occurring during early vision process. The related synoptic is shown in figure 5. Compared to the preliminary design (Le Meur et al., 2006), several improvements and modifications have been made. The first improvement brought to the model aims at building two chromatic saliency maps, enhancing its capacity to detect visually important locations. The second point concerns temporal saliency computation. The spatial model now

yields four saliency maps (achromatic, two chromatic and temporal saliency maps), which have to be coherently fused into one. Two combination strategies are proposed. Achromatic reinforcement and facilitative interactions, present in the preliminary version, are now disabled. Indeed, it would be redundant to use achromatic reinforcement since two chromatic saliency maps are now computed. Concerning the facilitative interactions, its impact is not significant enough to justify keeping it in the new design.

3.1 Dynamic salience model

3.1.1 Computation of the spatial salience

Psychovisual space

Despite the seemingly complex mechanisms underlying human vision, the visual system is not able to perceive all information present in the visual field with the same accuracy. Several mechanisms have been used and are accurately described in a recent paper (Le Meur et al., 2006).

First, the RGB picture is projected into the Krauskopf's color space (A, Cr_1,Cr_2) simulating the three different pathways used by the brain to encode the visual information (Krauskopf et al., 1982). The first pathway conveys the achromatic component (A), the second the red and green antagonist component (Cr_1) and the third the blue and yellow antagonist component (Cr_2).

In order to express all data in the same unit (in term of visibility), three contrast sensitivity functions are used, one per component. If components (A, Cr_1,Cr_2) can be described in terms of their sinusoidal Fourier components, then the visibility of each spatial frequency can be measured by applying a contrast sensitivity function. Each spatial frequency is then compared to a threshold CT_0. If the amplitude is above this threshold, the frequency is perceptible. This threshold is called the visibility threshold and its inverse defines the values of the contrast sensitivity functions (CSF) at this spatial frequency. While the CSF shows how sensitivity varies with spatial frequency and orientation, they do not take into account the changes in sensitivity caused by the context [1] . This modulation is commonly called visual masking. Figure 6 illustrates this property. It is necessary to replicate the hierarchical structure of the visual system to account for visual masking. Indeed electrophysiological measurements revealed that visual cells are tuned to certain types of visual information such as frequency, color and orientation. A hierarchical decomposition is then conducted 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. Psycho-visual spatial frequency partitioning for the achromatic component leads to 17 psycho-visual channels in

[1] CSF are deduced using psychophysic experiments involving very simple cues.

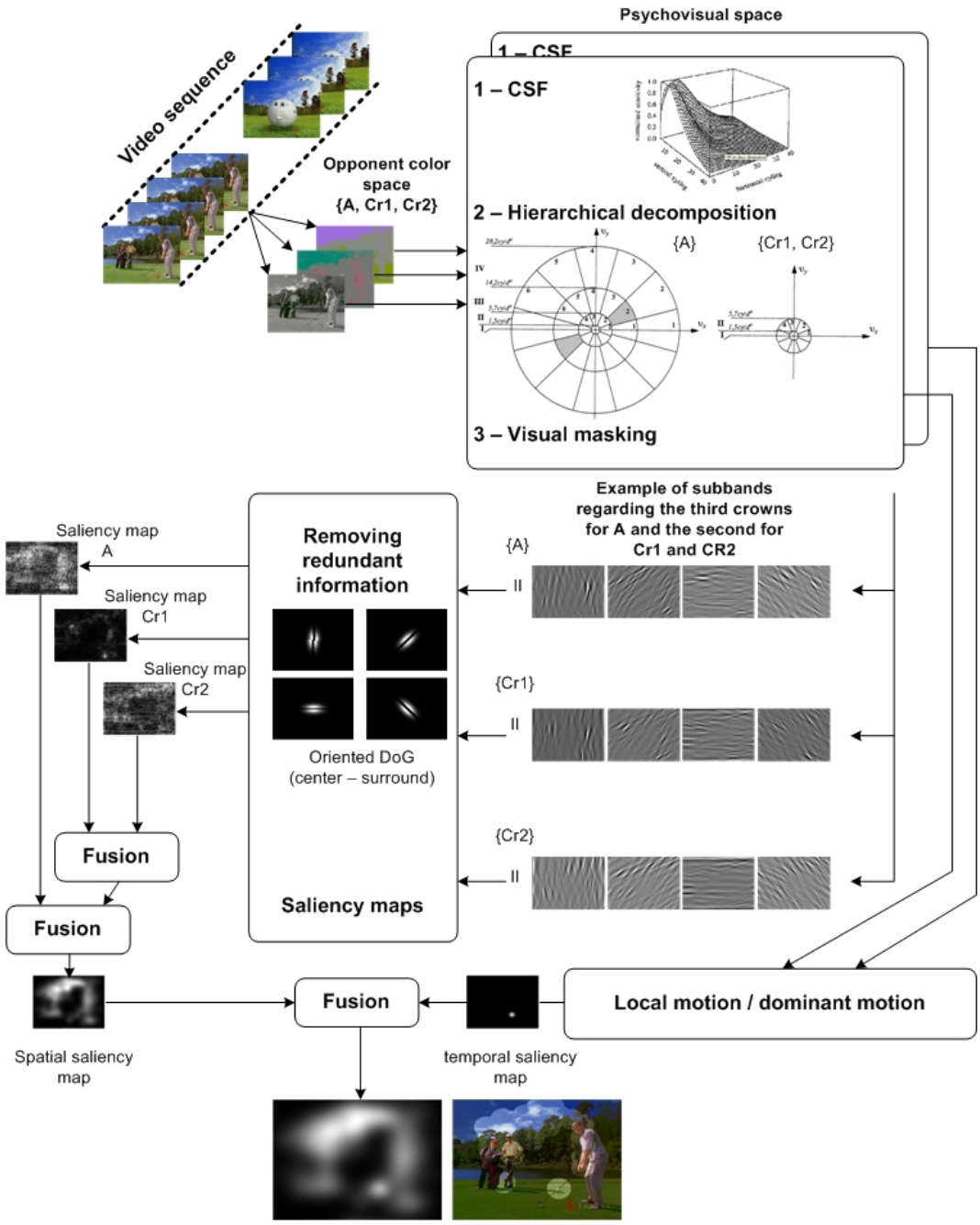


Fig. 5. Flow chart of the proposed spatio-temporal model. The model takes a video sequence as input and processes all the frames in three parallel channels using a range of spatial scales and orientation values. It yields a saliency map indicating the most salient region per image.

standard TV viewing conditions while only 5 channels are obtained for each chromatic component (see figure 5). The acromatic (respectively chromatic) channels are spread over 4 (respectively 2) crowns. Each resulting subband or channel may be regarded as the neural image corresponding to a particular population of cortical cells. These cells are tuned to a range of spatial fre-

(a) (b)

Fig. 6. Illustration of spatial visual masking: (a) original picture; (b) original picture corrupted by uniform noise. Interestingly, the noise is more noticeable in the sky than on the seashore.

quencies and to a particular orientation. Finally, the masking effect alters the differential visibility threshold of each subband. Three types of masking are considered in the proposed model: intra-channel intra-component masking, inter-channel intra-component masking and inter-component masking. Visual masking has been described elsewhere (Le Meur et al., 2006).

Removing redundant information

Since the visual system cannot process all visual information at once, two kinds of mechanisms are required to cope with this biological limitation. The first one selects a small part of the visual field on which a close inspection is performed. The second is more passive than the previous one. Its role is very important as it suppresses the redundancy of the visual information yielding an economical representation of the visual world. In a cluttered environment, this process allows the selection of the most informative areas (Tsotsos, 1990). To suppress irrelevant data, a center-surround filter is applied (Le Meur et al., 2006).

Saliency maps

After applying the center-surround filters, three saliency maps are derived: first, a two-dimensional achromatic saliency map, called S^A, was computed from the direct sum of the outputs of the achromatic channels belonging to the crown III. Second, two chromatic saliency maps were computed by the direct summation of the outputs of the chromatic channels belonging to the crown II:

$$S^A(s) = \sum_{s,\rho,\theta} \left(\alpha_A \times \tilde{R}^A(s,\rho,\theta) \right) \quad (1)$$

$$S^{Cr1}(s) = \sum_{s,\rho,\theta} \left(\alpha_{Cr1} \times \widetilde{R}^{Cr1}(s,\rho,\theta) \right) \tag{2}$$

$$S^{Cr2}(s) = \sum_{s,\rho,\theta} \left(\alpha_{Cr2} \times \widetilde{R}^{Cr2}(s,\rho,\theta) \right) \tag{3}$$

where, $\widetilde{R}^{x}(s,\rho,\theta)$ is the value of the site s of the component x modified by the center-surround filters. The values ρ, θ are respectively the radial frequency and the orientation of the considered subband. In the initial version, α_a, α_{Cr1}, α_{Cr2} are set to one.

3.2 Computation of the temporal salience

It is generally accepted that motion pathway in the monkey cortex sequentially involves areas V1, MT, MST and 7a. These areas contain population of neurons specialized for certain tasks. What is interesting to point out is that the task complexity as well as the receptive field size increase with this hierarchy. Although the literature is large on the topic, the objective here is not to provide a computationally plausible version of this motion processing hierarchy as in (Tsotsos et al., 2005), but rather to propose a straightforward approach. Nevertheless, from the moment it is possible, a comparison between the proposed approach and previous ones will be done.
The assumption here is that the motion contrast is one of the most important visual attractors. With regards to dynamic complex scenes, previous studies (Itti, 2005) have indeed shown that the motion contrast is a much more reliable predictor of salient areas than the others.
The basic aim of the temporal saliency map computation (Le Meur et al., 2005) rests on the relative motion occurring in the retina. The relative motion is the difference between the local and the dominant motion. The local motion $\overrightarrow{V}_{local}$ at each point s of an image (or the motion vector) is given by using a hierarchical block matching. It is computed through a series of levels (different resolution), each providing input for the next. In addition, on each level, the block matching is done for a certain neightboordhood size, that increases with the hierarchy level. In a way, these two points remind the properties of the motion processing in the monkey cortex.
The local motion does not necessarily reflect the motion contrast. It is the case when the dominant motion is null, meaning that the camera is fixed. As soon as the camera follows something in the scene, it is necessary to estimate the global transformation that two successive images undergo. This global transformation, or the dominant motion, is estimated from the previous estimated local motion. The dominant motion is represented by a 2D parametric model, noted $\overrightarrow{V}_{\Theta}(s)$; Θ is a parameter vector containing the 2D affine motion model $[a_1, a_2, a_3, a_4, a_5, a_6]$. These parameters are computed with a popular robust technique based on the M-estimators (Odobez & Bouthemy, 1995).

Finally, the relative motion representing the motion contrast is given by:

$$\overrightarrow{V}_{relative}(s) = \overrightarrow{V}_{\Theta}(s) - \overrightarrow{V}_{local}(s) \tag{4}$$

The relevance degree of a relative motion also depends on the average amount of relative displacement over the picture. For example, a high relative motion is more conspicuous when there are only few relative displacement (Wolfe, 1998). Recently, Fencsik et al. (Fencsik et al., 2005) confirm and extend the finding of Ivry and Cohen (Ivry & Cohen, 1992). This previous study showed that, when targets moved faster than distractors, target-detection time was minimally affected by set size. Targets pop-out. Nevertheless, search for a target moving slower than distractors is markedly more difficult.
The idea to partially reproduce this property is to predict the amount of relative motion. A good candidate is the median value of the relative motion, called in the following Γ_{median}. This value is readily deduced from a histogram. $\|\overrightarrow{V}_{relative}\|$ is then weighted by Γ_{median} in order to predict the temporal salience S^T:

$$S^T(s) = \frac{\|\overrightarrow{V}_{relative}(s)\|}{1 - \Gamma_{median}} \tag{5}$$

On one hand, the closer Γ_{median} to 0, the more the relative motion is perceptually important (the case of a moving target among stationary distractors). In the other hand, a high value of Γ_{median}, meaning that numerous parts of the image undergo a displacement, lessens the salience. Indeed it is easier to find a moving stimulus among stationary distractors (Γ_{median} close to 0) than a moving stimulus among moving distractors (high value of Γ_{median}).

3.3 The fusion

The combination of different saliency maps into a unique map is difficult. However, such process is mandatory when several maps are considered. This yields a single measure of interest for each location, regardless of which features contributed to the salience. Major problems arise when attempting to merge features stemming from different visual dimensions and having different dynamic ranges. Before going into the details of the proposed fusion, two influent attempts at solving this problem are described.
The most recent is the work of L. Itti et al. (Itti & Koch, 2001) in which three bottom-up combination strategies are proposed:

(1) the simple normalized summation (called NS in the following) is the

simplest method. All the conspicuous maps are normalized to the same dynamic range and are summed. This normalization scheme presents two main advantages: its simplicity and its rapid adaptability. Nevertheless, this type of strategy suffers from several drawbacks. First, if a conspicuous map is homogeneous, its contribution is relevant to the final saliency map, even if it is not relevant in itself. Second, if there are several conspicuous regions and if the saliency value of one of these regions is drastically greater than the others, this simple normalization scheme only promotes the region having the highest value. The others are lessened.

(2) the global non-linear normalization followed by summation is more elaborate than the previous one and attempts to correct the drawbacks of the previous fusion scheme. The principle is to promote the feature maps having a sparse distribution of saliency and to suppress the feature maps having numerous conspicuous locations. Therefore, all the feature maps, normalized to the same dynamic range, are weighted by their global maximum. There are two drawbacks clearly identified by the author. This method is not biologically plausible and is sensitive to noise. Another point has to be emphasized: a preliminary normalization is required in order to scale the feature maps to the same dynamic range. Therefore, the proposed method has to deal with the same problem encountered in the simple normalized summation.

(3) the local non-linear normalization followed by summation (called $LNLN$ in the following) is based on a simple iterative within-feature spatial competition. Similarly to the previous strategy and presenting the same advantages, the feature maps are now locally altered, considering the neighborhood around the current position, instead of the entire picture. Several iterations are required in order to converge to the most conspicuous regions. This method is insensitive to noise and is more biologically plausible than the previous ones. Nevertheless, as before, this strategy requires that all the feature maps have the same dynamic range. The global maximum of each map is used to normalize the dynamic range.

These techniques present some innovative points. However, they all suffer from a major drawback. In each of the three proposed strategies, the saliency maps are first normalized to the same dynamic range. As some of the saliency maps may be noisy or irrelevant, it is not really appropriate to start the fusion process with a global normalization scheme. Moreover, these approaches do not consider the complementarities between the different maps. The saliency maps might enhance the same regions of the picture, even if they are obtained from different approaches. This point is important and should not be overlooked.
R. Milanese has tackled the fusion issue more completely in 1993 (Milanese, 1993). The author proposed a coherent framework to compute a single saliency map. This map should be a "summarizing" function of the feature maps. The hallmark of his approach relies on both intra and inter-map competition. The former, the intra-map incoherence, is similar in spirit to the global non-linear

normalization scheme proposed by L. Itti. A uniform saliency map has no effect on the final result. The latter deals with the local inter-map incoherence that potentially exist in the set of the feature maps. The fusion process is thus driven by the knowledge of both the conflicting regions in different saliency maps and the regions where there is no ambiguity (all saliency maps present a relevant saliency at the same location).

The proposed normalization scheme is based on the two aforementioned types of competition. Its schematic diagram is shown in figure 7. The fusion of two feature maps, noted S_1 and S_2, is called $\mathcal{F}(S_1, S_2)$ and is given by

$$S(s) = \mathcal{F}(S_1, S_2)(s) = intraMap(s) + InterMap(s) \tag{6}$$

where, the term $intraMap$ (respectively $interMap$) pertains to the intra-map competition (respectively the inter-map competition).
Before looking into details of the intra and inter-map competitions, it is necessary to normalize the two feature maps S_1 and S_2 (the two normalized maps are noted S_1^N and S_2^N). The dynamic range of each saliency map are normalized by using the theoretical maximum of the considered feature rather than the global maximum.
In the proposed design, the three theoretical maximum values that characterize the maximum dynamic range of the three spatial saliency maps have been defined in a heuristic way. The method is described hereafter (A, Cr_1 or Cr_2). First, the maximum input dynamic range is calculated for each component (A, Cr_1 or Cr_2). The computation of the theoretical maximum related to the achromatic component is taken as an example. This component can convey data having maximum amplitude of 85.51 (this value (85.51) is the maximum output of the opponent color space regarding the component A). A test pattern is then built. It is strictly composed by achromatic target (square, circle...) having the highest possible amplitude (85.51). The maximum saliency value generated by this pattern is then noted. This experiment is repeated several times for each component with different targets having different sizes, locations, etc. From the collected data, the theoretical maxima [2] are 1, 18 and 26 respectively for the component A, Cr_1 and Cr_2. A slight modification of the theoretical maximum is not critical for the results accuracy. However, an over-estimation will favor the most important saliency peaks to the detriment of the others. Conversely, small theoretical maximum may promote saliency peaks that are not relevant.
Intra-map competition favors the most important salience locations. This process alters each spatial location in function of the value of the nearest local

[2] Compared to the chromatic theoretical maximum values, the achromatic maximum value is small. It is due to the fact that the achromatic CSF is applied on a contrast, whereas the chromatic CSFs are directly applied on the spectrum.

maximum. It is given by

$$IntraMap(s) = \frac{S_1^N(s)}{NearestMax_1} + \frac{S_2^N(s)}{NearestMax_2} \tag{7}$$

where the term $NearestMax_1$ (respectively $NearestMax_2$) indicates the value of the nearest local maximum regarding the current position for the component S_1 (respectively S_2). For each feature map, the first K maximums are computed. When the k^{th} local maximum is located and memorized, its neighbors are inhibited in order to determine the spatial location of the $(k+1)^{th}$ local maximum, as in a winner-takes-all algorithm (see figure 7). The size of the surrounding area equals 3 degrees of visual angle. This local normalization promotes local saliency value, allowing to process configurations where several strong local saliences are present. It is worth noting that all local maximum are not systematically taken into account. Local maximum that are considered are the local maximum for which the gradient of saliency is greater than a given threshold. This approach keeps only the most visually important locations.
As R. Milanese proposed, the inter-map competition relies on using complementarities and redundancies that the different feature maps could present. The complementarity of the feature maps is interesting because the visual attention can be attracted by only one visual dimension. For instance, a red circle sitting amongst a set of black circles stands out. In this case, the chromatic saliency maps contain the most relevant saliency. Moreover, the salience of a stimulus can also be located in more than one feature map. Taking into account these considerations, the inter-map competition is given by the multiplication of the two saliency values locally normalized:

$$InterMap(s) = \frac{S_1^N(s)}{NearestMax_1} \frac{S_2^N(s)}{NearestMax_2} \tag{8}$$

In the proposed model, the overall approach to combine the saliency maps is based on a hierarchical structure. The resulting saliency map S is given by

$$S(s) = \mathcal{F}(S^T, \mathcal{F}(S^A, \mathcal{F}(S^{Cr_1}, S^{Cr_2}))) \tag{9}$$

Firstly, a chromatic saliency map is computed from the maps S^{Cr_1} and S^{Cr_2}. It is quite logical to fuse first the two chromatic maps into one. As luminance and color information are of same type contrary to motion, it is also logical to fuse them into one map. A spatial saliency map is then deduced from the chromatic and the achromatic saliency maps. Secondly, the dynamic saliency map is the result of the fusion of the spatial and the temporal saliency maps.

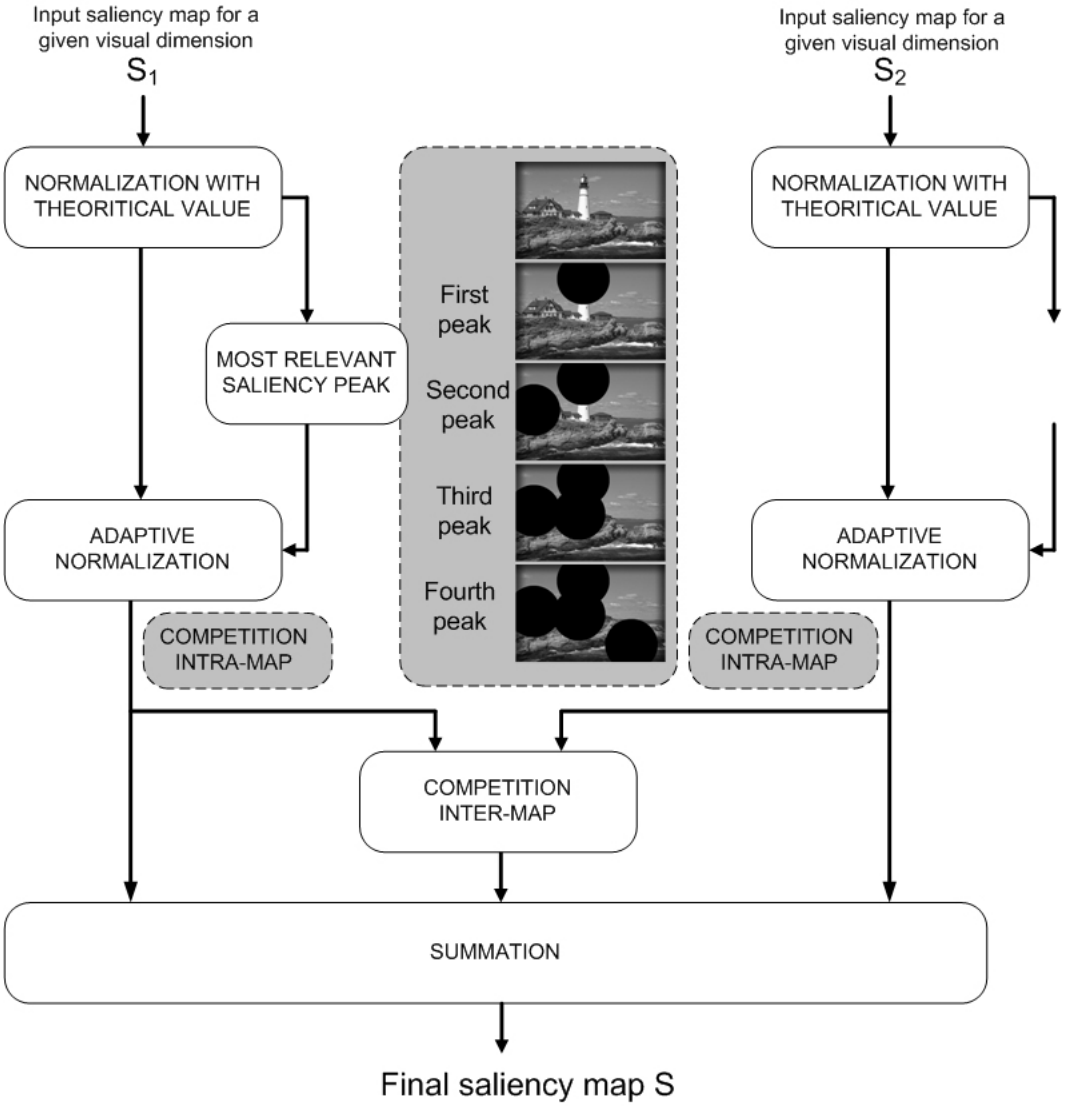


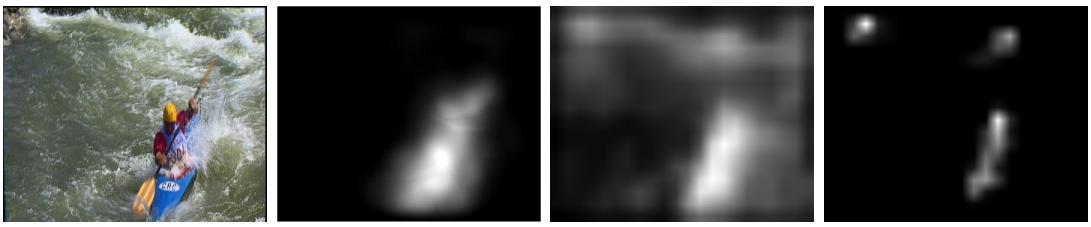
Fig. 7. Schematic diagram of the proposed fusion algorithm. In the proposed example, the fusion aims at building a final map from two saliency maps. Each of those saliency maps goes through several steps. The saliency map is first normalized to the theoretical maximum. The detection of the most relevant saliency peaks (by using a winner-takes-all algorithm) is performed in order to achieve an intra-map competition: the weakest saliency values are lessened whereas the strongest saliency peaks are promoted. An inter-map competition is finally used to detect the complementarity and the redundancy that the two saliency maps could present.

4 Human and predicted fixations: comparison and analysis

To quantitatively examine the similarity degree between the predicted and the priority maps, four completing metrics have been used.

4.1 Different models

In this study, different models are compared. Firstly, three versions of the proposed model are used: a model which includes all data (spatial and temporal, noted **ST**), a model based only on the spatial visual features (noted **S**) and a model based only on the temporal dimension (noted **T**). Secondly, the model proposed by L. Itti (Itti et al., 1998) is put to the test. This model is freely down loadable from the Internet and the command line used to conduct this experiment is: *ezvision.exe –wta-type=None –in=./kayak/kayak#.pnm –rescale-output=352x288 –save-salmap –out=raster*. By default, the fusion scheme is the the local non-linear normalization followed by summation ($LNLN$). By adding the option *–maxnorm-type=Maxnorm*, the fusion of the feature maps is based on the normalized summation (NS). Parameters are detailed on the iLab's web site [3] . Figure 8 gives an example of saliency maps obtained by the proposed approach, Itti's model with NS fusion and with $LNLN$ fusion.



(a) Source (b) Proposed (c) NS (d) LNLN

Fig. 8. Predicted saliency maps. (a) source (b) saliency map coming from the proposed model; (c) and (d) saliency maps coming from Itti's model respectively for NS and $LNLN$ fusion scheme. The difference between the two last saliency maps is obvious. The latter is more focused than the former.

These models have been designed by considering and simulating several fundamental properties of the human visual system. This is not the case for the three remaining models: the first one is the most simple as it computes a uniform saliency map. The second one favors the center of the screen. The centered saliency map is given by

$$S(s) = exp(-\frac{d(s, s_0)}{\sigma_c}) \tag{10}$$

s and s_0 are respectively the current and the center position. $d(s, s_0)$ is the Euclidean distance between the location s and s_0. σ_c controls the spread of

[3] http://ilab.usc.edu/

the saliency distribution. By default, σ_c is set to 50.
The two aforementioned models are limited to the spatial dimension. For completeness, a final model is defined. It is based on temporal changes (flicker) and is expected to yield interesting results during motion video sections. The saliency map computed from the flicker model saliency is given by

$$S(s) = |I_t(s) - I_{t-1}(s)| \tag{11}$$

I_t and I_{t-1} represent the frame at the time t and $t-1$ respectively. For the particular case of the first picture, the saliency map is uniform.

4.2 Linear correlation coefficient and KL-divergence

The first metric used here is the linear correlation coefficient, noted *cc*. This metric assesses the linearity degree between two data sets. The *cc* range is between -1 and 1. When the correlation value is close to -1 or 1, there is almost a perfect linear relationship between the two variables:

$$cc(p,h) = \frac{cov(p,h)}{\sigma_p \sigma_h} \tag{12}$$

with,
h and p respectively represent the priority map and the predicted density map, $cov(p,h)$ is the covariance value between p and h.
The second metric is the Kullback-Leibler divergence, noted KL. The KL-divergence (Cover & Thomas, 1991) estimates the dissimilarity between two probability density functions. It is not a distance, since the KL-divergence is not symmetrical, nor does it satisfy an inequality:

$$KL(p|h) = \sum_x p(x) Log(\frac{p(x)}{h(x)}) \tag{13}$$

with,
h the probability density deduced from the experimental priority map, and p the predicted probability density function.
When the two probability densities are strictly equal, KL-divergence value is zero.
Table 2 shows the overall similarity degree between different models in predicting attentional selection. As expected, the uniform model yields the worst results. This finding reinforces previous conclusions (Itti, 2005), suggesting human beings tend to look at salient objects in their visual environment on

dynamic color scenes. Concerning the flicker saliency maps, both *cc* and *kl* indicate that flicker, as it is defined in this study (absolute frame difference), is not a reliable predictor of human saccade targets. Indeed flicker only indicates temporal changes and therefore its predictions can be relevant only when the dominant motion is null.
It is worth stressing that as far as biologically plausible models are involved, the best performances stem from the proposed model incorporating all features (spatial and temporal). This is coherent with previous findings (Itti, 2005). This is not surprising, as emphasized in (Itti, 2005): during periods of rather still video, temporal saliency map yields no output. The best predictors are then those stemming from the spatial dimension (luminance and color).
When the temporal and the spatial saliency maps are mixed together, the gain is about 0.1 and at least 1.25, for *cc* and *kl* respectively. Whereas the best proposed model (noted Proposed ST in Table 2) yields 0.41 and 19.21, for *cc* and *kl* respectively, L. Itti's model with all feature channels (color, intensity, orientation, flicker and motion) gives 0.32 (NS version) and 22.21 ($LNLN$ version) respectively. It is interesting to note that the best performances in term of *cc* is obtained when considering Itti's model with the NS fusion. However, it is the other fusion scheme ($LNLN$) that provides the smallest KL-divergence value. It is not suprising because the saliency distribution given by the two fusion shemes (NS and $LNLN$) is significantly different. The saliency distribution obtained by the former (NS) is almost uniform whereas the latter is very sparse (see an example on figure 8).
Figures 9 and 10 show the average *cc* and *kl* per clip. Several observations can be made: the flicker model gives its best results on video sequences containing periods for which the dominant motion is null (still camera). This is the case for the sequences *Table* and *Titleist*. The best performances are given by the spatio-temporal model, except for the clip *Canoa* and *Kayak*. These video clips are particular in the sense that they consist of one contrasting actor moving across the scene. As the contrast between the background and the main actor is important, this can explain why the spatial proposed model performs better than the spatio-temporal one. Noting also that the proposed fusion outperforms the NS fusion, in term of *cc* and *kl*. However, the gain in *cc* brought in by the coherent fusion is not statistically significant.

4.3 Cumulative probability

The cumulative probability metric is close to those proposed by Parkhurst et al. (Parkhurst et al., 2002). Priority and predicted maps are first transformed into two probability density functions. Next, the coordinates (x^k, y^k) of the k^{th} most important saliency location are extracted from the experimental probability density function. To identity the $k{+}1^{th}$ most important saliency location,

Table 2
Compared models. **Proposed ST**: model with all feature channels (ST pertains for Spatio-Temporal). **Proposed S**: achromatic and chromatic components only (S pertains for Spatial). **Proposed T**: temporal dimension only (T pertains for temporal). The proposed fusion is used for these three models. **Proposed ST +** NS **fusion**: model with all feature channels, however, the final saliency maps is obtained by summing all saliency maps, that have been normalized to the same dynamic range. **L. Itti**: Itti's model with all feature channels (color, intensity, orientation, flicker and motion) with NS or $LNLN$ fusion. **Uniform**: uniform model. **Centered**: centered model. **Flicker**: model based on the frame difference. Average cc and kl, computed over all clips and for different models are given. Significance level, calculated from a non-parametric paired-sample test between the proposed spatio-temporal (Proposed ST) model and the others is given. Standard error of the mean (sem) is also given.

Model	cc	t-test	kl	t-test
	$mean \pm sem$		$mean \pm sem$	
Proposed ST	0.41 ± 0.001	$--$	19.21 ± 0.033	$--$
Proposed S	0.32 ± 0.003	$p < 0.086$	21.51 ± 0.038	$p < 0.013$
Proposed T	0.31 ± 0.001	$p < 0.00042$	20.46 ± 0.08	$p < 0.21$
Proposed ST + NS fusion	0.37 ± 0.001	$p < 0.36$	22.28 ± 0.01	$p < 0.002$
L. Itti NS	0.32 ± 0.002	$p < 0.14$	23.37 ± 0.012	$p < 0.0004$
L. Itti LNLN	0.28 ± 0.002	$p < 0.033$	22.21 ± 0.017	$p < 0.003$
Uniform	0.01 ± 0.00	$p < 10^{-6}$	25.36 ± 0.01	$p < 10^{-5}$
Centered	0.59 ± 0.001	$p < 0.0001$	16.88 ± 0.023	$p < 0.0004$
Flicker	0.09 ± 0.001	$p < 0.0001$	24.01 ± 0.057	$p < 0.05$

the location of k^{th} maximum is inhibited as well as its neighborhood. Its size in pixels is 30 (corresponding to $0.6° \times 0.6°$).
The cumulative probability is the sum of the predicted saliency included in a circle of 30 pixels of radius, centered on the most important experimental saliency locations. For the i^{th} picture, the cumulative probability C^i is given by

$$C_P^i = \sum_{k=1}^{N} \sum_{l=-r}^{r} \sum_{m=-r}^{r} P^i(x^k - l, y^k - m) \tag{14}$$

where, N is the number of the most important fixation points and r the radius of the circle. The subscript P indicates that the computation refers

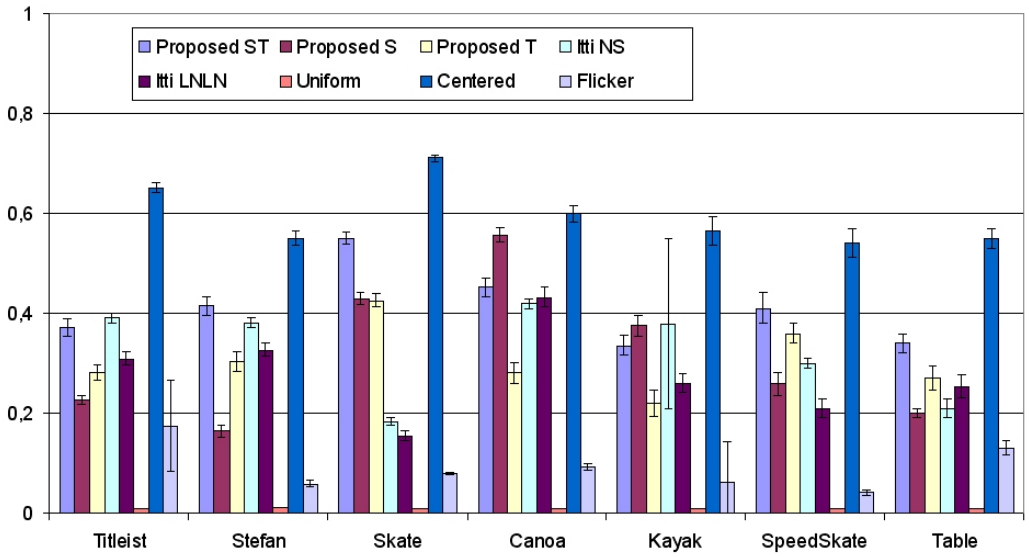


Fig. 9. Average correlation coefficient per clip. Error bars depicts the 95% confidence interval. Notation is as in Table 2.

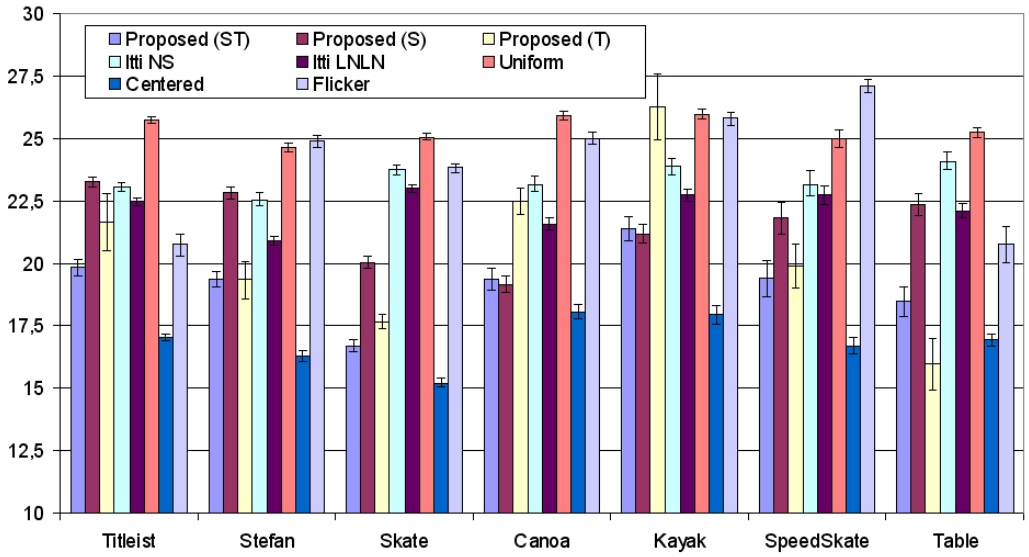


Fig. 10. Average KL-divergence per clip. Error bars depicts the 95% confidence interval. Notation is as in Table 2.

to the predictions. The overall predicted cumulative probability noted C_P is given by

$$C_P = \frac{1}{M} \sum_{k=1}^{M} C_P^i \tag{15}$$

Table 3
Comparison among models regarding the cumulative probability. The upper-bound, called **Human**, is the cumulative probability calculated from the experimental maps. * means that results coming from the proposed ST model are statistically better than those stemming from Itti's model (t-test, $p < 0.05$ or better). Notations are as in Table 2.

Model	$N = 5$	$N = 10$	$N = 15$	$N = 20$
	$mean \pm sem$	$mean \pm sem$	$mean \pm sem$	$mean \pm sem$
Human (upper-bound)	0.33 ± 0.008	0.55 ± 0.001	0.73 ± 0.007	0.81 ± 0.007
Proposed ST	$0.09 \pm 0.004^*$	$0.15 \pm 0.006^*$	$0.19 \pm 0.007^*$	$0.23 \pm 0.008^*$
Proposed S	0.06 ± 0.002	0.11 ± 0.003	0.15 ± 0.004	0.18 ± 0.005
Proposed T	0.09 ± 0.007	0.15 ± 0.01	0.20 ± 0.01	0.24 ± 0.01
L. Itti LNLN	0.06 ± 0.003	0.10 ± 0.004	0.14 ± 0.004	0.17 ± 0.005
Uniform	0.03 ± 0.000	0.06 ± 0.000	0.08 ± 0.000	0.10 ± 0.001
Centered	0.10 ± 0.003	0.18 ± 0.004	0.24 ± 0.004	0.27 ± 0.005
Flicker	0.05 ± 0.002	0.09 ± 0.004	0.12 ± 0.005	0.14 ± 0.005

where, M is the frame number.
A value of 1 indicates a perfect correlation between the experimental and the predicted saliency maps, whereas a value of 0 indicates an anti-correlation. Based on this metric, a lower and upper bound is also defined. The former is simply obtained by the use of an uniform saliency map, meaning that each pixel has the same probability to get fixated. This is the worst case. The upper-bound is obtained when the cumulative probability is not extracted from the predicted probability density function but rather in the experimental probability density function.

As previously and as expected, the uniform model yields the worst results, as shown in Table 3. In all tested cases, proposed computational models ST and T perform better than both proposed model S and Itti's model. The flicker model still remains a bad predictor of priority maps.
It is interesting to note the high difference between the upper-bound and the performances of the tested models. For example, if twenty points are considered, there is a ratio of 3.5 (0.81/0.24) between the upper-bound and the best biological model. There are at least three explanations: the eye tracking experiments have been conducted in a free-viewing task. However, it is impossible to prove that there is no top-down influence. Moreover, the extent to which the top-down mechanism influences the allocation of attention is impossible to quantify. One thing is certain: top-down influences can be very strong, directing the gaze to a particular location, irrespectively of the salience. The

second reason is related to the reaction time required to process the visual information following a scene cut. All subjects involved in the eye movement experiments were new to the content of the video clips. Therefore, each scene cut occurring in the clip induces a temporal masking lasting several frames. During these periods, the fixation immediately following the scene cut depends mainly on where observers gazed prior the cut. This is probably not the most important effect but its contribution cannot be neglected on clip containing numerous scene cuts. For instance, the sequence *Titleist* contains twelve scene cuts. Assuming the temporal masking lasts up to 100 ms on average, there are only 4% [4] of the results that are not relevant. The last plausible explanation concerns the intrinsic difference that exists between the experimental (see figure 3) and the predicted saliency maps (see figure 5). The former is sparse while the latter is more uniformly distributed. A post treatment of the prediction can probably reduce the gap.

4.4 The particular case of the centered model

Whereas biological models perform much better than the uniform and the flicker models, it is not the case anymore when the centered model is considered. It clearly yields the best performances for all metrics. The performances of the centered model reach 0.59 and 16.88, while the best biological model hardly yields 0.41 and 19.21, for *cc* and *kl*, respectively. What are the reasons of these performance differences? The general tendency for observers to fixate near the center of scenes, even if the salience is null could be a potential explanation. This effect is called the central bias. This tendency is due to a number of reasons notably detailed in (Tatler et al., 2005). The most important reason probably lies in the film makers tendency to place subjects of interest near the center. Therefore, attending to central regions represents an efficient information selection strategy, unconsciously deployed by observers. In addition, this tendency has been likely and unfortunately reinforced during the eye tracking experiments. Indeed each trial began with a centrally located fixation marker. This marker could be randomly positioned. However, studies which did not use a central fixation marker (Canosa, 2003) have also shown a central fixation bias.
As *cc*, *kl* and the cumulative probability metrics are sensitive to the dynamic range and to the salience distribution, it is also possible that these metrics were strongly influenced by this bias. Differences between experimental and predicted distributions are important. On one hand, the experimental saliency distribution is very sparse, due to non-spatially uniform distribution of saliency in natural scenes. On the other hand, predicted saliency distributions are rather smooth. It is important not to play down the influence of this point. In

4 $\frac{12cuts \times 100ms \times 25Hz}{741}$, 741 is the total frame number.

order to unravel the situation, a fourth metric, less sensitive to the dynamic range, is used.

4.5 ROC analysis

Receiver Operating Characteristic analysis (ROC analysis) consists in labeling (by thresholding) the predicted and the priority maps. In this study, all locations of saliency maps are labelled as fixated or not. Twenty values are used to threshold the predicted saliency maps. These thresholds are uniformly selected between the minimum and the maximum values of the predicted data. Only one value of threshold has been used to segment the priority map. The salience distribution of the experimental data is very sparse and the different salience values between a fixated and a non-fixated area is very high (see examples of priority maps in figure 3). It is for this reason that only one threshold has been used. ROC curves were obtained by varying the threshold and comparing the resulting pixel labels with the ground truth. Curves shown on figure 11, indicate the false alarm rate (labeling a non-fixated locations as fixated) as a function of the hit rate (labeling fixated locations as fixated). The more the top left-hand corner the curve approaches, the better the detection: the ideal discrimination is obtained by a false positive rate equal to 0 and a true positive range equal to 1. Figure 11 shows the ROC analysis results when con-

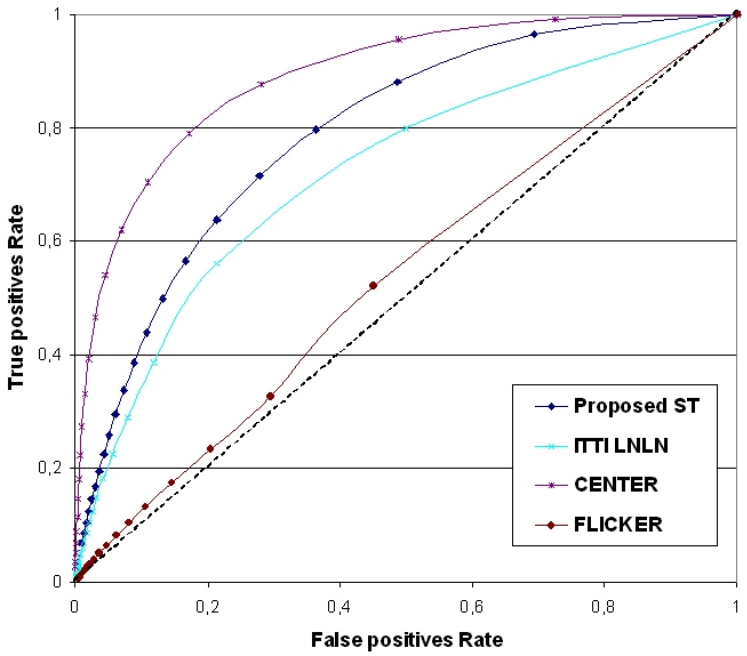


Fig. 11. Results of the ROC analysis on the set of sequences, for different models (results for **Centered**, **Proposed ST**, **Uniform** model and Itti's model ($LNLN$) are presented).

sidering the overall clips. As previously, the proposed spatio-temporal model outperforms all tested models except the centered model. The ROC analysis is very different from the previously tested metrics. Nevertheless, the hierarchy in terms of performances remains the same.
Whatever the metric, the centered model systematically outperforms the others models. It follows that the four metrics are probably not confounded by the central bias. The results just reflect the similarity degree that exists between the predicted and the priority maps. The central bias is likely to be the result of the interaction between two factors. Firstly, regions of interest are often located near the center of the screen, by film makers and photographers. Secondly, the central fixation marker that has been used during the eye tracking experiments could also share some responsibilities.

5 Discussion

The most important aspects of this study were both to quantify the agreement between experimental and predicted saliency maps and to benchmark several models on dynamic complex scenes. Except the particular case of the centered model, the new computational model of the bottom-up visual attention proposed in this paper outperforms flicker, uniform and L. Itti's models. It should be noted that this study does not offer an exhaustive survey of models predictions. Other models might yield better results.
Several side-results, detailed below, are obtained, reinforcing several previous findings and giving more substances to others.

Strong and persistent role of low-level visual features
Results described in this study first indicate that attentional allocation is strongly influenced by the low-level visual features during free-viewing of dynamics color scenes. This study confirms and reinforces the findings of L. Itti (Itti, 2005).

The best predictor of human fixation integrates all visual features
Whatever the metric, sensitive or not to the dynamic range of salience, the best predictor of human saccade targets is the model that incorporates all features. Again, these results reinforces previous findings.
A number of issues are brought in by the fact that several visual dimensions are considered (achromatic, chromatic and temporal). The main issue concerns the building of a unique saliency map: how to combine the different saliency maps, coming from different modalities and potentially having different dynamic ranges? The simple approach consists in summing all maps after that a suitable scaling has been applied on. Such approach presents several drawbacks. To cope with these issues, a new way of combining saliency maps is proposed, leading to a statistically significant gain, in terms of KL-divergence and ROC

analysis (Table 2 and figure 11). As the fusion issue is likely to be a key point in the visual attention modeling, further explorations are scheduled. In a future study, the spatial dimension will be split into color and luminance in order to quantify the extent to which the chromatic and achromatic features contribute to the attentional allocation.

The influence of images features does not change during viewing

It was previously deemed that stimulus dependence was not constant over viewing time and greatest just after stimulus onset. This is why numerous studies assess the similarity degree between experimental and predicted fixations either for short viewing durations (when the eye tracking experiments involve still pictures (Parkhurst et al., 2002; Jost et al., 2005)) or just after a scene cut for video clip (Carmi & Itti, 2006).

Assuming that the bottom-up influences are maximum just after the stimulus onset and decrease with viewing time, the similarity degree between prediction and experimental data should be maximum after the stimulus onset and should decrease over time. These were the conclusions of previous studies (Parkhurst et al., 2002; Jost et al., 2005). Nevertheless, as emphasized by B. Tatler (Tatler et al., 2005), the effect observed by Parkhurst et al. was probably due to an artifact of their methodology. Moreover, in a previous study dealing with still color pictures (Le Meur et al., 2006), the similarity degree for different viewing times (4, 10 and 14 seconds) was evaluated. The viewing duration was deliberately long, assuming that the stimulus dependence is almost constant over time. The performances were roughly the same, whatever the viewing duration (in fact, the performances increased because of observers continuing to gaze salient areas throughout the trial rather than to scan the whole picture). It indicated that the bottom-up influence still remains important over time. In other words, in a free-viewing task, attentional allocation was continuously and strongly driven by the low-level visual features. These previous results are coherent with the findings of B. Tatler (Tatler et al., 2005). Indeed, Tatler et al. have shown that *fixation location consistency changes between observers over time but the influence of image features does not*. This means that top-down mechanism can strongly influence the attentional allocation, leading to idiosyncratic patterns of eye movements. However, as soon as the top-down mechanism vanishes, bottom-up influences become dominant again, drawing the attention towards the most salient locations.

In conclusion, two key points have to be considered to compute the similarity degree that exists between saliency and priority maps: the initial fixation, prior the stimulus onset, is important, as emphasized in (Tatler et al., 2005). The second point concerns the temporal masking due to the stimulus onset or due to a scene cut. Temporal masking induces a significant reaction time and significantly influences salience measures (if the measure is done just after the scene cut). However, this aspect is often overlooked.

How to lessen the central bias?
The central bias influence is the most crippling aspect of this work. As previously described, the center bias has several origins and is likely reinforced in our experiments by the central position of the marker used just before stimulus onset.
In order to maximize agreement between human eye movements and biological plausible model predictions, a new protocol has to be defined. But, is it possible to design an eye tracking protocol that significantly lessens this bias? Several simple rules have to be followed. For instance, it is recommended to randomly position the marker before the stimulus onset. In addition, the selection of the video clips is important. Others rules could be defined to further reduce central bias effect, notably, on the relative alignment of the screen and video clip centers. This issue will be addressed in a future study.

References

Canosa, R. L., Pelz, J.B., Mennie, N.R., & Peak J. (2003). *High-level aspects of oculomotor control during viewing of natural-task images*, Proc. SPIE Human Vision and Electronic Imaging VIII (HVEI'03), vol. 5007, pp.240-251.

Carmi, R. & Itti, L. (2006). *Causal Saliency Effects During Natural Vision*, In: Proc. ACM Eye Tracking Research and Applications, pp. 11-18.

Cover, T. M., & Thomas, J. A. (1983). *Elements of Information Theory*, New York: Wiley.

Deco, G., & Schurmann, B. (2000). *A hierarchical neural system with attentional top-down enhancement of the spatial resolution for object recognition*, Vision Res, 40(20), pp. 2845-2859.

Fecteau, J.H., & Munoz, D.P., (2006). *Salience, relevance, and firing: a priority map for target selection*, Trends in Cognitive Sciences, 10(8), pp. 617-631.

Fencsik, D. E., Urrea, J., Place, S. S., Wolfe, J. M., & Horowitz, T., S.(2005). *Velocity cues improve visual search and multiple object tracking*, Visual Cognition, 14, pp. 92-95.

Findlay, J. M., (1997). *Saccade target selection during visual search*, Vision Research, 37, pp. 617-631.

Hoffman, J. E., & Subramanian, B. (1995). *Saccadic eye movement and visual selective attention*, Percept. Psychophys., 57, pp. 787-795.

Hoffman, J. E.(1998). *Visual attention and eye movement*, ed. by H. Pashler, Psychology press, 57, pp. 119-154.

Itti, L., Koch, C., & Niebur, E. (1998). *A model of saliency-based visual attention for rapid scene analysis*, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI), 20(11), pp. 1254-1259.

Itti, L., & Koch, C. (2001). *Feature Combination Strategies for Saliency-Based Visual Attention Systems*, Journal of Electronic Imaging, Vol. 10, No. 1, pp.

161-169.

Itti, L. (2005). *Quantifying the Contribution of Low-Level Saliency to Human Eye Movements in Dynamic Scenes*, Visual Cognition, Vol. 14, No. 4-8, pp. 959-984.

Ivry, R., & Cohen, A. (1992). *Asymmetry in visual search for targets defined by differences in movement speed*, Journal of Experimental Psychology: Human Perception and Performance, 18, pp. 1045-1057.

Jost, T., Ouerhani, N., Wartburg, R., Mri, R., & Hgli, H. (2005). *Assessing the contribution of color in visual attention*, Computer Vision and Image Understanding Journal (CVIU), Special Issue on Attention and Performance in Computer Vision, Elsevier, Vol 100 Issues 1-2.

Koch, C., & Ullman, S. (1985). *Shifts in selection in visual attention: towards the underlying neural circuitry*, Human Neurobiology, 4(4), pp. 219-27.

Krauskopf J., and Williams D. R., & Heeley D. W. (1982). *Cardinal direction of color space*, Vision Research, 22, pp. 1123-1131.

Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2005). *A spatio-temporal model of the selective human visual attention*, proceedings of the IEEE International Conference on Image Processing, pp. 1188-1191, Genoa, Italia.

Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). *A coherent computational approach to model the bottom-up visual attention*, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI), 28(5), pp. 802-817.

Mack, M., Castelhano, M. S., Henderson, J. M., & Oliva, A. (2003). *What the Visual System see: the Relationship between Fixation Positions and Image Properties during a Search Task in Real-Wolrd Scenes*, OPAM Annual workshop, Vancouver.

Maioli, C., Benaglio, I., Siri, S., Sosta, K. & Cappa, S. (2001). *The integration of parallel and serial processing mechanisms in visual search: evidence from eye movement recordings*, EUR. Journal Neuroscience, 13,pp. 364-372.

Milanese, R. (1993). *Detecting salient regions in an image: from biological evidence to computer implementation*, Ph.D, University of Geneva, Switzerland.

Odobez, J.M. & Bouthemy, P. (1995). *Robust multiresolution estimation of parametric motion models*, Journal of Visual Communication and Image Representation, 6(4), pp. 348-365.

Olshausen, B.A., Anderson C.H.,& Van Essen D.C. (1993). *A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information*, The Journal of Neuroscience, 13(11), pp. 4700-4719.

Parkhurst, D. J., Law, K., & Niebur, E. (2002). *Modeling the role of salience in the allocation of overt visual attention*, Vision Research 42, pp. 107-123.

Parkhusrt, D. J., & Niebur, E. (2003). *Scene contant selected by active vision*, Spatial Vision, vol. 16, pp.125-154.

Rajashekar, U., Cormack, L., & Bovik, A. (2003). *Image features that draw fixations*, proceedings of the IEEE International Conference on Image Processing, Barcelona, Spain.

Reinagel, P., & Zador, A.M. (1999). *Natural scene statistics at the centre of*

gaze, Network: Computational Neural Systems, 10, pp. 1-10.

Salvucci, D. D., & Goldberg, J. H. (2000). *Identifying fixations and saccades in eye-tracking protocols*, Proceedings of the Eye Tracking Research and Applications Symposium, pp. 71-78.

Sen, T., & Megaw, T. (1984). *The effects of task variables and prolonged performance on saccadic eye movement parameters*, In A. G. Gale & F. Johnson (Eds.) Theoretical and Applied Aspects of Eye Movement Research, pp. 103-111.

Serences, J.T., & Yantis, S. (2006). *Selective visual attention and perceptual coherence*, Trends in Cognitive Sciences, 10, pp. 38-45.

Seyler, A. J., & Budrikis, Z.L. (1959). *Measurements of temporal adaptation to spatial detail vision*, Nature Rev Neuroscience, 184, pp. 1215-1217.

Seyler, A. J., & Budrikis, Z.L. (1965). *Details perception after scene changes in television image presentations*, IEEE Trans. Inform. Theory, 11(1), pp. 31-43.

Tam, W.J. (1995). *Visual masking at video scene cuts*, SPIE Human Vision and Electronic Imaging, 2411, pp. 111-119.

Tatler, B. W., Baddeley, R. J., & Gichrist, I. D. (2005). *Visual correlates of eye movements: Effects of scale and time*, Vision Research, 45 (5), pp. 643-659.

Treisman, A. M., & Gelade, G. (1980). *A feature-integration theory of attention*, Cognit. Psychol., vol. 12(1), pp. 97-136.

Tsotsos, J. K. (1990). *Analysing vision at the complexity level*, Behavioral and Brain Sciences 13(3), pp. 423-445.

Tsotsos, J. K., & Liu, Y. Martinez-Trujillo J. C.,& Pomplum, M., & Simine, E., Zhou, K. (2005). *Attending to visual motion*, Computer Vision and Image Understanding, 100, 3-40.

Wandell, B. (1995). *Foundations of vision*, Sunderland, Massachusetts:sinauer Associates.

Wolfe, J. M. (1998). *Visual Search*, in Pashler, H. editor, Attention, pp. 13-74, Psychology Press.