Modelling spatio-temporal saliency to predict gaze direction for short videos

Sophie Marat, Tien Ho Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin, Anne Guérin-Dugué

MODELLING SPATIO-TEMPORAL SALIENCY TO PREDICT GAZE DIRECTION FOR SHORT VIDEOS

S. Marat, T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin, A. Guérin-Dugué

GIPSA-Lab/Department Images-Signal
BP 46, 38402 Grenoble Cedex, FRANCE
phone: +334-76-574355, fax: +334-76-574790, email: sophie.marat@gipsa-lab.inpg.fr

The original publication is available at www.springerlink.com
DOI 10.1007/s11263-009-0215-3

ABSTRACT

This paper presents a spatio-temporal saliency model that predicts eye movement during video free viewing. This model is inspired by the biology of the first steps of the human visual system. The model extracts two signals from video stream corresponding to the two main outputs of the retina: parvocellular and magnocellular. Then, both signals are split into elementary feature maps by cortical-like filters. These feature maps are used to form two saliency maps: a static and a dynamic one. These maps are then fused into a spatio-temporal saliency map. The model is evaluated by comparing the salient areas of each frame predicted by the spatio-temporal saliency map to the eye positions of different subjects during a free video viewing experiment with a large database (17000 frames). In parallel, the static and the dynamic pathways are analyzed to understand what is more or less salient and for what type of videos our model is a good or a poor predictor of eye movement.

keywords: Saliency, Spatio-temporal model, Gaze prediction, Video viewing

1. INTRODUCTION

Usually, people do not look at every object in the visual field but concentrate on some salient regions. In the visual field, the spatial regions which attract attention, and therefore the eyes, are usually called salient. The emerging problem is how to design a model that puts salient areas in conspicuous locations. The answer relates to modeling human visual attention with saliency maps; this has been of growing interest to many researchers for the last few decades. The saliency of a spatial location depends mainly on two factors: one is task-independent and the other is task-dependent. The first one is often called bottom-up and is mainly driven by low-level processes depending on the intrinsic features of the visual stimuli. The latter refers to top-down processes. It is more complex to model because it integrates high-level processes (task, cognitive state...) [1].

Most computational models of visual attention are bottom-up and are inspired by the concept of Feature Integration Theory (FIT) of Treisman and Gelade [2]. The first model was described by Koch and Ullman [3]; like most of the models, it concentrates on spatial image features such as color, contrast, orientation... Several models [4], [5] are inspired by this theory; the most popular is the one proposed by L. Itti et al. [6] and it has become a reference for all research on saliency. Motion feature has been added to this model more recently [7] and to others [8], [9] to create saliency models for videos.

Following a similar approach, a spatio-temporal bottom-up saliency model is proposed. This model differs from existing ones on several points:

- The model of the two outputs of the retina which provides two different signals: using a retina model, the signal processed by the static pathway differs from the one processed by the dynamic pathway. The useful information is separated to provide more efficient signals to both pathways [10].
- The compensation of the camera motion: using camera motion compensation, we detect only the areas that move against the background. Not only are moving areas detected, but we define a motion contrast map by estimating the module of the motion for each pixel.
- The method of fusion of static and dynamic saliency maps: a new fusion to combine the static and the dynamic pathway outputs is proposed. This fusion modulates the different saliency maps with adaptive coefficients for each frame. These coefficients were chosen by analyzing simple statistics (mean, maximum and skewness) on both outputs. We classified the videos using these statistics and we analyzed in detail to what extent low-level descriptors may contribute to the guidance of eye movement.

For this model, we only concentrated on some basic features: signal orientations and spatial frequencies for the static saliency, and the module of motion for the dynamic saliency. We chose to concentrate only on these basic features and not to add color, stereo or other features, first, because the chosen features are predominant for saliency and, second, because we wanted to understand better how these features are correlated with human eye movement and how to combine them to create a spatio-temporal saliency map. Other features could be added in further research.

The proposed model is described in section 2. Section 3 presents an experiment that records the eye movements of fifteen people looking at a large number of videos (17000 frames). In section 4, an evaluation of the proposed model is drawn, and after a detailed analysis of the static and dynamic pathways a new fusion method is presented to combine both outputs to create a spatio-temporal saliency map.

2. MODEL

The proposed model is inspired by the first steps of the human visual system, from the retina cells to the complex cells of the primary visual cortex. The visual information goes through the retina preprocessing to the cortical-like filter de-

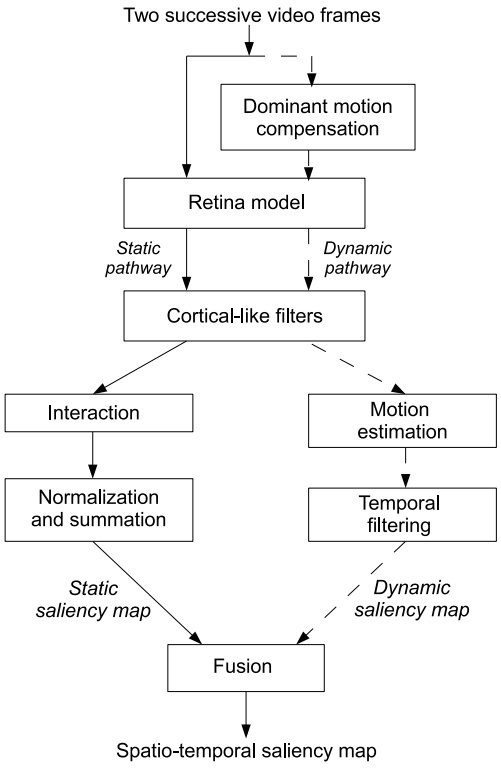


Figure 1: Schema of the proposed spatio-temporal saliency model

composition [10],[11]. The retina extracts two signals from each frame that correspond to the two main outputs of the retina [12]. Each signal is then decomposed into elementary features by a bank of cortical-like filters [13]. These filters are used to extract both static and dynamic information, according to their frequency selectivity, providing two saliency maps: a static and a dynamic one. Both saliency maps are combined to obtain a master spatio-temporal saliency map per video frame (Fig. 1). This map predicts the gaze direction to particular areas of the frame analyzed.

2.1 Retina model

The retina, which has been described in detail in [14], [12], [15], is composed of different neural layers. The flow of information goes from the photoreceptors to the horizontal cells that provide a local average of the incoming information. The bipolar cells take the difference of the outputs of the photoreceptors and the horizontal cells. Amacrine cells provide a second local average of the bipolar cells output.

The retina has two outputs formed by different ganglion cells: parvocellular output and magnocellular output. Parvocellular output provides detailed information which can be simulated by extracting the high spatial frequencies of an image. This output enhances frame contrast, which attracts human gaze in static frame [16]. Magnocellular output responds rapidly and provides global information which can be simulated by using lower spatial frequencies. The proposed model (Fig. 2) decomposes the input frame into different frequency bands: a high spatial frequency one to provide a "parvocellular-like" output and a lower spatial frequency

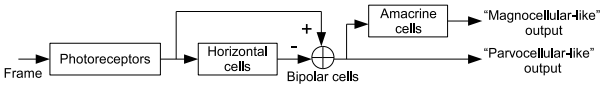


Figure 2: Retina model

one to simulate the "magnocellular-like" output (Fig. 3).

2.1.1 The retina "parvocellular-like" output

First, luminance coming from the real word is captured by the photoreceptors which act as a low-pass filter with a high cut-off frequency. Horizontal cells play the role of a low-pass filter of the photoreceptor's output and are modeled by a gaussian filter.

Bipolar cells calculate the difference of the outputs of the photoreceptors and the horizontal cells, which corresponds to a high-pass filtering of the frame. Bipolar cells *On* retain the positive part of this difference while bipolar cells *Off* retain the absolute value of the negative part:

$$bipolar\ On = max\{0, y-h\}$$
$$bipolar\ Off = max\{0, h-y\}$$

The output of the bipolar cells is given by the difference of the bipolar cells *On* and the bipolar cells *Off*:

$$bipolar\ cells = bipolar\ On - bipolar\ Off$$

The output of the ganglion cells, formed by the bipolar cells, is used to model the parvocellular output of the retina. Therefore, the "parvocellular-like" output reveals frame contrast and helps to whiten its spectrum. This output is the first stage of the static pathway of the model (Fig. 3(c)).

2.1.2 The retina "magnocellular-like" output

Human beings see stable and moving components in a moving scene effortlessly. An object tracked by the camera is seen as moving even if it is stationary on the frames. We assume that visual attention is attracted by motion contrast and we define it as the motion of regions against background. The first step, before the retina filter, is the compensation of the background motion to estimate the relative motion of regions against background.

Background is supposed to represent more than half of the frame's pixels. In this case, background motion is also called dominant motion and is computed using the 2D motion estimation algorithm developed in [17]. This algorithm provides dominant motion compensation between two successive frames by carrying out a robust multi-resolution estimation of an affine parametric motion model. The parametric model chosen here is an affine one with 6 parameters:

$$\begin{cases} v_x = a_1 + a_2.x + a_3.y \\ v_y = a_4 + a_5.x + a_6.y \end{cases}$$

where $(a_1,...,a_6)$ are the estimated parameters and v_x and v_y are the vectorial components of the dominant motion computed at position (x,y) using the previous parameters.

After the camera motion compensation, the two frames (the current frame and the next compensated frame) go through the retina filter. The bipolar cells calculate the difference between the photoreceptors and the horizontal cells'

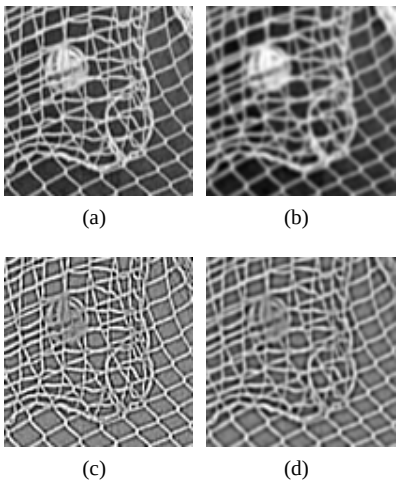


Figure 3: Retina model: a) Input image, b) Horizontal cell response, c) "Parvocellular-like" output, d) "Magnocellular-like" output.

outputs. This difference acts as a high pass filter that whitens the energy spectrum of the frame. Then, the amacrine cells act as a low-pass filter that eliminates high frequencies (gaussian filter). The resulting equivalent filter is a band-pass filter. This output corresponds to the Magnocellular output of the retina and is the first stage of the dynamic pathway of the model (Fig. 3(d)).

2.2 Cortical-like filters

Visual information is decomposed into different spatial frequencies, orientations, colors and motion in the primary visual cortex (V1) [18],[13],[19]. In this model, we choose to not study color information and Gabor filters are used to model V1 cells to extract frequencies, orientations and motion information. These filters are a good compromise of resolution between the frequential and spatial domains. Each filter G_{ij} (Eq.1), at orientation i and at frequency j, is determined by its central radial frequency f_j and its standard deviations σ_{ij}^{θ} and σ_{ij}^{f} in orientation θ_j and its orthogonal orientation, respectively $i = 1,..,N_\theta$, $j = 1,..,N_f$ and $\frac{f_j}{f_{j-1}} = 2$ with $f_{N_f} = 0.25$. We chose $\sigma_{ij}^{\theta} = \sigma_{ij}^{f}$, which is justified in the next section.

The number of orientations and frequencies were respectively fixed at $N_\theta = 6$ and $N_f = 4$, for the static pathway according to preliminary experiments (Fig. 4). For the dynamic pathway, the spatial resolution is lower; so only the three low frequency bands were used (f_1, f_2 and f_3.)

$$G_{ij}(u,v) = exp\left\{-\left(\frac{(u'-f_j)^2}{2\sigma_{ij}^{f\,2}} + \frac{v'^2}{2\sigma_{ij}^{\theta\,2}}\right)\right\} \qquad (1)$$

with:

$$\begin{cases} u' = ucos(\theta_i) + vsin(\theta_i) \\ v' = vcos(\theta_i) - usin(\theta_i) \end{cases}$$

The output of each filter corresponds to an intermediate map m_{ij}. These maps are the equivalent of some of the elementary features of Treisman's Theory [2].

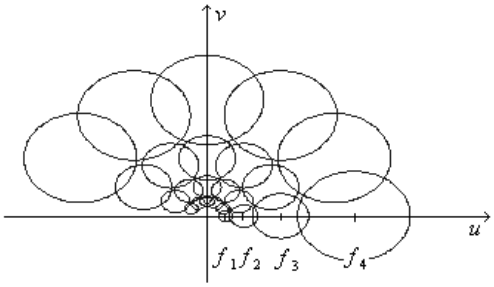


Figure 4: Configuration of Gabor filters in the frequential domain: 6 orientations and 4 frequency bands.

2.3 The static pathway

2.3.1 Interactions between filters

Neuron responses in the primary visual cortex are influenced by other neurons as far as excitation and inhibition are concerned. We considered two types of interactions based on the range of the receptive fields. Short interactions reinforce objects belonging to a specific orientation while long interactions are used for contour facilitation [20].

Short interactions introduce inhibition between neurons of neighboring orientations and overlapping receptive fields. For the standard deviations of the cortical-like filters, if $\sigma_{ij}^{\theta} > \sigma_{ij}^{f}$ it is more orientation-selective but reduces the inhibitive interaction. So, we chose $\sigma_{ij}^{\theta} = \sigma_{ij}^{f}$. Short interactions occur with the same pixel in different intermediate maps m_{ij}. Each pixel is excited by similar pixels in the other maps of the same orientation but different frequencies and inhibited by those of different orientations but similar frequency (Fig. 5).

The second interaction type is long range interaction which occurs among collinear neurons beyond the receptive fields. This type of interaction is worked out in each intermediate map by convolution with a "butterfly" mask [20]. This mask (Fig. 6) consists of an excitory part in the corresponding orientation of the intermediate map m_{ij} and an inhibitive part in other orientations. It was normalized in such a way that its summation was equal to one. The mask size is inversely proportional to the frequency of the corresponding intermediate map m_{ij}.

2.3.2 Normalization and summation

A region is salient if it is different from its neighbors. Thus, to strengthen the intermediate maps that have spatially distributed maxima, the method proposed by Itti [6] is used. After being normalized in $[0,1]$, each map m_{ij} was multiplied by $(max(m_{ij}) - \overline{m}_{ij})^2$ where $max(m_{ij})$ and $\overline{m}_{ij}$ are its maximum and average respectively. Then, all values in each map that were smaller than 20% of its maximum were set to 0.

Finally, all intermediate maps were added together to obtain a static saliency map $M_s(x,y,k)$ for each frame k (Fig. 7(a), 7(b)).

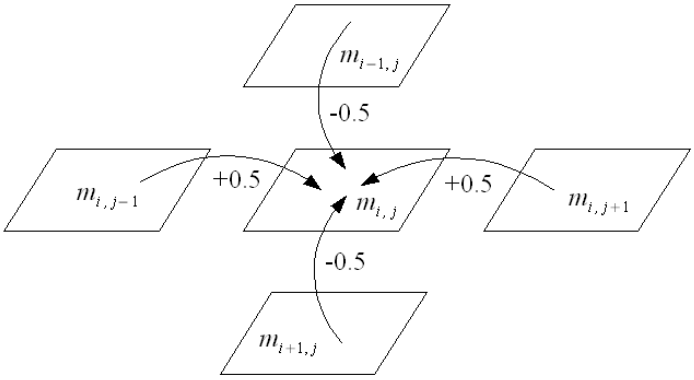


Figure 5: Short interactions for intermediate maps m_{ij}.

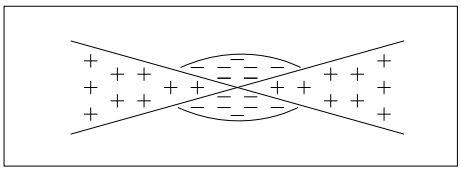


Figure 6: Butterfly mask used for long interactions.

2.4 The dynamic pathway

Dynamic saliency is linked to motion and particularly to the motion of a region against the background. The speed of moving region against background was computed using a motion estimator on compensated frames at the "Magnocellular-like" output of the retina.

2.4.1 Motion estimation

A differential approach, described in detail in [21], was used. It relies on the assumption of luminance constancy.

The motion at location *(x,y)* in frame *t* is given by vector *V(x,y,t)* which satisfies the optical flow constraint equation (Eq. 2)

$$\nabla I(x,y,t).V(x,y,t) + \frac{\partial I(x,y,t)}{\partial t} = 0 \qquad (2)$$

where *I(x,y,t)* is the luminance of the pixel at the position *(x,y)* in the frame *t*.

For each frame, the optical flow constraint was applied to each output of the cortical-like filters, with the same radial frequency, leading to an over-determined system of equations allowing the aperture problem to be overcome. For each pixel *(x,y)*, a motion vector (v_x, v_y) was computed, solving the system (Eq. 3) with a least square estimation using Biweight Tuckey's function.

$$\begin{bmatrix} \Omega_1^x & \Omega_1^y \\ \vdots & \vdots \\ \Omega_{N_\theta}^x & \Omega_{N_\theta}^y \end{bmatrix} \cdot \begin{bmatrix} v_x \\ v_y \end{bmatrix} = - \begin{bmatrix} \Omega_1^t \\ \vdots \\ \Omega_{N_\theta}^t \end{bmatrix} \qquad (3)$$

where $\Omega_i^p = \frac{\partial (I * G_{i.})}{\partial p}$, $G_{i.}$ is one of the cortical-like filter at the orientation *i*, and *I* is the "Magnocellular-like" output of the retina. The optical flow constraint creates an accurate estimation only for low motion. A robust multiresolution scheme is needed to estimate a large scale of speed. A first approximation of motion was done with frames at coarse resolution to estimate fast motion; this displacement was then compensated and the residual motion was estimated at a finer resolution.

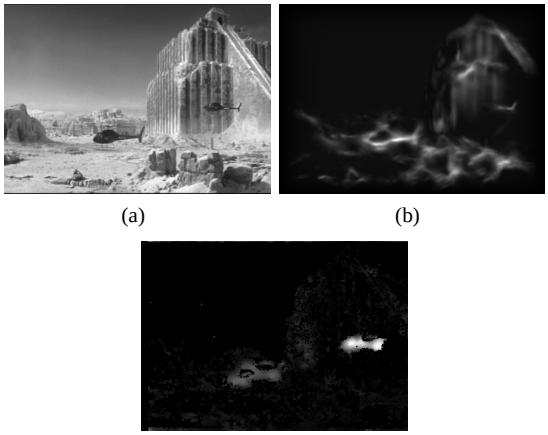


Figure 7: Example of a natural scene (a) with its static saliency map M_s (b) and its dynamic saliency map M_d (c).

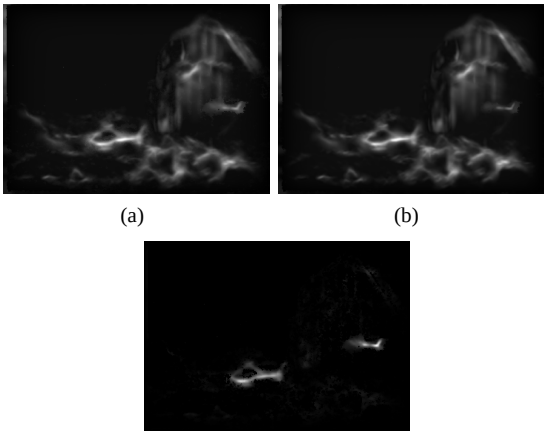
A motion vector was defined (per pixel) by its module, corresponding to the speed, and its angle, corresponding to the motion direction. As we assume the motion saliency map of a region is linked to its speed against background, we only used the module of this motion vector to define the dynamic saliency of the area.

2.4.2 Temporal filtering

A temporal median filter was applied to remove noise. If a pixel had a motion in one frame but not in the previous ones, it is most probably noise resulting from the motion estimation. This temporal filter was applied on five successive frames (the current frame and the four previous ones) and the filter was reinitialized after each shot cut to avoid artifacts. A dynamic saliency map $M_d(x,y,k)$ was obtained for each frame *k* (Fig. 7(a), 7(c)).

2.5 Fusion

The saliency maps obtained at the outputs of the static and the dynamic pathways do not have the same range of values. To carry out the fusion, the raw saliency information, without normalization, was retained to take advantage of this difference and to promote the more accurate saliency map. However, the range of values of the static saliency map and the dynamic one is compatible. Neither of these two kinds of maps had systematically outperforming values. For the static saliency map the normalization was done on the intermediate maps $m_{i,j}$, the maximum of static saliency values were around 1.7 and could go up to 2.7. The dynamic saliency maps, had to respect the theorem of shannon in the temporal domain, the maximum speed would be limited, considering the frame rate. The maximum of dynamic saliency values was around 2.7 and could reach 9.



(a) (b)

(c)

Figure 8: Examples of fused saliency maps: (a) M_{mean}, (b) M_{max} and (c) M_{and} of a natural scene (Fig. 7(a)). The NSS are respectively 1.07, 0.67 and 2.20.

Three different fusions were proposed:

- a *mean* fusion, taking the pixel average of the two saliency maps:
  $M_{mean} = \frac{M_s + M_d}{2}$
- a *max* fusion, taking for each pixel the maximum of the two saliency maps:
  $M_{max} = Max(M_s, M_d)$
- a pixel by pixel multiplicative fusion corresponding to a logical *and*:
  $M_{and} = M_s \times M_d$

Examples of these fusions are given in Fig. 8. The M_{mean} fusion modulates one map with the other. If an area is salient for the static map but not for the dynamic one, the fusion saliency is lower than it was in the static one. For the M_{max} fusion, an area has the highest saliency between static and dynamic maps and is less selective. In the M_{and} there are small salient regions, which could be spread all over the saliency map. This multiplicative fusion is the most selective one. In this case, an area needs to be salient simultaneously in the static and the dynamic maps to be salient in the fused map. The usually used M_{mean} [22] and M_{max} give saliency maps that are close. For the latter, results are presented with the M_{and} fusion only as it gives the best results experimentally.

3. EXPERIMENT

The goal of this part is to compare the results given by our model to the human eye position density map obtained through an eye movement experiment. Both bottom-up and top-down influence eye position. The bottom-up saliency model proposed is used to quantify the contribution of low-level saliency to eye movements. To better prevent top-down processes, we did not use classical videos but instead we used small clips, as was done by Itti [23]. The aim is to remove all semantic content from videos as far as possible. For that, videos were split into small clips and these clips, from different video sources, were put together [24].

Participants:
Fifteen human observers (3 women and 12 men, aged from 23 to 40 years old). All participants had normal or corrected to normal vision, and were not aware of the purpose of the experiment. They were asked to look at videos without any particular task.

Apparatus and experimental design:
Eye tracking was performed by an Eyetracker Eyelink II (SR Research). During the experiment participants were sitting with their chin supported in front of a $21''$ color monitor (75 Hz refresh rate) at a viewing distance of 57 cm ($40^\circ \times 30^\circ$ usable field of view). A 9-point calibration was carried out every five stimuli and a control drift was done before each stimuli.

Stimuli:
This experiment is inspired by an experiment of Carmi and Itti [23]. Fifty-three videos (25 fps, 720×576 pixels/frames) were selected from heterogeneous sources including movies, TV shows, TV news, animated movies, commercials, sport, music clips. These 53 videos gathered indoor, out-door, day-time and night-time sources. The 53 videos were cut every 1-3 seconds (1.86 ± 0.61) into *324 clip snippets*. The length of the clip snippet was chosen randomly, the only constraint was to obtain a snippet without any shot cuts. These clip snippets were then strung to form *20 clips* of 30 seconds (30.20 ± 0.81). Each clip contained at most one clip snippet from each continuous source. The choice of the clip snippets and their duration were random to prevent subjects anticipating shot cuts. As the proposed model is bottom-up, clip snippets were used to minimize potential top-down influence on eye movements. Stimuli (17000 frames) were presented on gray level without audio as the model did not consider color and audio information.

Human eye position density maps:
We recorded and analyzed the eye positions. The eyetracker records the eye position at 500 Hz. The eyetracker records 20 eye positions per frame for the two eyes. The median of all these positions was taken (with X-axis median and Y-axis median) for each subject and for each frame. For each frame the points of all the subjects were gathered. Then we applied a 2D gaussian function to each point to obtain the human eye position density map, $M_h(x,y,k)$. The standard deviation of the gaussian was chosen to have a diameter at half the height of the gaussian equal to 0.5° of visual angle.

4. RESULTS

We analyzed the eye positions rather than the fixation points for two reasons. First, we had more data when choosing all the eye positions, and so we could extract one point per frame and per subject. Second, in most of the cases, eye positions and fixations are very close except during smooth pursuit. So by retaining all the eye positions we obtained data even during smooth pursuit. The aim is to compare the salient areas given by the model, with the fixated areas.

Various criteria have been proposed and used for this comparison: the Kullback-Leibler distance [25] or the Receiver Operator Curve (ROC) [26] (others examples can be found in [23], [8], [27]). In this paper, we chose to focus on the correlation coefficient and the Normalized Scanpath Saliency (NSS) [28]. This last criteria was especially designed to study eye movement data and so, the corresponding results can be easily interpreted. The correlation coefficient and the NSS lead to the same conclusion on the data analysis. However, the correlation coefficient is very dependent on the

standard deviation of the gaussian applied to gaze positions to compute the human eye position density map. The NSS was therefore preferred. The NSS criteria is a Z-score, (also called standard score). This Z-score expresses the divergence of the experimental result from the model mean as a number of standard deviations of the model. The larger the value of Z, the less probable it is that the experimental result is due to chance. The NSS is computed using the equation:

$$NSS(k) = \frac{\overline{M_h(x,y,k) \times M_m(x,y,k)} - \overline{M_m(x,y,k)}}{\sigma_{M_m(x,y,k)}} \quad (4)$$

where $M_h(x,y,k)$ is the human eye position density map normalized to obtain unit mean, and $M_m(x,y,k)$ the model saliency map. The model saliency map can be the static, dynamic or fused maps. This is equivalent to normalize each saliency map to have a zero mean and a unit standard deviation and to retain, on the normalized saliency map, the value corresponding to gaze locations of subjects and then to average the retained values over subjects. If the mean of the saliency values at eye position is equal to the mean of saliency value on the whole frame (the NSS is null), there is no link between eye position and saliency. If the mean of the saliency values at eye position is lower than the mean of saliency values on the whole frame (the NSS is negative), eye position tends to be on non-salient regions. If the mean of the saliency values at eye position is higher than the mean of saliency values on the whole frame (the NSS is positive), eye position tends to be on salient regions.

4.1 Global analysis

The NSS was computed for each frame of every clip (17000 frames). In order to test if our model is a good predictor of human eye movements, the mean NSS value is calculated using the model saliency map (static, dynamic, fusion of both) and the experimental data. To compare our model we used two sets of experimental data. For the first set, we associated to each frame of a clip snippet the corresponding eye movement of subjects when they were looking at the videos. This first set is called the real eye movements, in opposition to the second set called the partially randomized eye movements. For this second set, we associated to a frame the eye movement of subjects when they were looking at another clip snippet. We only kept the frame position inside a clip snippet. This second comparison is to ensure that our model predicts the salient areas of specific frame and not simply predicts subjects eye movements without any correlation with the content of the frame. If our model is a good predictor of salient areas and because the NSS value was defined to compare computational saliency map with eye movements, we should observe low values of NSS when comparing our model with partially randomized eye movements and high NSS values when comparing our model with real eye movements. Using partially randomized eye movements also prevents the effect of central bias of eye position on model evaluation [26]. Partially randomized eye movements were obtained using subject eye position and not using random sampling: the same bias as real eye movement was then observed. Thus, the difference of NSS between real eye movement and partially randomized eye movement is not due to the fact that subjects are more likely to stare at the center of the image than to look at the image randomly. The mean NSS value is given for three models of saliency maps (static, dynamic and the fusion of both) in comparison with real eye movements and the partially randomized eye movements, in Table 1.

Saliency maps	M_s	M_d	M_{and}
Real eye movements	0.68	0.87	0.96
Partially randomized eye movements	0.33	0.14	0.14

Table 1: Mean NSS value on all the clips for the three models of saliency maps (static, dynamic and multiplicative fusion)

Comparing M_{and} with real eye movement gives the best results. The mean of the saliency values at the eye position was around one standard deviation away from the mean of saliency values on the whole frame. As expected, NSS values are higher when comparing the three models with real eye movements than when comparing with partially randomized eye movements ($F(1,84968)=10497.07$; $p\equiv 0$).

If we analyze what happened with the partially randomized eye movements, the mean NSS value for the static saliency map is more than twice the mean NSS value for the dynamic saliency map. This can be explained by the fact that, the static and the dynamic saliency maps have different appearance. For most of the frames, the static saliency map highlights areas spread over the whole frame. In fact all, the frames represent naturalistic scenes with textured area. On the contrary, the dynamic saliency map can exhibit small and compact areas corresponding to moving objects. By so doing, random eye position is, on average over all frames, more likely to be on a salient region in the static saliency map than in the dynamic saliency map. If all the subjects had the same eye movement pattern, the results may be more closely linked to this pattern than to the actual saliency on the videos. By testing our model with partially randomized eye movements, we can see that there is no such common eye scan path, and that the high mean NSS is caused by the relevance of the saliency model and not a plausible strategy of eye position during video viewing.

Moreover the dynamic pathway gives better results than the static one ($F(1,28328)=275.40$; $p\equiv 0$). The fusion gives the best result ($F(2,42482)=276.06$; $p\equiv 0$), both pathways are needed to obtain improved results.

The model is also compared to simple heuristics as it is usually done in the literature [23], [16]. This comparison tests if our model is more accurate in predicting saliency than a simple model only based on low level image descriptors such as luminance. Two static naive heuristics and one dynamic naive heuristic were tested. The two static naive heuristics were the entropy H and the standard deviation SD of pixels' luminance. The image was split into patches of 16×16 pixels and the entropy or standard deviation was computed on each patch. This value was propagated on the corresponding pixels of the patch to form an entropy saliency map M_{snH} and a luminance standard deviation saliency map M_{snSD}. A gaussian filter was applied to each map to spread the patch border effect. The dynamic naive heuristic M_{dn} was the absolute difference of the pixels of two consecutive frames. This difference highlights the moving pixels. Note that no dominant motion compensation was done before. In fact, we just wanted to compare our model with a simple naive heuristic and so without using motion 2D [17] which is an elaborate preprocessing.

The mean NSS value is given for the saliency map obtained for the three simple heuristics in comparison with real eye movements, in Table 2. The proposed fused saliency model (mean $NSS = 0.96$ (Table 1)) gives more accurate saliency prediction than the simple heuristics tested (F(3,56658)=1046.64; p≡0).

Saliency map	M_{snH}	M_{snSD}	M_{dn}
Real eye movements	0.54	0.44	0.54

Table 2: Mean NSS value on all the clips for two static naive heuristics (entropy H and standard deviation SD of pixels' luminance) and one dynamic naive heuristic (absolute difference of the pixels of two consecutive frames)

All these observations were in agreement with our expectation and means that our model can predict real eye movements. Research had shown that motion is the feature that attracts the human gaze the most. This has been shown experimentally but also using saliency models [23], [29]. As expected, the NSS value is higher when comparing dynamic saliency to static. However, we need to take into account both pathways to obtained the highest NSS value.

4.2 Temporal analysis

The previous results are an average over all frames. It can be interesting to see if there is an evolution of the NSS value during clip snippets. Such a study was carried out previously for static images [27], [30] and for videos [8], [23]. Because the proposed model is a bottom-up one, it can only predict human gaze for the first fixations of subjects when looking at a static image or for the first few frames when looking at a video. Figure 9 presents the NSS value as a function of frame position inside a clip snippet for the static saliency maps M_s, the dynamic ones M_d and the fusion of both M_{and} with real eye movements. We also present the results for partially randomized eye movements for the static pathway $M_{s'}$ and for the dynamic one $M_{d'}$. The curves with partially randomized eye movements do not have the same shape as the curves with real eye movements, as the first ones hardly vary with time. These curves present very low NSS values that are typical of no correspondence between the model and the experimental data. All the other curves obtained for real eye movements have the same shape. The maximum NSS value is reached for all the curves (Fig. 9) at about the 13^{th} frame, which corresponds to 520 ms, then curves decrease slowly. The shape of these curves can be explained by the fact that, at the beginning, only bottom-up influences occurred, followed by top-down processes. Studies have found that bottom-up influences act faster than top-down processes [1], [31]. If this is the case, we should observe that bottom-up influences occur before top-down processes. In [30], [23] saliency effects were stronger just after the stimuli onset than later on, while in [26] no saliency dependencies on time were found. Our results are in accordance with [30], [23].

We analyze the bottom-up influence with another indicator: the dispersion of eye position between subjects. This dispersion is plotted for all the clip snippets (Fig. 10). The dispersion D is defined by:

$$D = \frac{1}{N^2} \sum_{i,j<i} d_{i,j}^2$$

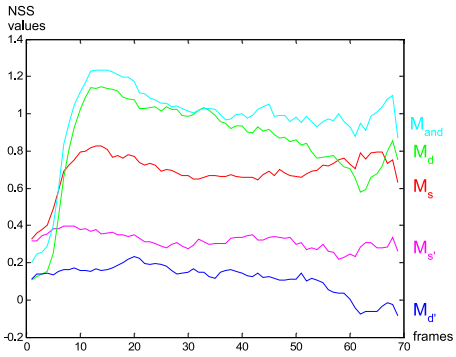


Figure 9: NSS as a function of frame. NSS is averaged on 324 clip snippets for different saliency maps: static M_s, dynamic M_d, fusion of both M_{and}, static with partially randomized eye movements $M_{s'}$, dynamic with partially randomized eye movements $M_{d'}$

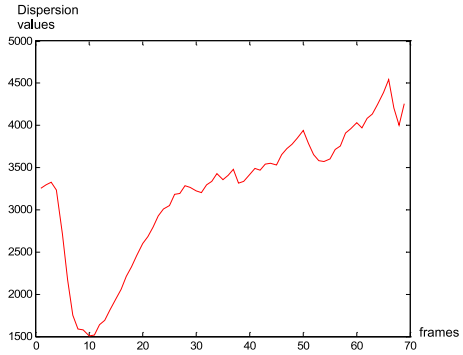


Figure 10: Dispersion D of eye positions as a function of frame. The dispersion is averaged on 324 clip snippets

where N is the number of subjects, $d_{i,j}$ is the distance between the eye position of subject i and j. A low value of dispersion corresponds to close fixations by subjects. Low values of dispersion are more probably explained by bottom-up influences: the saliency is given by the intrinsic features of the stimuli and is the same for all observers. On the other hand, the top-down processes involve the cognitive state and the prior knowledge of each subject, and tend to be different for different subjects. The dispersion curve is in accordance with the NSS curve. With bottom-up influence, subjects look at salient regions and the dispersion of their gaze decreases as NSS increases. Top-down processes occur too, and dispersion increases as subjects gaze at different regions, NSS decreases.

During the first few frames, NSS is low and dispersion is high. This can be explained by the fact that after each shot cut, the gaze stays at the previous position during a few frames and then moves to a salient region. We can assume that there is some time shifting between the time a region, present on the screen, is salient and the time this region is fixed by subjects. Figure 11 shows NSS as a function of time for the dynamic saliency map without and with an offset, i.e. a shift of three frames (=120 ms) between the saliency map

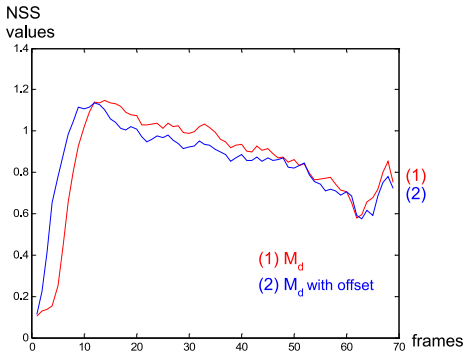


Figure 11: NSS as a function of frame for the dynamic saliency map with and without offset.

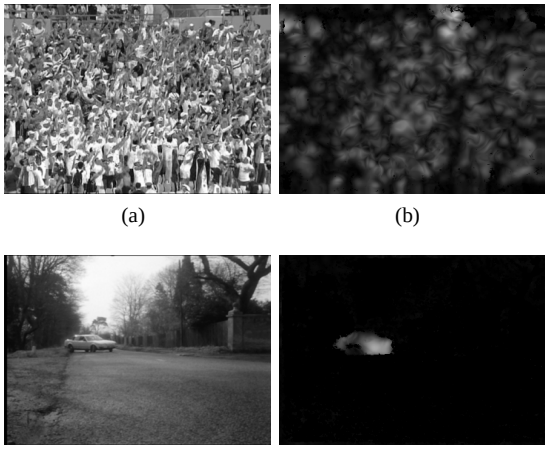


Figure 12: Examples of a natural scene (a) with a dynamic saliency map (b) with low skewness (1.00) and a natural scene (c) with a dynamic saliency map (d) with high skewness (8.62)

and the fixation considered (the NSS of the first frame of a clip snippet was computed considering the third fixation frame). With the offset, the curve is shifted to the left, which means a greater correspondence between the human fixations and the saliency model earlier in the clip. This result was expected as we know there is a saccade latency around 150 ms.

The duration of preponderance of bottom-up influences on static images usually reach 150 to 300 ms after stimulus onset [32], [33]. If we take into account the offset, the first maximum of the NSS curve corresponds to the first minimum of the dispersion curve and is about 8 frames; this corresponds to 320 ms which is above the usual reported time of 150ms for static images. However, video stimuli are used instead of still images, and to our knowledge, no time values for the predominance of bottom-up influences using video stimuli were previously reported.

The evolution of the NSS value as a function of the frame position inside a snippet fits well with the fact that (1) our model is a bottom up model so it can only predict eye movements for the first frames, and (2) the dispersion of the eye positions of subjects increases with frame position (Fig. 10).

4.3 Detailed analysis of the two pathways

The proposed model is a good predictor of the eye movements of subjects when looking freely at clip snippets. However, this model is a better predictor for the first frames of a snippet. It can be interesting to inquire now what is more attractive in the static saliency map and what is more attractive in the dynamic saliency map.

As we said before, the static and the dynamic saliency maps do not have the same appearance. On one side, the static saliency map exhibits a large number of salient areas, corresponding to textured areas, which are spread over the whole image. On the other side, the dynamic saliency map can exhibit only small and compact areas corresponding to moving objects. Concerning the question "what is more salient in the static and the dynamic pathways?" we can suppose:

- *for the static map*: a frame would be salient if its static saliency map has a high value and not if its static saliency is spread. The saliency of a frame would be correlated with the maximum of its corresponding static saliency map.
- *for the dynamic map*: a frame would be salient if its dynamic saliency map has small and compact areas. The saliency of the frame would be linked to the number and the size of the salient areas in its corresponding dynamic saliency map.

To test these two hypotheses, we introduced two statistics: the maximum and the skewness of the saliency map [24]. The maximum is characteristic of the static map. The skewness is characteristic of the dynamic saliency map. They should indicate the saliency prediction efficiency of the map.

In fact, we observed that a dynamic saliency map with a high skewness corresponds, in general, to a map with only small and compact areas. Whereas a dynamic saliency map with a low skewness value corresponds to a map with spread salient areas, or to a map with a small salient area but with different motion values (different gray levels). These observations correspond to the fact that skewness is a measure of the degree of asymmetry of a distribution; here, the distribution of the gray values of dynamic saliency map pixels is considered. If a frame contains a small moving area, its dynamic saliency map would exhibit only a small area. Its distribution would have a high and sharp peak close to zero and another peak, smaller and wider, close to the motion value; this would induce a mean greater than the mode and so, a high skewness. On the other hand, for a frame with a lot of moving areas, the dynamic saliency map distribution would still have a peak close to zero but would have another more spread out peak around the motion value (more spread than in the previous example); this would reduce the asymmetry and so decrease the skewness. The skewness is also decreased in the case of a small area with a diffuse motion. These examples are illustrated in figure 12.

Using these two statistics, on the static saliency map, in one part, and on the dynamic saliency map, in a second part, we classified our videos into four groups: 1)high skewness high maximum, 2) high skewness low maximum, 3) low skewness high maximum, 4) low skewness low maximum. Each snippet is labeled 1) 2) 3) or 4) on static saliency

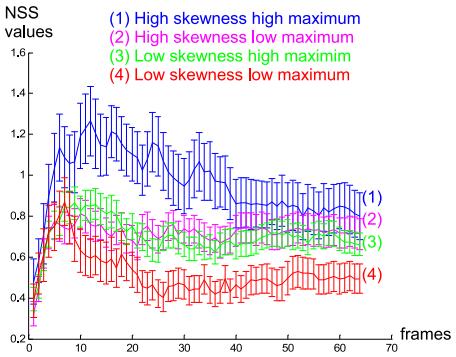


Figure 13: NSS as a function of frame for clip snippets categorized using maximum and skewness for static saliency map

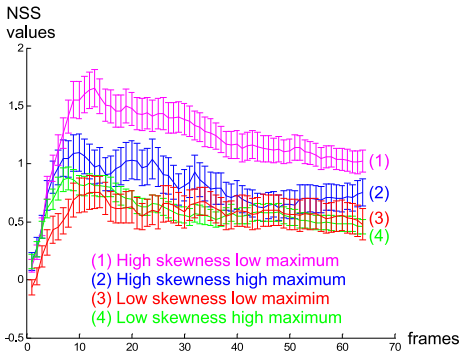


Figure 14: NSS as a function of frame for clip snippets categorized using maximum and skewness for dynamic saliency map

map (Fig. 13) and on the dynamic one (Fig. 14). The NSS is plotted as a function of frame for the four categories of snippets. If our hypotheses are verified, we should observe the highest NSS values for the videos with a high maximum value for the static saliency map and the highest NSS values for the videos with a high skewness value for the dynamic saliency map. The static pathway is more predictive for snippets with high maximum in static saliency (maximum of the curve above 0.8). The main information is given by the maximum saliency; if a frame has a high maximum saliency value, there is an attractive region in this frame. On the other hand, if the frame has a low maximum saliency the most attractive region is less attractive than in the previous frame (Fig. 15). The dynamic pathway is more predictive for snippets with higher skewness (maximum above 1.1). The dynamic map $M_d(x,y,k)$ gives motion information, but now the salient regions may be small (Fig. 12 a, c). If there is only a small moving region, the saliency must be concentrated on this region. If there is only a dynamic salient region, the gaze of all the subjects would be concentrated there. However if there are several regions with equivalent dynamic saliency, subjects' gaze would be spread over these different regions. The fact that NSS is higher with lower motion (low maximum) can be explained by the fact that if the speed of the moving region is too high it is difficult to track it [34].

A fusion taking advantage of the characteristics of the static and the dynamic saliency maps is then proposed (Fig. 16):

$$M_{skew-max} = \alpha M_s + \beta M_d + \gamma M_s \times M_d$$

with:

$$\begin{cases} \alpha = max(M_s) \\ \beta = skewness(M_d) \\ \gamma = max(M_s).skewness(M_d) \end{cases}$$

The static pathway is modulated by its maximum value α. The dynamic saliency map is modulated by its skewness value β. The reinforcement term γ gives more importance to the areas that are salient both in a static and dynamic way, and so to the small moving region with high static saliency. This fusion has a mean NSS=1.01 and is significantly above the M_{and} fusion (mean $NSS = 0.96$ (Table 1)) (F(1,28308)=10.54; p=0.0012).

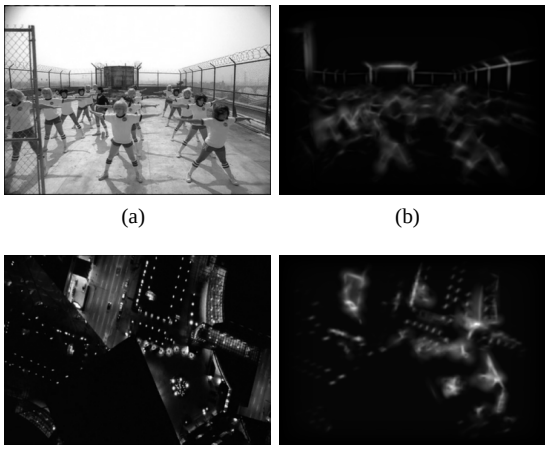


Figure 15: Examples of a natural scene (a) with a static saliency map (b) with low maximum (1.53) and a natural scene (c) with a static saliency map (d) with high maximum (2.03)

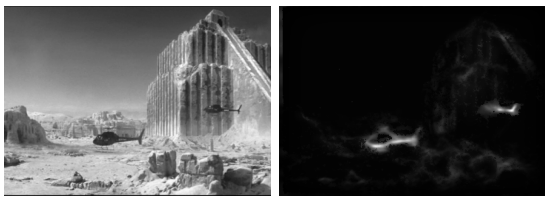


Figure 16: Example of a natural scene (a) with its saliency map $M_{skew-max}$ (b)

5. CONCLUSION

This study presents a new bottom-up saliency model inspired by the biology of the first steps of the human visual system. The model presents a simulation of the two pathways (magnocellular and parvocellular) of the human visual system based on their main known properties. These two pathways can be seen as static and dynamic pathways. A video, coming through the model, is split into spatial and motion information. This split starts with the two main outputs of the retina and continues with the cortical cells, sensitive to different spatial frequencies for the different pathways. At the output of each pathway, a saliency map is extracted. The static and the dynamic saliency maps are fused to create a spatio-temporal saliency map. The model associates a saliency map to each frame of a video. This map is used to predict the areas that would be gazed at by people when looking at the videos.

In order to test the exactness of the proposed model, we ran an experiment to record eye movements of subjects when looking freely a large base of videos. We did not use "classical" videos, but instead, inspired by Itti's experiment, we used small clip snippets [23]. This allowed us to compare experimental data with our model, which is a bottom-up one. In fact, the model can predict the gaze of people for the first few frames of a video. Different comparisons to validate the model were carried out. First, the output of the model was compared with subject's gaze and with partially random gazes. We also compared the model with other simple heuristics. Each comparison showed that our model is significantly better than the others.

Moreover, we studied the evolution of the model prediction as a function of time (for frames inside each small clip snippet). The model is more accurate at the beginning of a clip snippet. This expected result can be explained by two facts: (1) our model is a bottom up model, which means that it can only predict the fixations of people for the first frames of a clip, (2) for the first frames all the subjects were looking at the same locations (the dispersion between their fixation points was weak) but they were looking at different locations at the end of a clip snippet (higher dispersion).

We showed that the dynamic map is more predictive than the static one. However the fusion of both maps gives the best results. We also showed that the static saliency covers spread areas on a frame, on the contrary, the dynamic saliency map gives motion information on areas that can be small. If there was only a small moving region, the saliency would be concentrated on this region. The maximum is then more relevant for the analysis at the static saliency maps and the skewness is more relevant for dynamic saliency maps. These characteristics are then used to compute a new fusion $M_{skew-max}$ taking advantage of both pathways. Saliency maps are modulated with maximum and skewness information and a reinforcement term that gives more importance to the areas that are salient both in statically and dynamically, so to the areas with localized moving region with high static saliency.

In this model, we chose to concentrate only on basic features that are predominant for static saliency. The efficiency of these features has been shown and we understood the importance of each feature better. In future work it would be interesting to add more features such as color or a spatially varying sampling of the retina depending on eye positions to reinforce our model. The model could also be used to improve video compression or be added to camera motion analysis [35] to help select frames for a summary of the video.

REFERENCES

[1] J. M. Henderson, "Human gaze control during real-world scene perception," *Trends in cognitive sciences*, vol. 7, pp. 498-504, 2003.

[2] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97-136, 1980.

[3] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219-227, 1985.

[4] O. Le Meur, P. Le Callet, and D. Barba, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. on PAMI*, vol. 28, pp. 802-817, 2006.

[5] J. K. Tsotsos, S. M. Culhane, Y. K. W. Winky, Y. Lai, N. Davis and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, pp. 507-545, 1995.

[6] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. on PAMI*, vol. 20, pp. 1254-1259, 1998.

[7] R. J. Peters and L. Itti, "Applying computational tools to predict gaze direction in interactive visual environments," *ACM Trans. on Applied Perception*, vol. 5, 2008.

[8] O. Le Meur, P. Le Callet and D. Barba,"Predicting visual fixations on video based on low-level visual features," *Vision Research*, vol. 47, pp. 2483-2498, 2007.

[9] Y.-F. Ma, X. Hua and H. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. on multimedia*, vol. 7, 2005.

[10] S. E. Palmer, "Vision science: Photons to phenomenology," *MIT Press*, Cambridge, 1^{st}, 1999.

[11] C. Massot and J. Hérault, "Model of frequency analysis in the visual cortex and the shape from texture problem," *IJCV*, vol. 76, pp. 165-182, 2008.

[12] W. H. Beaudot, "The neural information in the vertebra retina: a melting pot of ideas for artificial vision," *PHD thesis, Tirf laboratory*, Grenoble, France, 1994.

[13] R. L. DeValois, "Orientation and spatial frequency selectivity: properties and modular organization," *In A. Valberg and B. B. Lee (Eds). From Pigment to perception*, New York: Plenum, 1991.

[14] W. H. A. Beaudot and P. Palagi, J. Hérault, "Realistic simulation tool for early visual processing including space, time and colour data," *IWANN, in LNCS*, vol. 686, pp. 370-375, Springer-Verlag, Barcelona, June 1993.

[15] S. H. Schwartz, "Visual perception: a clinical orientation," *McGraw-Hill*, New-York, 3^{rd}, 2004.

[16] P. Reinagel, and A. Zador, "Natural scene statistics at the center of gaze," *Network: Computation in Neural Systems*, vol. 10, pp. 341-350, 1999.

[17] J.-M. Odobez and P. Bouthemy, "Robust multiresolution estimation of parametric motion models," *Journal of visual communication and image representation*, vol. 6, pp. 348-365, 1995.

[18] D. H. Hubel and T. N. Wiesel, “Functional architecture of macaque visual cortex,” *Proc. of the Royal Society of London*,B, 198, pp. 1-59, 1977.

[19] J. G. Daugman, “Two-dimensional spectral analysis of cortical receptive field profiles,” *Vision Research*,vol. 20, pp. 847-856, 1980.

[20] T. Hansen, W. Sepp, and H. Neumann, “Recurrent long-range interactions in early vision,” *Emergent Neural Computational Architectures Based on Neuroscience*, LNCS/LNAI 2036, pp. 139-153, 2001.

[21] E. Bruno and D. Pellerin, “Robust motion estimation using spatial Gabor-like filters,” *Signal Processing*, vol. 82, pp. 297-309, 2002.

[22] R. Milanese, H. Wechsler, S. Gil, J.-M. Bost and T. Pun, “Integration of bottom-up and top-down cues for visual attention using non-linear relaxation,” *Proc. CVPR*, pp. 781-785, 1994.

[23] R. Carmi and L. Itti, “Visual causes versus correlates of attentional selection in dynamic scenes,” *Vision Research*, vol. 46, pp. 44333-4345, 2006.

[24] S. Marat, T. Ho Phuoc, N. Guyader, D. Pellerin, A. Guérin-Dugué, “Spatio-temporal saliency model to predict eye movements in video free viewing,” *EUSIPCO’08 - 16th European Signal Processing Conference*, Lausanne, Switzerland,2008.

[25] U. Rajashekar, L. K. Cormack and A. C. Bovik, “Point of gaze analysis reveals visual search strategies,” *Human vision and electronic imaging IX 2004, Proc. of SPIE*, vol. 5292, pp. 296-306, 2004.

[26] B. W. Tatler, R. J. Baddeley and I. D. Gilchrist, “Visual correlates of fixation selection: effects of scale and time,” *Vision Research*, vol. 45, pp. 643-659, 2005.

[27] A. Torralba, A. Oliva, M. S. Castelhano and J. M. Henderson “Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search,” *Psychological Review*, vol. 113, pp. 766-786, 2006.

[28] R. J. Peters, A. Iyer, L. Itti and C. Koch “Components of bottom-up gaze allocation in natural images,” *Vision Research*, vol. 45, pp. 2397-2416, 2005.

[29] J. M. Wolfe, K. R. Cave and S. L. Franzel, “Guided search: an alternative to the feature integration model for visual search,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 15, pp. 419-433, 1989.

[30] 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.

[31] J. M. Wolfe, G. A. Alvarez and T. S. Horowitz, “Attention is fast but volition is slow,” *Nature*, vol. 406, pp. 691, 2000.

[32] J. K. Tsotsos, A. J. Rodríguez-Sánchez, A. L. Rothenstein and E. Simine, “The different stages of visual recognition need different attentional binding strategies,” *Brain Research*, vol. 1225, pp. 119-132, 2008.

[33] H.E. Egeth and S. Yantis, “Visual attention: control representation and time course,” *Annual review of psychology*, vol. 48, pp. 269-297, 1997.

[34] S. G. Lisberger, E. J. Morris and L. Tychsen, “Visual motion processing and sensory-motor integration for smooth pursuit eye movements,” *Ann. Rev. Neurosci.*, vol. 10, pp. 97-129, 1987.

[35] M. Guironnet, D. Pellerin, N. Guyader, P. Ladret, “Video summarization based on camera motion and a subjective evaluation method,” *EURASIP Journal on Image and Video Processing*, vol. 2007, Article ID 60245, 12 pages, 2007.