Spatio-temporal saliency model to predict eye movements in video free viewing

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

SPATIO-TEMPORAL SALIENCY MODEL TO PREDICT EYE MOVEMENTS IN VIDEO FREE VIEWING

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

Grenoble Images Parole Signal Automatique (GIPSA-Lab)/Images and Signal Department
46 avenue Felix Viallet - INPG, 38031, Grenoble, France
phone: + (33) 4 76 57 43 55, fax: + (33) 4 76 57 47 90, email: sophie.marat@gipsa-lab.inpg.fr
web: www.gipsa-lab.inpg.fr

ABSTRACT

This paper presents a spatio-temporal saliency model that predicts eye movements. This biologically inspired model separated a video frame into two signals corresponding to the two main outputs of the retina (parvocellular and magnocellular outputs). Both signals are then decomposed into elementary feature maps by cortical-like filters. These feature maps are then used to form two saliency maps: a static one and a dynamic one. These maps are fused into a spatio-temporal saliency map. The model is evaluated by comparing the salient areas of each frame predicted by these saliency maps (static, dynamic, spatio-temporal) to the eye positions of different subjects during a video free viewing experiment with a large database (17000 frames).

1. INTRODUCTION

Usually, people do not look at all objects in the visual field but concentrate on some salient regions. We call salient regions areas in the visual field that attract attention and so, the eyes. The emerging problem is how to design a model that puts in conspicuous locations salient areas. The answer relates to model human visual attention with saliency maps; this has been of interest to many researchers for the last few decades. The saliency of a spatial area depends mainly on two factors, one is task-independent and the other is task-dependent. The first one is often called the bottom-up and is mainly driven by low-level processes with the intrinsic features of the visual stimuli. The latter refers to top-down processes. It is more complex to model because it must integrate high-level processes (task, cognitive state...).

Most computational models of visual attention are bottom-up and are inspired by the concept of Feature Integration Theory (FIT) of Treisman and Gelade [1]. The first model was described by Koch and Ullman [2]. Most of the models concentrate on spatial features like color, contrast, orientation... The most popular one inspired by this architecture is the model proposed by L. Itti et al. [3] and has became a reference for saliency models. Motion feature has been added recently to this model [4] and to other models [5, 6] to obtain saliency models for videos.

This paper presents a new spatio-temporal saliency model close to biological knowledge.Visual information is decomposed into two signals: one signal carries spatial information of the visual scene and the other carries motion information. Retina and primary visual cortex are modeled according to the biological particularities of the static and dynamic pathways. This model is validated by an eye movement experiment with a large video database. The model is described in section 2. Section 3 presents an experiment that records eye movements of 15 people looking at videos and then, some evaluations of the proposed model are drawn.

2. MODEL

The proposed model (Fig.1) simulates a part of the human visual system. Visual information goes through the retina and then are processed by cortical-like filters. The static and dynamic pathways are modeled and further combined to obtain a master spatio-temporal saliency map per video frame. This map predicts the gaze direction to some particular areas of the frame analyzed.

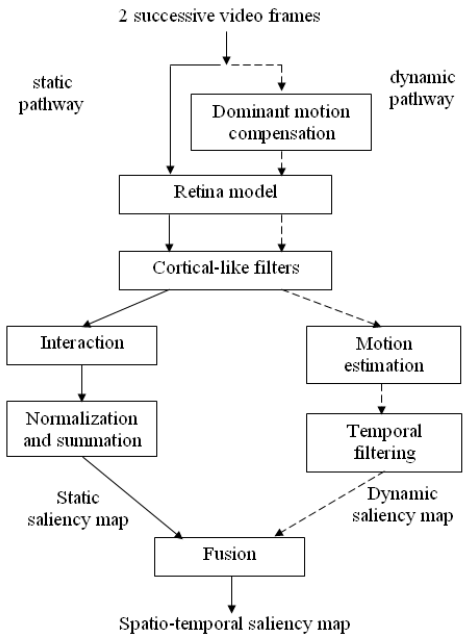

Figure 1: Spatio-temporal saliency model

2.1 Retina model

The retina, which has been described in detail in [7], is composed of different neuronal layers (Fig.2). The retina has two main outputs: (1) The parvocellular which gives precise information on analyzed regions and which is used to model the static pathway. (2) The magnocellular which responds rapidly and keeps lower spatial frequencies and which is used for the dynamic pathway. The flow of information goes through the photoreceptors then the horizontal cells calculate a local average of the coming information. The bipolar cells take the difference of the outputs of the photoreceptors and the horizontal cells. Amacrine cells calculate a second local average of the information.

2.1.1 The parvocellular pathway

In static frames, contrast attracts human gaze [8]. The retina filter enhances frame contrast and helps to find out salient areas. At first, photoreceptors enhance contrast through a nonlinear function (Eq.1) increasing luminance of dark regions without saturating the bright ones:

$$y = \frac{255 + x_o}{x + x_o} x \tag{1}$$

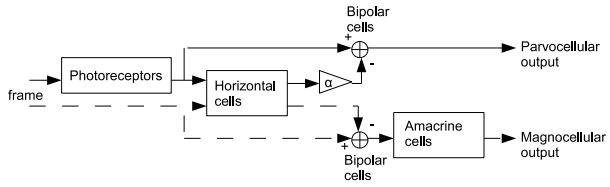

Figure 2: Retina model

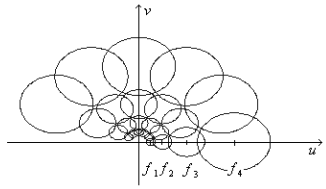

Figure 3: Configuration of Gabor filters: 6 orientations and 4 frequency bands.

where x is the luminance of the initial frame, x_o represents its average local luminance and y is the photoreceptor output.

Horizontal cells act as a low-pass filter of photoreceptor output and are modeled by a gaussian filter (σ=0.96). 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. In human visual perception, low frequencies precede high frequencies [9]. Low frequencies are added to the high frequency frame using an α coefficient (α=1/3). The Parvocellular pathway reveals frame contrast. This pathway is used to compute the static saliency map.

2.1.2 *The magnocellular pathway*

We assumed that visual attention is attracted by motion contrast and we defined it as the motion of regions against background. The first step, before the retina filter, is the compensation of the background motion to estimate motion of regions against background.

Background is supposed to represent more than half of the frame's pixels, in that case background motion is also called dominant motion and is computed using the 2D motion estimation algorithm developed in [10]. This algorithm provides dominant motion compensation between two sucessive frames by carrying out a robust multi-resolution estimation of an affine parametric motion model. Then the two frames (the current frame and the next compensated frame) go through the retina filter.

The magnocellular pathway responds rapidly but not precisely in the spatial domain, the nonlinear function is thus not necessary as it gives more details in a frame. The bipolar cells calculate the difference between the current frame and the output of the horizontal cells. The low-pass filter is then turned into a high-pass filter, which whitens the energy spectrum of the frame. Then the amacrine cells act as a low-pass filter which eliminates high frequencies (gaussian filter with σ=0.62). The resulting equivalent filter of the magnocellular pathway is a band-pass filter. This pathway is used to compute the dynamic saliency map.

2.2 Cortical-like filters

Visual information is processed in different frequencies, orientations, colors and motion in the primary visual cortex (V1) [11]. In this model, Gabor filters are used to model frequency and orientation processing in V1. These filters are a good compromise of resolution between the frequential and spatial domains. Each filter G_{ij} at the orientation i and at the frequency j is determined by its central radial frequency f_i and its standard deviations σ_{ij}^{θ} and σ_{ij}^{f} in orientation θ 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 numbers of orientations and frequencies were fixed at $N_\theta = 6$ and $N_f = 4$ respectively, for the static pathway, according to preliminary experiments. The output of each filter is an intermediate map m_{ij}. This map corresponds to the elementary feature of Treisman Theory [1].

For the dynamic pathway the spatial resolution is lower; so only the three low frequency bands are used (f_1, f_2 and f_3).

2.3 The static pathway for the visual attention model

2.3.1 *Interactions between filters*

Neuron responses in the primary visual cortex are influenced as far as excitation and inhibition are concerned by other neurons. We consider two types of interaction based on the range of the receptive fields [12].

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 interaction occurs with the same pixel in different intermediate maps m_{ij}. Each pixel is excited by the similar pixels in the other maps of same orientation but different frequency and suppressed by those of different orientations but similar frequency.

The second interaction type is long range interaction which occurs among collinear neurons beyond the receptive fields and is often used for contour facilitation [12]. This type of interaction is worked out in each intermediate map by convolution with a "butterfly" mask, an excitory part in the corresponding orientation and an inhibitive part in other orientations (having its summation equal to 1). The mask size was inversely proportional to the frequency of the corresponding intermediate map m_{ij}.

2.3.2 *Normalization and summation*

The intermediate maps m_{ij} were normalized to the same maximal value before summation. Moreover, an object is more salient if it is different from its neighbors. We used the method proposed by Itti [3] to strengthen the intermediate maps which had few maxima. After being normalized in $[0,1]$, each map m_{ij} was multiplied by $(m_{ij}^{*} - \overline{m}_{ij})^2$ where m_{ij}^{*}, $\overline{m}_{ij}$ are its maximum and average respectively. Then, all values in each map which were smaller than 20% of its maximum were set to 0.

Finally, all intermediate maps are added together to obtain the static saliency map $M_s(x,y,k)$ at each frame k (Fig.4).

2.4 The dynamic pathway for the visual attention model

Dynamic saliency is linked to motion and particularly to the motion of region against background. The speed of moving region against background was computed using a motion estimator on compensated frames.

2.4.1 *Motion estimation*

A differential approach, described in detail in [13], was used. It relies on the assumption of luminance constancy. The motion at location *(x,y)* in frame *t* is given by the vector *V(x,y,t)* which satisfies the optical flow constraint equation (Eq.2)

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

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

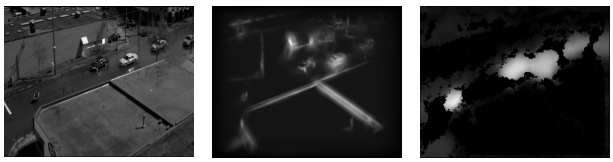

Figure 4: Examples of static M_s (middle) and dynamic M_d saliency maps (right) of a natural scene (left)

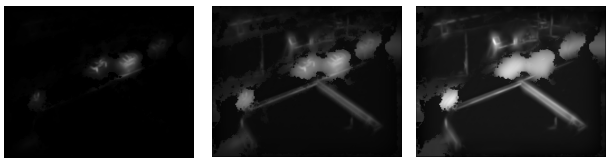

Figure 5: Examples of M_{and} (left), M_{moy} (middle) and M_{max} (right) of a natural scene (Fig.4 left)

For each frame, the optical flow constraint was applied to each output of 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).

$$\begin{bmatrix} \Omega_1^x & \Omega_1^y \\ \vdots & \vdots \\ \Omega_N^x & \Omega_N^y \end{bmatrix} \cdot \begin{bmatrix} v_x \\ v_y \end{bmatrix} = - \begin{bmatrix} \Omega_1^t \\ \vdots \\ \Omega_N^t \end{bmatrix} \quad (3)$$

with $\Omega_i^p = \frac{\partial(I * G_{i.,j})}{\partial p}$, $G_{i.,j}$ is the cortical-like filter at the orientation θ. Then, a robust multiresolution scheme allows the motion estimation from low frequencies (f_1) to higher (f_3) radial frequency to be obtained.

A motion vector was defined per pixel. As the motion saliency of a region is linked to its speed against background we used the module of this motion vector. (Higher the pixel is, brighter the pixel is).

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 filter was applied to a window of five frames (the current frame and the four previous ones) and the window is reinitialized after each shot cut to avoid artifacts. The dynamic saliency map $M_d(x,y,k)$ was then obtained for each frame k (Fig.4).

2.5 Fusion

The saliency maps obtained at the outputs of the static and dynamic pathways do not have the same range of value.

There are always textured and contrasted regions in frames, so there is always static saliency information. The static saliency maps are normalized between 0 and 255.

Considering that dynamic saliency is linked to the speed of moving regions against background, the normalization is done to be adaptive to motion. We assume that (1) the faster a region is, the higher its saliency is, and (2) if there is only one moving area the higher its saliency is. Motion previously seen is also taken into account: if the motion decreases, the saliency also decreases but if the motion is constant or increases, the saliency will remain at maximum value (= 255). Each dynamic map is normalized between 0 and the maximum value present on a temporal window of duration L=25 frames (24 previous and the current frame) i.e. 1 second of video. As videos are continuous in time, there is no abrupt change of speed in such a short portion of shot. The window was reinitialized after each shot cut, to avoid shot cut artifacts.

Three different fusions are proposed:

- a *mean* fusion, taking the average of each pixels 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.5. The 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 for this fusion. The *max* and *mean* fusions are less selective. The *mean* fusion modulates a 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 static. For the *max* fusion an area has the highest saliency between static and dynamic and is less selective.

3. EXPERIMENT AND RESULTS

The goal of this part is to compare the results given by our model to the human eye position density map obtained with an eye movement experiment.

3.1 Experiment

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:
Eyetracking 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 viewing distance of 57cm ($40° \times 30°$ usable field of view). A 9 point calibration was made every five stimuli and a control drift was done before each stimuli.

Stimuli:
This experiment is inspired by an experiment by Carmi and Itti [14]. Fifty three videos (25fps, 720x576 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) in *305 clip snippets*. The length of the clip snippet was chosen randomly, the only constraint was to obtain snippet without any shot cut. These clip snippets were then strung to form *20 clips* of 30 seconds (30.20 ± 0.81). Each clip contains at most one clip snippet of each continuous source. The choice of the clip snippets and their duration were random to prevent subject to anticipate shot cut. As the proposed model is bottom-up, clips 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 500Hz. The eyetracker records 20 eye positions per frame for the two eyes. The median of all these points was taken for each subject and for each frame. A point per frame and per subject was obtained and enabled the gaze to be tracked even during smooth pursuit movement which can occur during viewing video stimuli. For each frame the points of all the subjects were gathered. Then we added a 2D gaussian function to each point to obtain the human eye position density map, $M_h(x,y,k)$.

3.2 Results

We analyzed the eye positions rather than the fixation points for two reasons.
First we have much more data when choosing all the eye positions; and we can can have one point per frame and per subject.
Second in most of the cases, eye positions and fixations are very close except during smooth pursuit. This method allows us to obtain the eye positions even during smooth pursuit. In this section the contribution of the two pathways is analyzed, as well as different fusions.

After testing different criteria, results are presented using the Normalized Scanpath Saliency (NSS) [4]. The NSS is used to measure the correspondence between the saliency map computed by a model and the human eye position map. It is computed by:

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

with $M_h(x,y,k)$ is the human eye position map and $M_m(x,y,k)$ the model saliency map ($m = s$, d, $mean$, max or and). NSS(k) value of zero indicates no correspondence between saliency and eye positions, higher NSS, that can be above one, suggests a greater correspondence.

3.2.1 *Global analysis*

The NSS(k) was computed for each frame of all the clips (17000 frames). The mean value on these clips is given for each model of saliency map in Tab.1.

saliency map	M_s	M_d	M_{mean}	M_{max}	M_{and}
$\overline{NSS}$ value	0.79	0.81	0.99	0.84	1.00

Table 1: Mean NSS value on all the clips

For the dynamic pathway, results are as expected. The $\overline{NSS}$ value is high, as motion is an important gaze attractor. For the static pathway, results are better than those usually presented [14]. The $\overline{NSS}$ value for M_s is close to the M_d one. Every kind of fusion gives better NSS scores than a static or dynamic saliency map alone. M_{and} gives the highest score. Notice that NSS(k) values were computed for random fixations and they were close to 0 for all the models.

The NSS(k) values are then presented as a function of frame. The NSS(k) value at frame k on Fig.6 is the average of the NSS(k) values on every clip snippets at frame k. For the long clip snippets, the 65 first frames are kept, which correspond to 2.6 seconds.

All the curves have the same shape. The first values are weak and increase rapidly. This can be explained by the fact that after each shot cut, the gaze stays at the previous position during few frames and then moves to a salient region. The maximum NSS(k) value is reached for all the curves at about 13 frames, which corresponds to 520 ms, then curves decrease slowly. The shape of these curves can be explained by the fact that at the begining only bottom-up influences occurred, followed by top-down processes. M_s starts with a higher NSS(k) value, indeed the gaze stays at the previous position, which is more likely to be on a static salient region than a moving area. The dynamic saliency map has a higher maximum NSS(k) value and it decreases more rapidly than the static saliency one. After being attracted by a moving area, the gaze will go to other locations which are less salient for dynamic saliency but with high static saliency.

The $M_{max}(x,y,k)$ saliency corresponds to the maximum of static and dynamic for each map and then for each curve. The $M_{and}(x,y,k)$ and $M_{moy}(x,y,k)$ fusions are clearly above both static and dynamic saliency curves. The model of saliency maps that gives the highest NSS(k) during the first 25th frames is the M_{and}:

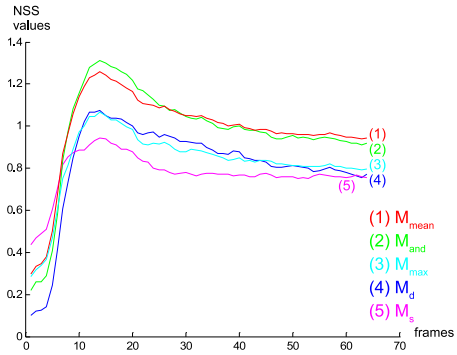

Figure 6: NSS as a function of frame. The NSS is averaged on 305 clip snippets for different saliency maps

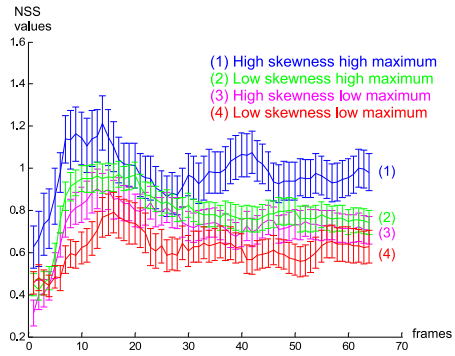

Figure 7: NSS as a function of frame for clip snippets categorized using maximum and skewness for static saliency map

as it retains what is salient in both static and dynamic saliency maps (Fig.5). Salient areas, defined by human eye positions, are salient both in the static and dynamic pathways. Both the static and dynamic pathways are correct predictors. After around 25 frames (1s) we can suppose that top-down is present, and subject gaze would go to less salient regions. As the M_{moy} fusion covers areas that are salient either in the static or the dynamic pathway, gaze is more likely to be on a salient area in M_{moy} than in M_{and}. Static and dynamic pathways are both needed to detect salient areas.

3.2.2 *Detailed analysis of the two pathways*

The static and dynamic pathways have been evaluated, we now want to find which information in the static and dynamic saliency maps is the most gaze-attractive.

The important characteristics of saliency maps are the maximum saliency value on the map and the dispersion of salient regions. For the dispersion information skewness is used: it is the third moment on the distribution of the $M_m(x,y,k)$ model saliency maps. Skewness gives information on the dissymetry of a distribution. If the saliency map contains every saliency level (gray level) in the same proportion, the skewness is low; on the other hand if the saliency map contains only a small salient region and all the rest is not salient, the skewness is high. The clip snippets are split into 4 categories. Each snippet is labeled with maximum (resp. skewness) information: above or below the median maximum (resp. skewness) value on all the snippet (Fig.7, 8). The maximum (resp. skewness) information for each snippet is the average of the maximum (resp. skewness) value on this snippet.

The static pathway is more predictive for snippets with high

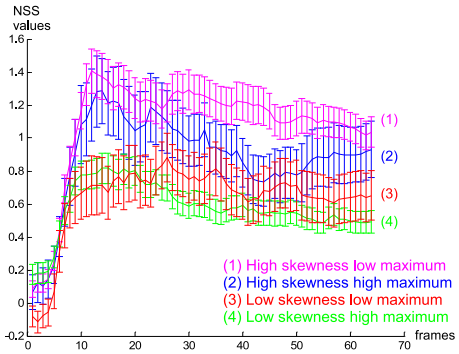

Figure 8: NSS as a function of frame for clip snippets categorized using maximum and skewness for dynamic saliency map

maximum in static saliency (maximum above 0.9), and then snippets with higher skewness are even more predictive (maximum at 1.1). Static saliency maps $M_s(x,y,k)$ give static salient information, if a region is textured for example. Static saliency covers in general more area on the frame and is not localized (Fig.4). The main information is given by the maximum of saliency value. If a frame has a high maximum value of saliency, there is an attracting region in this frame. On the other hand, if the frame has a low maximum of saliency the most attractive region is less attractive than in the previous frame.

The dynamic pathway is more predictive for snippets with higher skewness (maximum above 1.2) and then snippets with lower maximum are more predictive (maximum at 1.4). Map $M_d(x,y,k)$ gives motion information, but this time the salient regions may be localized (Fig.4). If there was only a small moving region, the saliency would be concentrated on this region. If there was only a dynamic salient region the gaze of all the subjects would be concentrated there. However if there were several regions with equivalent dynamic saliency, subjects' gaze would be spread over these different regions. The fact that NSS(k) is higher, with lower motion, can be explained by the fact that when a region moves rapidly, the eyes anticipate the motion, and gaze precedes the moving region. Gaze is then on a less salient region in our model.

4. CONCLUSION

In this paper a new bottom-up saliency model inspired by the biology of the human visual system is proposed. This model decomposed a visual signal into spatial information conveyed by a static pathway and motion information conveyed by a dynamic pathway. This decomposition starts with the two main outputs of the retina and continues with the cortical cells, sensitive to different spatial frequencies and orientations, and which are modeled using the same bank of Gabor filters for both static and dynamic pathways. The two pathways give two saliency maps that are fused into a spatio-temporal saliency map. The model is evaluated by comparing the salient predicted areas with human eye positions.

The proposed model is very efficient for a large number of frames (17000) coming from heterogeneous and realistic videos. The M_{and} fusion, which retains only salient regions in both static and dynamic, gives the best results, showing that salient regions are salient for both the static and the dynamic pathways. The analysis of NSS(k) as a function of time shows the effectiveness of prediction of our model during the period corresponding to the bottom-up mechanism. The categorization based on intrinsic information of the video shows the influence of particular characteristics of stimuli on human behavior.

In future work it would be interesting to use the spatio-temporal saliency maps obtained to improve video compression or watermarking. These saliency maps may also be useful for cropping frames to their most interesting part for video viewing on small display, or to help selecting video frames to make a video summary.

Acknowledgment

This work is partially supported by the LIMA project (Loisir et IMAges, http://liris.cnrs.fr/lima/) of the region Rhone-Alpes (France).

REFERENCES

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

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

[3] 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.

[4] 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.

[5] O. Le Meur, D. Thoreau, P. Le Callet and D. Barba,"A spatio-temporal model of the selective human visual attention," *ICIP*, vol. 3, pp. 1188-1191, 2005.

[6] Y. 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.

[7] W.H.A. Beaudot, P. Palagi, J. Herault, "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.

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

[9] D. Navon,"Forest before trees: the precedence of global features in visual perception," *Cognitive Psychology*, vol. 9, pp. 353-383, 1977.

[10] J.M. Odobez, P. Bouthemy, "Robust multiresolution estimation of parametric motion models," *JVCIR*, vol. 6, pp. 348-365, 1995.

[11] S. Marcelja, "Mathematical description of the responses of simple cortical cells" *J. Opt. Soc. Am. A*, vol. 70, pp. 1297-1300, 1980.

[12] 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.

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

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