A SPATIO-TEMPORAL MODEL OF THE SELECTIVE HUMAN VISUAL ATTENTION

O. Le Meur[1,2], *D. Thoreau*[1], *P. Le Callet*[2], *D. Barba*[2]

[1]THOMSON, 35511 Cesson-Sevigne; [2]IRCCyN UMR n°6597 CNRS, 44306 Nantes, France.

ABSTRACT

A new spatio-temporal model for simulating the bottom-up visual attention is proposed. It has been built from numerous important properties of the Human Visual System (HVS). This paper focuses both on the architecture of the model and on its performances. Given that the spatial model of the bottom-up visual attention has already been defined [1], the temporal dimension is more accurately described. A qualitative and quantitative comparison with human fixations collected from an eye tracking apparatus is undertaken. From the former, the quality of the prediction is deemed very good whereas the latter illustrates that the best predictor of the human fixation consists of the sum all visual features (achromatic, chromatic and motion).

1. INTRODUCTION

Despite the subjective impression that we see everything around us with a high quality, a relatively small portion of visual information is accurately treated. It is well known that the selective visual attention is driven both by a signal-based (Bottom-up) and by a task-based mechanism (Top-Down). The design of computational models simulating the human selective visual attention is a difficult issue. A very fine understanding of numerous properties of the HVS is mandatory in order to build a reliable model. The challenge really matters because the potential applications are considerable (video quality assessment, watermarking and video compression,...). Most of recent approaches can be divided into two categories: the first one concerns a statistical signal-based approach [2] whereas the second one, on which we work, attempts to simulate the major properties of the visual system.

The purpose of this study is to present a biologically-plausible signal-based model of selective visual attention selecting locations of high salience in a color video clip. The most important dimensions of the visual field are the achromatic, the chromatic and the motion. Such dimensions are taken into account leading to two saliency maps: spatial saliency map and tempo-ral saliency map. From these maps, a spatio-temporal saliency map is computed. The proposed model, and especially its extension to the temporal dimension, is described in section two. Section three refers to the eye tracking experiments. Section four presents the assessment method we use in order to compare the predictions with human fixations.

2. VISUAL ATTENTION MODELLING

The process allowing to build the saliency maps are detailed in the following sections, with a greater attention to the temporal saliency map determination (the construction of spatial saliency map have been already described [1, 3]). Figure 1 gives the flow chart of the proposed model.

2.1. Spatial saliency map

According to an important psychovisual backing, the spatial model consists of three sequential stages: visibility, perception and finally perceptual grouping stage.

The visibility part attempts to simulate the biologically limited sensitivity of the visual system: first, a transformation of the RGB component into the Krauskopf color space composed of the cardinal direction A (achromatic), Cr_1 (red and green opponent component) and Cr_2 (the blue and yellow opponent component) is achieved. Second, early visual features extraction is achieved by a perceptual channel decomposition (DCP) consisting in splitting the 2D spatial frequency domain both in spatial radial frequency and in orientation. This decomposition is applied on each of the three perceptual components leading to 17 psychovisual channels (distributed over 4 crowns) and 5 channels for chromatic component. Third, contrast sensitivity functions (CSF) are used to assess the visibility of natural images components, taken into account that we are not able to assess all details with the same accuracy. Three anisotropic CSF are used to weight the components (A, Cr_1, Cr_2). Finally, visual Masking effect refers to the modulation of the differential visibility

threshold due to the influences of the context (spatial, intra-inter channels, inter components interactions).

The perception step attempts to simulate the perception process. The perception produces from the psychovisual space a description useful to the viewers and not cluttered with irrelevant information. By simulating the center/surround suppressive organization of cortical cells, two relevant structural descriptions (one for the achromatic component and one for the chromatic components) are built.

The perceptual grouping step computes a spatial saliency map. This two dimensional map is obtained by linearly combining the two structural descriptions.

We have shown [3] that this model provides a correlation coefficient close to 0.72 between the output of the model and the experimental results. It is noticeable that this model outperforms Itti's model [4].

2.2. Temporal saliency map

Figure 1 gives the synoptic of the proposed algorithm used to build the temporal saliency map. The method is based on the fact the visual attention is attracted by motion contrast. Such contrast is deduced from the local and the global motion. The proposed technique is similar in spirit to the approach proposed in [5] in which the saliency of a motion region is inversely proportional to its occurring probability.

2.2.1. Hierarchical block matching motion estimation

Motion estimation plays a very important role in the construction of the saliency map. Apparent motion is computed between two successive frames using a hierarchical block matching method. In general, the hierarchical decomposition is performed by a dyadic Gaussian pyramid: the input image is first filtered by a 2D separable filter and then subsampled (horizontally and vertically by a factor of two). This process is iteratively applied to build up each level of the Gaussian pyramid. Here, we take advantage of the perceptual channel decomposition performed during the first steps of the spatial saliency map creation.
Two neighboring pictures are used to form two pyramids. $\overrightarrow{V}^i_n$ denotes the motion field for the nth frame, at the ith level of the pyramid. At the lowest resolution, the motion vector providing the smallest sum of absolute difference is kept, up sampled and transmitted to the next higher resolution. Refinement algorithm and decision are used to form the final motion field $\overrightarrow{V}_{local}$.

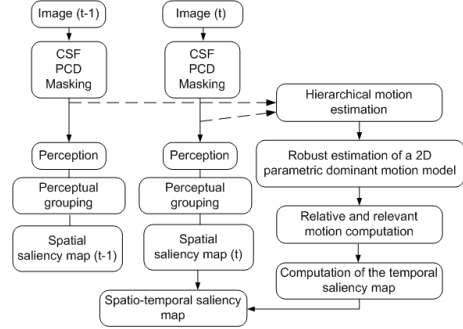**Fig. 1**. Architecture of the proposed model.

2.2.2. Estimation of a 2D parametric dominant motion model

In order to detect the temporal conspicuous areas of a video sequence, we have first to cancel the motion inherent to the camera. Assuming that the dominant motion is due to the camera, we estimate the global transformation between successive images based on a previous estimated motion fields. The displacement $\overrightarrow{V}_\Theta(s)$, at a pixel site s related to a motion model parametrized by Θ is given by a 2D affine motion model:

$$\overrightarrow{V}_\Theta(s) = \begin{pmatrix} a_1 + a_2 x + a_3 y \\ a_4 + a_5 x + a_6 y \end{pmatrix} \qquad (1)$$

where $\Theta = [a_1, a_2, a_3, a_4, a_5, a_6]$ represents the affine parameters of the model. The affine parameters are computed with a popular robust technique based on the M-estimators. They lessen the outliers effects by replacing the squared residuals errors by another function. The estimated affine parameters $\widehat{\Theta}$ have to minimize: $\widehat{\Theta} = arg\min_\Theta \sum_{s_i \in S} \rho(r(s_i))$. $r(s_i) = \tilde{I}(s_i + \overrightarrow{V}_\Theta(s_i), t + 1) - \tilde{I}(s_i, t)$ represents the displaced frame difference. ρ is the Tukey's biweight function.

2.2.3. Relative and relevant motion

From the knowledge of the apparent dominant displacement $\overrightarrow{V}_\Theta$ and of the local displacement $\overrightarrow{V}_{local}$ for each pixel s, the relative motion $\overrightarrow{V}_{relative}$ is computed by $\overrightarrow{V}_{relative}(s) = \overrightarrow{V}_\Theta(s) - \overrightarrow{V}_{local}(s)$. As the perception of a moving object heavily depends on whether or not the object is tracked by the eyes, we introduce the maximal pursuit displacement capability of the eyes. The relative displacement greater than the maximal pursuit displacement is discarded. For video, this value belongs to the range 8 to 10 deg/s.
The relevance degree of a relative motion also depends on the average amount of the relative displacement computed across the picture. For example, a high relative motion is very conspicuous when there is only

few relative displacement. It is intuitively clear that it will be easy to find a moving stimulus among stationary distractors [6], [7]. To model such property, a linear quantification of $\|\overrightarrow{V}_{relative}\|$ is achieved in order to build a histogram. The median value of the histogram, called Γ_{median}, is a reliable estimator of the relative motion amount. $\|\overrightarrow{V}_{relative}\|$ is then weighted by Γ_{median}. The closer Γ_{median} to 0, the more the relative motion is perceptually important. Finally, the normalized temporal saliency map S^T is computed by:

$$S^T(s) = \frac{\overrightarrow{V}_{relative}(s)}{\Gamma_{median} \times \lambda}, \; \lambda = \max_s(\frac{\overrightarrow{V}_{relative}(s)}{\Gamma_{median}}) \quad (2)$$

2.3. Spatio-temporal saliency map

To cope with the difficult problem of the saliency maps fusion, the following relation is proposed:

$$S(s) = \alpha S^T(s) + (1 - \alpha)S^S(s) + \beta S^T(s)S^S(s) \quad (3)$$

where S^S and S^T are the normalized spatial and temporal saliency map respectively. α is deduced from the spatio-temporal activity estimated by the frame difference (FD) value (See figure 2). β controls the strength of the reinforcement. Figure 3 shows for the sequence
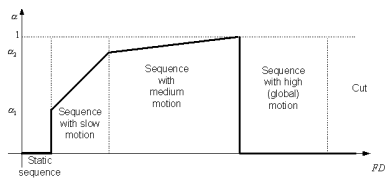

Fig. 2. Computation of the α coefficient depending on the spatio-temporal activity of the sequence.

Stefan the different saliency maps. On these maps, the most conspicuous areas detected by the model are highlighted.

Fig. 3. Example of spatio-temporal saliency maps for the sequence *Stefan*. From the left to the right, we have spatial, temporal and spatio-temporal saliency map.
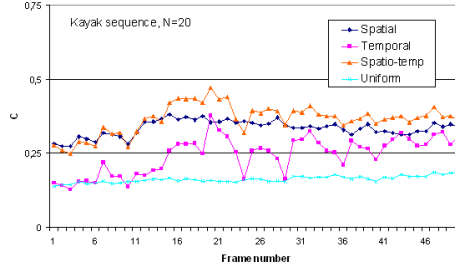

Fig. 4. Temporal evolution of C value for different models. Results concerns the sequence *Kayak*.

3. EYE TRACKING EXPERIMENTS

In order to track and record real observers eye movements, experiments have been performed with an eye tracker from Cambridge Research Corporation. This apparatus is mounted on a rigid headrest for good accuracy on the measurement of the fixation point (σ is less than 0.5 degree). Experiments were conducted in normalized conditions (ITU-R BT 500-10) at viewing distance of 4 times the TV monitor height. Four natural color sequences of various contents have been selected. Each sequence was seen in random order by up to 30 observers each in a task-free viewing mode. The collected data correspond to the regular time sampling (20 ms) of eye gaze on the monitor. For each particular picture of a video clip and for each observer, a fixation map, which encodes the conspicuous locations, is computed from the collected data. Afterwards, all fixation patterns are added together providing a spatial distribution of human fixation. The resulting map is then smoothed using a 2-Dimensional Gaussian filter. The standard deviation is determined in accordance with the accuracy of the eye-tracking apparatus. The appliction of the Gaussian filter is necesary to simulate the fact that we stare at a particular area rather than at a particular point. The fixated area has a size close to the fovea's size. A fixation density map representing the observer's regions of interest is finally obtained.

4. QUANTITATIVE COMPARISONS

In order to quantify the correlation C between the predicted saliency map and the human map, the following algorithm, inspired from the work [8], is conducted for each picture of the considered sequence. To begin, the two maps are transformed into two dense probability functions. Next, the coordinates (x^k, y^k) of the kth most important fixation location are extracted from the human map H. The salience, which is in fact a probability, is then extracted from the predicted probability function P. Finally, for the ith picture, the extracted
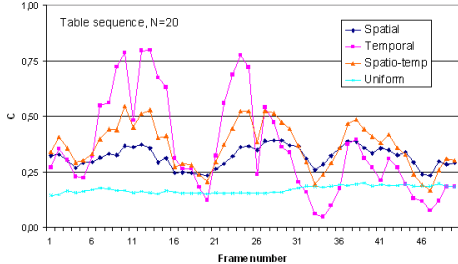
Fig. 5. Temporal evolution of C value for different models. Results concerns the sequence *Table*.

Table 1. Average value of C for the predictions and accuracy value obtained from a confusion matrix.

N=20	S^S	S^T	S	Uniform	Accuracy
Kayak	0,33	0,16	0,36	0,16	0,84
Table	0,32	0,35	0,37	0,17	0,71
PatSpeed	0,34	0,32	0,4	0,18	0,82
Mobile	0,28	0,39	0,33	0,21	0,72

salience values are summed: $C^i = \sum_{k=1}^{N} P^i(x^k, y^k)$, N is the number of the most important fixations that we used. An uniform model which is the worst case allows to bound the similarity degree. Table 1 gives the average value of C computed over the whole sequence (for $N = 20$) whereas the figures 4 and 5 illustrate the temporal evolution of the similarity degree for different configurations of the proposed model. In average, the best model is the model incorporating all the visual features; the sum of all visual features provides the best predictor of human fixations. It is obvious that, in presence of still video content or of complex dynamic scenes, the strength of each visual feature is not the same. The fusion step, based on the spatio-temporal activities, deals with all of these cases. Moreover, in a general case, motion is good predictor of human saccades. Finally, although that top-down influences can be very strong, it is now clear that bottom-up attention continuously alerts us to salient items in the visual field. These results are in agreement with the studies [8] and [9]. Table 1 gives also the accuracy of the model, computed from a confusion matrix. It contains information about actual and predicted classifications achieved by a classification system. Two categories are considered: a pixel could be either of interest or not. It is noticeable that the average value of the accuracy of the prediction (i.e the proportion of the total number of correct predictions) is about 80 percents. Given that the proposed model is only based on the low level attributes, this result is very encouraging. The best results are obtained for the sequences *Kayak* and *PatSpeed*.

5. CONCLUSION

In this paper, a spatio-temporal model simulating the bottom-up selective visual attention has been proposed. From human fixations, collected from an eye tracking experiments, a qualitative and a quantitative comparisons have been done. The predicted locations of interest are qualitatively considered to be in good agreement with the experimental results. Quantitative analysis shows that the prediction is improved when all visual features are used. Given that the proposed model is only based on the low level features, the accuracy of the prediction is really good: 80 percents of the human fixations are well detected. As in [5], this technique must be improved to deal with the case of relative large foreground. In this case, the regions of the background could be considered as of a great interest. In future works, we will cope with the aforementioned problem. In addition, we will pursue to assess and to optimize the proposed model. A convenient manner will be to compare it with other models.

6. REFERENCES

[1] O. Le Meur, P. Le Callet, D. Barba, D. Thoreau, "From low level perception to high level perception, a coherent approach for visual attention modeling", HVEI-04, CA, 2004.

[2] U. Rajashekar, L. Cormack, and A. Bovik, "Image features that draw fixations", ICIP-2003.

[3] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "Performance assessment of a visual attention system entirely based on a human vision modeling", ICIP-2004.

[4] L. Itti, C. Kock, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis", IEEE Trans. Pattern Anal. Mach. Intell. (PAMI), vol. 20, N°11, pp. 1254-1259, 1998.

[5] Y. Ma, L. Lu, H. Zhang and M. Li. "An Attention Model for Video Summarization", ACM, pp. 533-542, France, 2002.

[6] J. Wolfe, "Visual Search", in Pashler, H. editor, Attention, pp. 13-74, Psychology Press, 1998.

[7] C. Royden, J. Wolfe, and N. Klempen, "Visual search asymmetries in motion and optic flow fields", Percep. and Psycho., vol. 63, pp.436-444, 2001.

[8] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention", Vision Research 42, pp. 107-123, 2002.

[9] L. Itti, "Quantifying the Contribution of Low-Level Saliency to Human Eye Movements in Dynamic Scenes", Visual Cognition, 2005.