PERFORMANCE ASSESSMENT OF A VISUAL ATTENTION SYSTEM ENTIRELY BASED ON A HUMAN VISION MODELING

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

[1]-THOMSON R&D France, 1 Avenue de Belle Fontaine, 35511 Cesson-Sévigné cedex, FRANCE.
[2]-IRCCyN UMR n°6597 CNRS, Ecole Polytechnique de l'Université de Nantes, rue Christian Pauc, La chantrerie, BP50609, 44306 Nantes cedex, FRANCE.

ABSTRACT

It is now commonly assumed that the human visual attention, which is a selecting process of the most relevant locations in a scene according to a particular behavior, is driven by both top-down (task-dependent) and bottom-up (signal-dependent) control. A new model attempting to simulate the bottom-up process has been designed [1]. This model is purely based on visual system properties that provides noticeable advantages compared to the classical published approaches. This paper focuses on the performance assessment of this model by achieving a comparison with real fixation points stemming from eye-tracking apparatus both subjectively and objectively.

1. INTRODUCTION

At any time, the human visual system (HVS), which is intrinsically limited, has to deal with a great quantity of visual information. To tackle this problem, the process of visual attention, which allows human to select the most important information in cluttered visual environments, is assumed to be controlled by two different mechanisms called bottom-up (signal-dependent) and top-down (task-dependent) control, respectively. In this paper, we are concerned with the first mechanism, the bottom-up control.

Among the widespread potential applications, image and video ROI-based (region of interest) coding schemes can take benefit of bottom-up visual attention models.

For a decade now, several models have been developed to simulate visual attention mechanism leading to generate a set of conspicuous locations or saliency maps. Only some of these provide some interesting but limited results pointing out the real difficulty to incorporate all the different dimensions of this complex task. Most of the recent approaches can be divided into two categories. One category concerns a statistical signal-based approach [2] while the other consists in incorporating more or less part of human vision properties [3,4].

Recently, we proposed a model [1] which belongs to the last category with a real emphasis on human vision properties comparing to the others and especially the reference model proposed by L. Itti [3]. The novelty lies on the fact we use a fully psychovisual space on which visual phenomena can be easily described and combined. Currently, only a subset of visual phenomena has been modeled. Nevertheless, the performances of the model are in a good agreement [1] with our own subjective opinion and with Itti's model.

The objective of the paper is to assess the real performances of our method by comparing its outputs with data computed from real measurements provided by an eye-tracking experiments from human observers. Since the comparison is not a trivial process, several methods to achieved it, will be used and tested. The proposed model is briefly described in section two. Readers could find more details in [1]. The third section presents the eye-tracking experiments and the way to extract the interesting features from the rough data. Section four describes the mean to perform the comparison and to obtain the model's performances.

2. FROM AN IMAGE TO ITS SALIENCY MAP

The elements of the model we use have been described in a previous paper [1]. We recall briefly in this paper the fundamental basis of the proposed method but highlight its main differences with classical approach like Itti's one. According to a psychovisual backing, the model consists of three main steps : visibility step, perception step and perceptual grouping step. The whole synoptic is shown in figure 2.1.

The visibility step attempts to simulate the limited sensitivity of the human visual system (HVS). Despite the seemingly complex mechanisms underlying the human vision, it is clear that the visual system is not able to

* Olivier.le-meur@thomson.net

0-7803-8554-3/04/$20.00 ©2004 IEEE.

perceive all information present in the visual field with the same accuracy. To take into account these intrinsic limitations, the visibility step includes the following set of basic mechanisms entirely identified from psychophysic experiments (conducted at the IRCCyN's laboratory):

- Transformation of the RGB luminance into the Krauskopf's color space composed of the cardinal direction A, Cr1 and Cr2. This conversion allows us to simulate the 3 different pathways used by the brain to encode the visual information. The first pathway conveys the achromatic component (A), the second the red and green opponent component (Cr1) and the third the blue and yellow opponent component (Cr2).
- Perceptual channel decomposition, which consists in splitting the 2D spatial frequency domain both in spatial radial frequency and in orientation, is applied on each of the three components. Psychovisual spatial frequency partitioning for the achromatic component leads to 17 psychovisual channels in standard TV viewing conditions while only 5 channels are obtained for chromatic component.
- Contrast sensitivity function (CSF) is applied on each components to weight the magnitude of the components (normalization with the differential visibility threshold without masking effects).
- 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). Readers could find more details in [5].

All these previous mechanisms lead to transform the image data into a fully psychovisual space. Since all features are expressed in term of differential visibility threshold, it is possible to manage visual features stemming from different modalities. For instance, chromatic information could be directly compared to achromatic ones while these interactions cannot be easily implemented in other approaches.

The second part of our model deals with perception. The perception is a process that produces from the psychovisual space a description useful to the viewers and not cluttered with irrelevant information.

In a first approach, we only built a structural description of the achromatic component. This description is obtained by two mechanisms :

- Achromatic reinforcement by chromatic context : it consists in increasing the magnitude at each site of the achromatic channels by taking into account the local oriented smooth gradient of the chromatic components.
- Center/surround suppressive interaction : it emulates the action of the classical receptive field (CRF) of visual cells belonging to the primary visual cortex which is modeled by using a difference-of-Gaussian function.

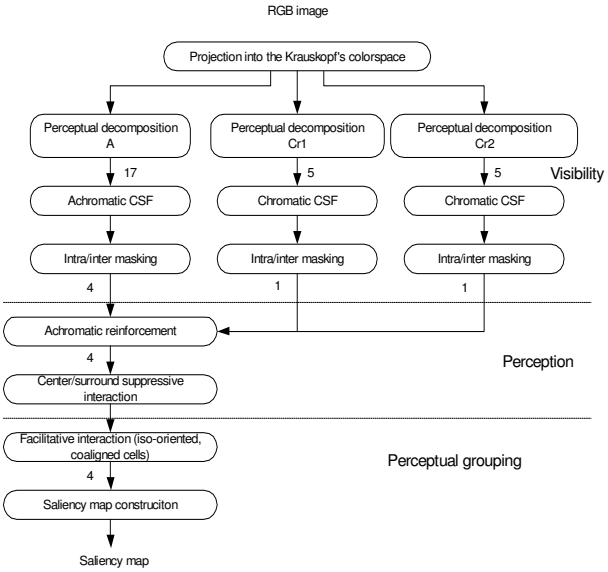

figure 2.1 : general synoptic of our model.

The last step focuses on some aspects of the perceptual grouping action. It refers to the human visual ability to group and to bind visual features to set up a meaningful higher-level structure. One important aspect we considered here is the enhancement of iso-oriented and co-aligned linear structures.

A two-dimensional saliency map is thus computed by summing directly the outputs of the different achromatic channels. This simple sum up is possible since in all the previous steps a coherent normalization has been used. This property denotes again a main great difference with techniques proposed in other approaches. An anisotropic Gaussian function is finally applied to favour the central part of the picture.

3. EYE TRACKING EXPERIMENTS

3.1. Apparatus and procedure

In order to track and record real observers eye movements, we have conducted experiments with an eye tracker from Cambridge Research Corporation. This apparatus is mounted on a rigid headrest allowing good accuracy on fixation point measure (less than 0.5°). Experiments were made in normalized conditions (ITU-R BT 500-10) at viewing distance of 4 times the TV monitor height. We have selected 46 natural color and grayscale images with various contents. Every image was seen in random order by up to 40 observers during 15 seconds each in a task-free viewing mode. The collected data correspond to the regular time sampling (20 ms) of eye gaze on the monitor.

3.2. Human fixation map computation

From the collected data, we compute a fixation map which encodes the conspicuous locations.

For a particular picture and per observer, we first filter out the samples which correspond to saccades. All fixations patterns coming from different observers for a particular image are summed. Finally, a 2D gaussian filter with a standard deviation of half degree of visual angle is applied in order to approximate the size of the fovea.

The human fixation map depends on the viewing time. Since we have the time stamp of the eye tracker sample, we are able to build fixation map for different viewing time. Figure 3.1 exhibits examples of human fixation maps for two different viewing times (1s and 14s respectively).

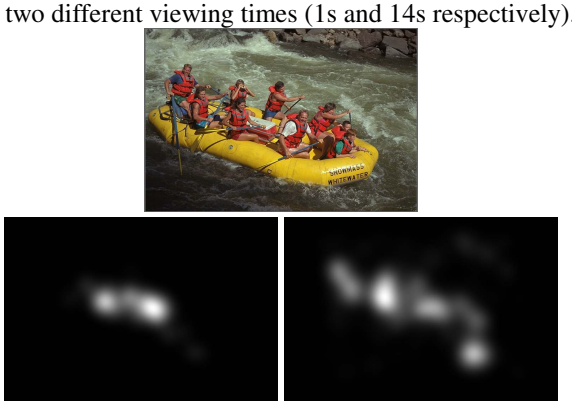

figure 3.1: first row : source picture; second row : human fixation map with gaussian filtering, viewing time: left 1s, right 14s.

From the collected data, we can also derive the coverage value which is a measure of the amount of the original stimulus covered by the observers. Obviously this value depends on the viewing time since the proportion of the map covered by humans increases with the viewing time and depends also with the picture contents. This dependency is shown in figure 3.2.

4. COMPARISONS BETWEEN HUMAN AND PREDICTION FIXATIONS

Evaluating the performance of our model by comparing observer fixations and fixation predictions is an unavoidable step. Several methods for comparing human fixations to predictions are discussed in [6]. However, the quantification of the similarity still remains a tickly problem and there is not a reasonable consensus about a particular method one should use. Consequently, we present hereafter a subjective comparison and an objective comparison.

4.1. Subjective comparison

In a task-free viewing, it is well known that human beings pay attention only to few parts of an image and pursue to focus on these few areas rather than scanning the whole image. Figure 3.2 emphasizes this particular property. After 14 seconds, the coverage is only about 45% on average. Thus, when the viewing time increases, it is easier to detect or to segment the most important regions of interest of a picture.

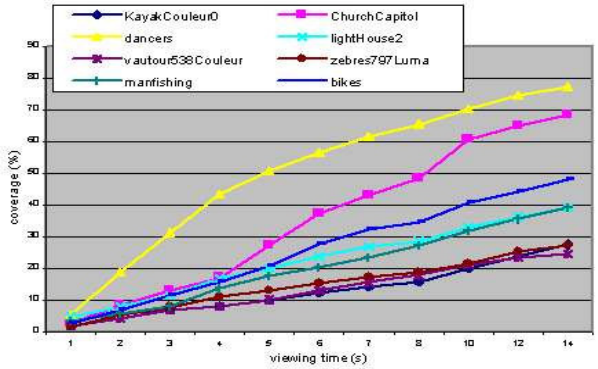

figure 3.2 : coverage value in percent as a function of viewing time, for different images.

The subjective comparison consists in screening the locations of the first 4 and 6 biggest values in the two density maps (humans and model). Our fixation predictions are indeed compared with the human fixations obtained from 2 different viewing times : 6 and 14 seconds. Figure 4.1 emphasizes the most salient regions for picture *Kayak*. The bright circles correspond to human fixations while the dark circles represent fixations given by the model.

figure 4.1 : comparison between human and model fixation points (the first 4 and 6 fixations) on picture Kayak. First row : viewing time =6s; second row : viewing time = 14s.

The region of interest for picture *Kayak* (figure 4.1) is well detected by our method and its good degree of the similarity with the human fixations for the two viewing times shows our method is in good agreement with the real human behavior. Nevertheless, the similarity is highest for the greatest viewing time due to the fact that observers continue to focus on the region having the greatest interest (the kayak). Implausible points tend to disappear.

4.1. Objective comparison

In order to provide an objective measure of the dissimilarity degree, the Kullback-Leibler divergence is

used. It consists in considering our saliency maps as two dense probability density functions. The degree of dissimilarity between two probability functions h (the human probability density function) and p (the prediction probability density function) is given by the Kullback-Leibler metric (noted KLD) [7]:

$$KLD(p|h) = \sum_{x \in \aleph} p(x) \log(\frac{p(x)}{h(x)})$$

When the two probability densities are strictly equal, the KLD value is null.

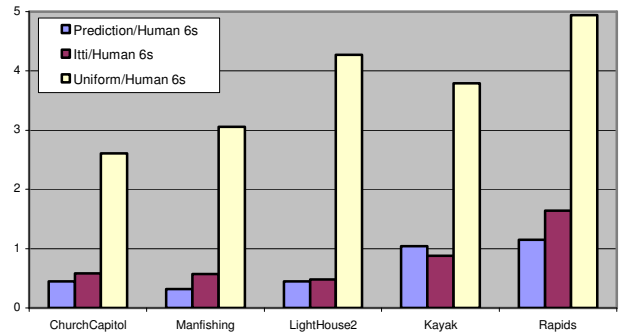

figure 4.2 : distance value based on the Kullback-Leibler divergence for a 6s viewing time.

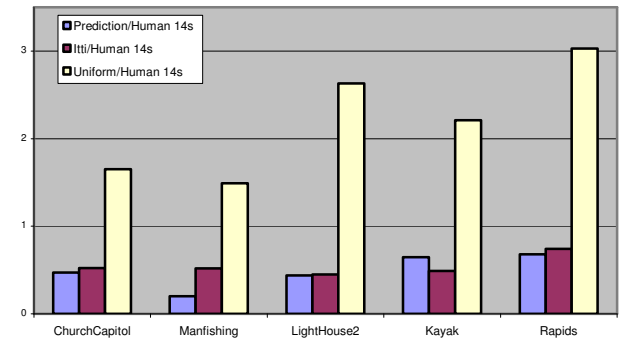

figure 4.3 : distance value based on the Kullback-Leibler divergence for a 14s viewing time.

Degree of dissimilarity are shown in figures 4.2 and 4.3 for 2 different viewing times. A reference measure is provided by computing the degree of dissimilarity (noted KLD_r) between the probability density function coming from the human fixations and an uniform probability density function. The degree of dissimilarity is thus bounded by 0 and by the KLD_r value. Moreover, the comparison is also achieved with the modified model of L. Itti [3] proposed by [8]. Objective results confirm the observations previously emphasized by subjective assessments. Best results are obtained when the greatest viewing time is considered. The best degrees of dissimilarity are obtained for pictures *LightHouse2* and *Manfishing*. Moreover, on a set of 10 pictures, the improvement of the average KL is about 12% compared to the best version of L. Itti's model.

5. CONCLUSION

In this paper, we have examined the degree of similarity between human fixations and predicted fixations given by a bottom-up model of visual attention.

Currently, the proposed model extracts fixations emphasizing only achromatic structures. But, since it is coherently built on a modeling of the human visual system, it will be easy to include new visual phenomena.

Because of the difficulties to compare human fixations and predicted fixations, subjective and objective comparisons have been led. The former method consists in comparing the first real fixation points with fixation points produced by the model. Two viewing times have been considered pointed out that the results are closest for the greatest viewing time. The latter aimed at quantifying the degree of similarity by using the Kullback-Leibler divergence. Objective measurements often confirm the good agreement between human fixations and predicted fixations.

In future works, we will pursue the improvement of our model by integrating new visual phenomena especially the temporal dimension.

6. REFERENCES

[1] O. Le Meur, P. Le Callet, D. Barba, D. Thoreau and E. François, "From low level perception to high level perception, a coherent approach for visual attention modeling", *in proc. SPIE Human Vision and Electronic Imaging*, San Jose, CA, 2004.

[2] U. Rajashekar, L. K. Cormack and A. C. Bovik, "Image features that draw fixations", *in proc. ICIP'03*, Barcelone, 2003.

[3] 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.

[4] B. Bruce, E. Jernigan, "Evolutionnary design of context-free attentional operators", *in proc. ICIP'03*, Barcelone, 2003.

[5] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "Masking effey in visual attention modeling", *WIAMIS*, Lisboa, Portugal, 2004.

[6] D. S. Wooding, "Eye movements of large population : II. deriving regions of interest, coverage, and similarity using fixation maps", *Behavior Research Methods, Instruments and Computers*, 34(4), pp. 509-517, 2002.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 1991.

[8] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention", *Vision Research*, Vol. 42, N°1, pp. 107-123, 2002.