A SELECTIVE ATTENTION MODEL FOR PREDICTING VISUAL ATTRACTORS

Éric Dinet, Emmanuel Kubicki

LIGIV, Université Jean Monnet
18 rue Benoît Lauras 42000 Saint-Étienne, France

ABSTRACT

The huge amount of visual information continuously received by an observer cannot be wholly analyzed by the brain. In order to interact efficiently with the environment, an observer has to select region of interests in the visual scene. Only the regions of interest will be processed in details by cortical structures.

This paper aims at introducing a selective attention model able to predict the location of visual attractors in natural scenes. The underlying idea is to extract and combine, in a competitive process, early visual features such as color and spatial arrangements to construct a saliency map coding interest areas in correlation with human visual behavior. The purpose is to effectively locate which region of a scene would attract the gaze of an observer and then where computational resources should be directed for a selective image processing.

Index Terms— Visual system, Image processing, Visual attention, Eye tracking, Salient regions

1. INTRODUCTION

With visual attention mechanisms that select parts of the scene to be foveated and processed with high resolution, an observer can focus his attention only on local regions of interest [12, 15]. In other words, visual attention mechanisms allow to direct and to concentrate processing resources only on important information.

The selection of regions of interest is driven both by neurological and cognitive resources [7, 9, 16]. Neurological resources refer to bottom-up (stimuli-based) information when cognitive resources refer to top-down (task-dependent) cues. Bottom-up information is controlled by low-level image features that stimulate achromatic and chromatic parallel pathways of the human visual system. Top-down cues are controlled by high-level cognitive strategies largely influenced by memory and task-oriented constraints.

During the last decade, knowledge in visual attention has been exploited in computer vision for selective image processing. Different biologically plausible models of attention were introduced to identify local visual attractors over the entire scene [1, 3, 8]. Local visual attractors are commonly extracted by quantifying signal properties such as intensity variations or contour orientation.

Models of visual attention would be useful in many fields of computer vision. For instance, the quality of image parts with high saliency could be preserved during a compression process or could be enhanced when displayed. The main underlying idea is that a given processing strategy could be efficiently adapted to the characteristics of each input image.

In this paper, we propose a bottom-up approach for modeling visual attention. The proposed approach is designed to detect regions of interest in images that are the most attractive for observers. These attended regions are directly given by a saliency map. The salience of every scene location is computed at multiple spatial scales for chromatic and achromatic variations. Different from the previous methods, our computational model of attention is consistent with neural mechanisms of the human visual system. Several important physiological features such as the relative ratio of photoreceptors, the normalization mechanisms of the parallel visual channels or the cell sensitivity in terms of frequency bands are considered. Chromatic variations are quantified through a multilayer perceptual representation of photoreceptor signals that fits to human perception.

The proposed model provides a map of perceptual saliency values resulting of a competitive process between bottom-up cues. Regions with higher scalar values in the final saliency map are more likely to correspond to the regions firstly picked by observers. One major advantage of our approach compared to the existing algorithms is that the output map provides region level attention in addition to the pixel level saliency value. To validate our model of visual attention, we tested it with ninety different images of natural scenes and paintings. The resulting saliency maps were compared to data recorded during eye tracking experiments conducted with thirty-three observers. Very encouraging results have been obtained.

The remainder of this paper is organized as follows. The computational model is presented in Section 2. Section 3 describes eye tracking experiment procedure. Section 4 is dedicated to the comparison between computed results and recorded data. A brief discussion will conclude this paper in section 5.

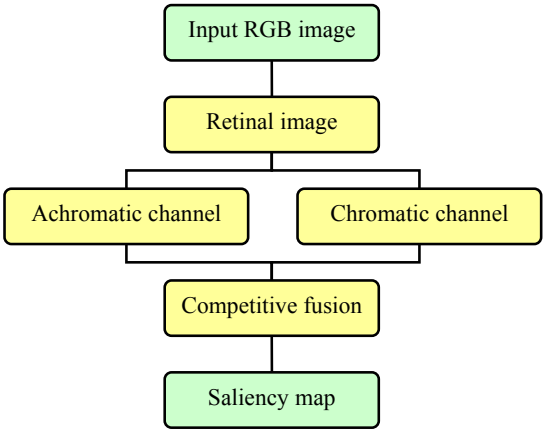


Figure 1. General work flow of the proposed model of attention. It is based on two parallel channels processing independently color information and achromatic variations. The outputs of these two channels are combined using a competitive approach to produce the final saliency map.

2. COMPUTATIONAL MODEL

The general work flow of the proposed computational model of visual attention is presented in Figure 1. Human visual system is sensitive to the chromatic and achromatic contrast of visual signals [13, 14]. Chromatic and achromatic signals are independently processed by distinct cortical structures. Taking this key feature, we propose to integrate two distinct modules in our model. The first one derives intensity and orientation variations when the second one processes color information. As in the human visual system, the two channels are fed by a common input and their outputs are fused into a single saliency map [17].

2.1. Retinal image

In photopic vision, it is generally assumed that the human visual perception is based on the signals provided by three types of photoreceptors called cones *L*, *M* and *S*. In a first stage, the responses of these photoreceptors are integrated in the computational model by transforming the original RGB image into a retinal image. This first process consists of projecting the raw RGB signals of the input image into the LMS color space using the linear transformation procedure recommended by the CIE [4].

Inspired by experiments based on small Mondrians suggesting that human visual system independently normalizes to the maxima for each cone photoreceptor type, color data are normalized in the LMS space prior to input to the achromatic and the chromatic channels [11].

Different studies show that the *S* cones constitute less than 10 % of the total cone population and that the *L* cones are roughly twice as numerous as the *M* cones [5]. Therefore in the selective attention model we will assume that the proportions of *L*:*M*:*S* cones are on average 10:5:1.

2.2. Achromatic channel

Cones are contacted by horizontal and bipolar cells. Bipolar cells difference the direct input from the photoreceptors and the indirect input transmitted from neighboring photoreceptors through horizontal cells. The anatomical connections define concentric regions organized in a centre-surround structure well-known as receptive field. Such a receptive field is a general architecture in the retina but also in primary visual cortex [15].

The centre-surround oppositions allow to efficiently detect local spatial discontinuities. As discontinuities are more likely to attract visual attention, receptive fields participate in detecting locations which are more salient. The centre-surround oppositions are implemented in the model with a Laplacian pyramid. As suggested by Itti et al., such a multi-resolution approach is interesting to simulate the multi-scale feature extraction performed by the human visual system [8].

Horizontal interactions between all cones make the pyramidal cell's response dependent on both the amplitude of signals and the local spatial organization of stimuli. Then the pyramidal structure of the achromatic channel codes the most salient variations of achromatic contrasts.

Psychophysical and physiological evidences indicate that orientation-sensitive neurons are present in the primary visual cortex. The receptive field sensitivity of these neurons can be well approximated by bidimensional Gabor filters [2, 10]. The output of the achromatic pyramidal structure is then convoluted with a bank of Gabor filters chosen to fit psychophysical data. Gabor filters were tuned with the parameters given in [3].

The achromatic channel of our model of attention provides an intermediary map that identifies the most salient achromatic visual stimuli in the input color image as it could be performed in the primary visual cortex.

2.3. Chromatic channel

Anatomical evidence reveals the presence of different spectrally opponent cells in the human visual system. These cells respond in opposite directions to light wavelength shifts. This suggests that signals from the output of each cone type are merged in different opponent strategies. In 1993, Russell and Karen de Valois proposed an efficient model of these early opponent strategies [6]. Inspired by such a model, the chromatic channel is based on a pyramidal structure that simulates the response of cone-opponent cells.

The pyramidal processing of chromatic data in which cone-opponent comparisons are made codes the chromatic variability within the scene. The chromatic channel of the proposed model provides an intermediary map that identifies

the most salient chromatic visual stimuli in the input color image as it could be performed in the primary visual cortex.

2.4. Competitive fusion

The outputs of the achromatic and the chromatic channels have to be fused into one saliency map. In cortical areas, neural signals are in competition and they are merged according to their relative strength. Such a merging scheme is reproduced to compute the output saliency map of our model. The strongest salient locations are promoted in each intermediary map and a competition process reveals the redundancy of information. The final saliency map identifies regions with high energy extracted by opponent strategies and preserved through the multiple scales of the model.

3. EYE TRACKING EXPERIMENT

The relevancy of regions of interest extracted by the computational model we propose can be determined by how they fit to human visual behavior. Therefore, we recorded where the gaze of observers is attracted through an eye tracking experiment.

Thirty-three subjects, ranging from 18 to 65 years, participated as unpaid volunteers. All had normal or corrected to normal vision and all had no color vision defect. All observers were naïve to the experiment.

Eye movements of subjects were recorded using a QuickClamp eye tracker from Arrington Research Inc. The eye tracker system is mounted on a stable head and chin stabilizer. It incorporates an infrared camera and an infrared LED as illumination source. The apparatus has a theoretical accuracy of 0.5 deg. raw eye positions and the camera records with a rate of 60 Hz.

Ninety color images with a resolution of 1280×1024 pixels have been selected as visual stimuli. Selected images have various contains such as landscapes, roads, crowds, animals, flowers, abstract or figurative paintings, etc (see Figure 2 for some examples). They were presented on a calibrated 20-inch CRT monitor at a viewing distance of 70 cm with a frame rate of 85 Hz.

The subjects were placed in a darkened room. The height of the head and chin stabilizer was adjusted so that each observer was comfortable and with the gaze at the level of the center of the monitor. The subjects were instructed not to move during the experiment.

Before the beginning of each experiment, a calibration procedure was performed. During such a procedure, the subjects are asked to fixate sixteen points that are sequentially presented at different locations of the screen. The calibration procedure is intermittently repeated during the experiment. Then the sequence of events was as follows: firstly a white fixation disk (2° visual angle) on a dark grey background was presented at the center of the monitor during 300 ms. This ensures that the observer starts his viewing task at the same position for each image and it allows to check if there is no shift in the calibration. Then an image was displayed during 3000 ms. Observers were instructed to freely look at the image.

Fixations were derived from each raw eye tracking records. According to the mean fixation duration (300 ms), our results are in accordance with those obtained in other works. As we want to evaluate the effectiveness of our model of attention, individual scan-paths are ignored and valid data are merged.

4. COMPARISON

Figure 2 presents some typical results obtained with visual stimuli used in our experiments. The final saliency map given by our selective attention model codes the regions of interest of input images. These regions should correpond to the visual attractors that are more likely to attract attention.

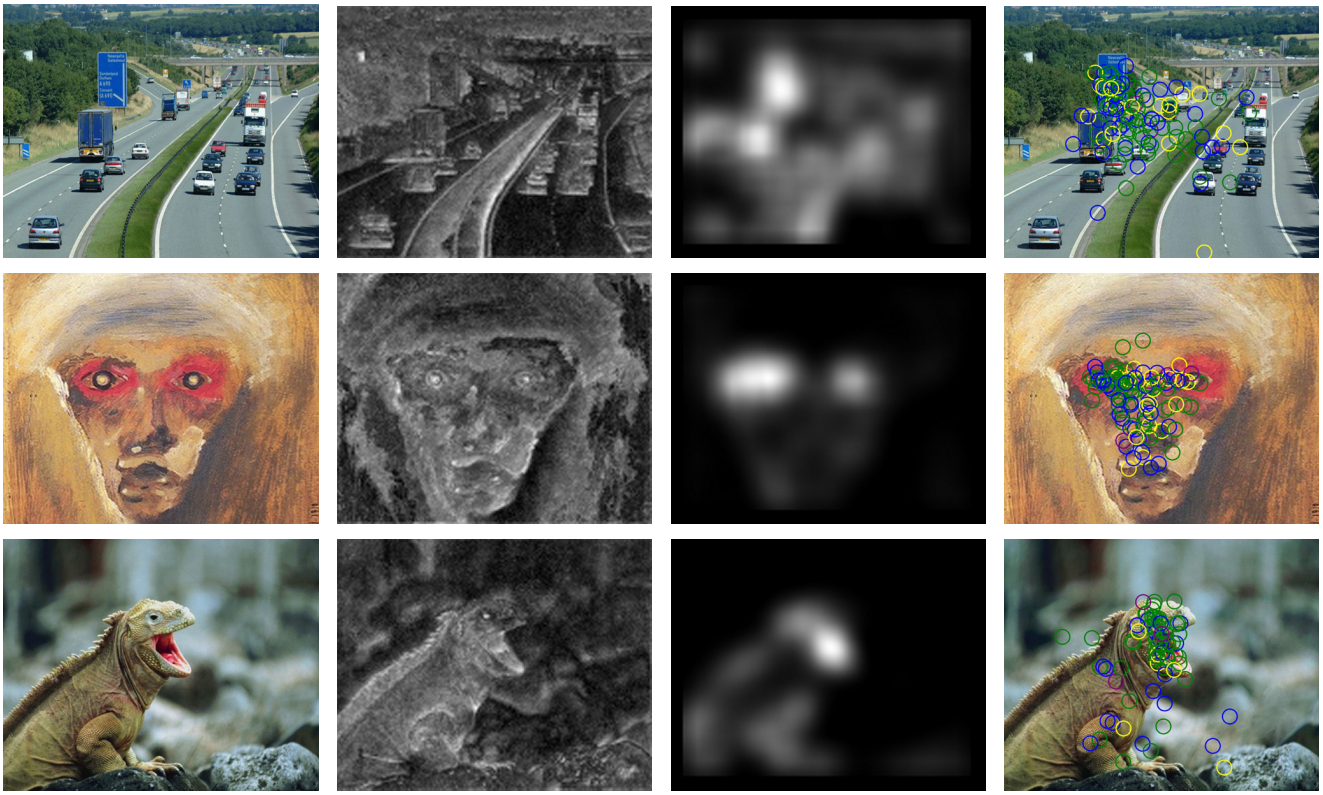
In order to judge the effectiveness of the proposed model of attention, saliency maps computed with the well-established algorithm developed by Itti *et al.* [8] are presented in column (*b*) of Figure 2. The recorded fixations are also drawn on original images in column (*d*). It should be noticed that in the two models, the selection of regions of interest is directed by bottom-up cues. This is why only the locations of the first saccades are considered in order to minimize the top-down influences.

For the natural scene shown in the bottom row, the first recorded fixation points are essentially located on the head of the iguana. The predictions of our model correspond to this location. We can clearly see that the saliency map is strongly correlated with the positions of the fixations of observers. Similar comments hold about the other examples: visual attractors are significantly extracted by the computational model we propose in this paper.

In comparison, regions of interest are not precisely localized in Itti's approach. Some areas of fixation are not extracted by this algorithm when they are clearly stood out by our model. The blue traffic sign along the highway is certainly the most significant example in images presented in this paper. According to the recorded fixation points, the blue traffic sign is undoubtedly a visual attractor for observers during the first saccades and it is pointed out as such by our model but not by the model of Itti *et al.*

5. BRIEF DISCUSSION

Results obtained with the selective attention model we presented in this communication are very encouraging. They demonstrate the effectiveness of our approach. The output saliency map efficiently identifies the bottom-up visual attractors in a natural scene. The first subjective comparisons have to be completed with an objective quantification of the agreement between experimental and computed data. This issue is currently under study.



(*a*) (*b*) (*c*) (*d*)

Figure 2. Typical results for visual stimuli used in experiments. Column (*a*) shows examples of color input images; column (*b*) shows the saliency maps given by the computational model developed by Itti *et al*. [8]; column (*c*) shows the saliency maps provided by our approach; column (*d*) shows the first five fixations of all subjects drawn on original input images.

6. REFERENCES

[1] Canosa, R.L., "Modeling selective perception of complex natural scenes," *International J. on Artificial Intell. Tools* 14 pp. 233-260, 2005.

[2] Carandini, M., and D.L. Ringach, "Predictions of a recurrent model of orientation," *Vis. Res.* 37, pp. 3061-3071, 1997.

[3] Chauvin,, A., J. Hérault, C. Marendaz, and C. Peyrin, "Natural scene perception: visual attractors and images processing," ed.: W. Lowe, and J. Bullinaria, *Connectionist Models of Cognition and Perception*, World Scientific, pp. 236-248, 2002.

[4] CIE 109-1994, "A method of predicting corresponding colours under different chromatic and illuminance adaptations," 1994.

[5] Curcio, C.A., K.A. Allen, K.R. Sloan, C.L. Lerea, J.B. Hurley, I.B. Klock, and A.H. Milam, "Distribution and morphology of human cone photoreceptors stained with anti-blue opsin," *J. of Comparative Neurology* 312(4), pp. 610-624, 1991.

[6] de Valois, R.L., and K.K. de Valois, "A multi-stage color model," *Vis. Res.* 33(8), pp. 1053-1065, 1993.

[7] Duncan, J., and G.W. Humphreys, "Visual search and stimulus similarity," *Psychological Review* 96, pp. 1-26, 1989.

[8] Itti, L., C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pat. Anal. Machine Intell.* 20(11), pp. 1254-1259, 1998.

[9] Maioli, C., I. Benaglio, S. Siri, K. Sosta, and S. Cappa, "The integration of parallel and serial processing mechanisms in visual search: Evidence from eye movement recordings," *European Journal of Neuroscience* 13, pp. 364-372, 2001.

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

[11] McCann, J., "Lessons learned from Mondrians applied to real images and color gamuts," *Proc. IS&T/SID 7 Color Imaging Conference*, Scottsdale, AZ, pp. 1-8, 1999.

[12] Suder, K., and F. Worgotter, "The control of low-level information flow in the visual system," *Reviews in the Neuro-sciences* 11, pp. 127-146, 2000.

[13] Theeuwes, J., "Perceptual selectivity for color and form," *Perception & Psychophysics* 51(6), pp. 599-606, 1992.

[14] Turatto, M., and G. Galfano, "Color, form and luminance capture attention in visual search," *Vis. Res.* 40, pp. 1639-1643, 2000.

[15] Wandell, B.A., *Foundations of vision*, Sunderland, Massachusetts: Sinauer Associates, 1995.

[16] Wolfe, J.M., "Visual search," ed.: H.E. Pashler, *Attention*, London, UK: University College London Press, pp. 13-74, 1998.

[17] Zhaoping, L., "A saliency map in primary visual cortex," *TRENDS in Cognitive Science* 6(1), pp. 9-16, 2002.