WHAT WE SEE IS MOST LIKELY TO BE WHAT MATTERS: VISUAL ATTENTION AND APPLICATIONS

Olivier Le Meur

Thomson R&D
1 av. Belle Fontaine
35776 Cesson Sevigne - France

Patrick Le Callet

IRCCyN, UMR CNRS 6597
Rue C. Pauc
44360 Nantes - France

ABSTRACT

The computational modeling of the visual attention is receiving increasing attention from the computer vision community. Several bottom-up models have been proposed. In spite of their complexities, these models are still a basic description of our visual system. Review of resulting approaches of these efforts are presented in the first part of this paper. Limitations of these approaches are introduced and several research trends are given. Among them, the most important one might be the use of prior knowledge, conjointly with the low-level visual features. Concomitantly with visual attention (VA) modeling progress, the image and video processing community is increasingly considering VA models in different fields or services. Current and future applications of VA models are discussed in the second part.

Index Terms— Visual attention, bottom-up, top-down, Visual attention driven applications

1. INTRODUCTION

Over the last few years, considerable efforts have been devoted to the human Visual Attention (VA). While these efforts were initially and almost exclusively made by psychologist [1, 2], there exists today a wide diversity of work on this topic, leading to a new interdisciplinary research. This new and growing interest involving both fundamental (neuroscience, anatomical, physiological, etc.) and applied (computer vision, image processing, etc.) disciplines is a strength because the problem we are facing is located at the boundaries of various disciplines.

Taking some benefits from the efforts on fundamental understanding of VA, some computational models of the VA have been proposed. The goal of these models is to detect the locations that attract the gaze of an observer. Most of the models compute a saliency map indicating where the most visually interesting parts are located. The quantitative assessment of their performances is still an open issue. However, their predictions, as we will see, are, qualitatively speaking, very convincing.

The capacity to automatically detect the spatial locations of the region of interest provides a great interest for numerous image and video applications. As the introduction of some properties of the human visual system in image and video applications (contrast sensitivity function, visual masking in video compression scheme) has led to successful outcomes during the last two decades, it will be not surprising that future breakthrough in computer vision will be closely related to the use of computational models of visual attention.

The paper is composed of the following sections. In the first part, we present a taxonomy of widespread computational models of visual attention as well as their limitations and some perspectives for improvement. The second part is related to new saliency-based applications and to the role of the saliency in the improvement of existing solutions.

2. COMPUTATIONAL MODEL OF VISUAL ATTENTION

2.1. A taxonomy of computational model of bottom-up visual attention

Since 1998, the year where the first computational and biologically plausible model of bottom-up visual attention was published by L. Itti and C. Koch, there has been a growing interest on the subject. Indeed, several models, more or less biological and based on different mathematical tools, have been investigated. These models can be classified into three different categories (see Table 1 for examples and main properties of models of each categories):

- Hierarchical models (HM) have roughly the same computational architecture, characterized by a hierarchical decomposition, whether it involves a Gaussian, a Fourier-based or wavelet decomposition. Most of the time a difference of Gaussian is applied on the computed subbands to estimate the salience decomposition level. Different techniques are then used to aggregate this information across levels in order to build a unique saliency map;
- Statistical models (SM) are based on a probabilistic framework deduced from the content of the current image. The saliency is then defined as a measure of the deviation between the features of a current location and features present in its neighborhood. It is worth noting that this neighborhood can be local or as large as the image;
- Bayesian models (BM): the Bayesian framework can bring a real benefit since it allows the combination of bottom-up saliency with prior knowledge. This prior knowledge concerns for instance the statistic of visual features in natural scene, its layout or its spectral signature... This is probably one of the most important factors that affects our perception. Prior knowledge coming from our perceptual learning would help the visual system to understand the visual environment and it could be compared to a visual priming effect that would facilitate the scene perception.

All these models are close in term of performance, whatever the assessment technique (e.g. quantitatively or qualitatively); examples of saliency maps for each of models presented in Table 1 are given in figure 1.

Table 1. Main features of computational models of bottom-up visual attention.

HM	**Visual dimension**	**Operations**	**Prior knowledge**
Itti et al. [3]	Intensity, two chromatic channels, orientations, flicker.	Dyadic Gaussian and Gabor pyramids, center/surround filters, peak-to-peak normalization, pooling.	None
Le Meur et al. [4, 5]	Luminance, two chromatic channels, motion.	Oriented subband decomposition in the Fourier domain, Contrast Sensitivity Functions, Masking, center/surround filters, long-term normalization, pooling.	None
Bur et al. [6]	Intensity, two chromatic channels, orientations, motion	Dyadic Gaussian and Gabor pyramids, center/surround filters, long-term normalization, pooling.	None
SM	**Visual dimension**	**Operations**	**Prior knowledge**
Oliva et al. [7]	R, G, B	Saliency of a location is inversely proportional of its occurrence probability in the image. The probability distribution is only based on the statistic of the current image.	None
Bruce et al. [8]	R, G, B	Saliency is based on the self-information computation. Joint probability of the feature, deduced from a given neighborhood.	None
Gao et al. [9]	Intensity, two chromatic channels, orientation, and motion	Dyadic Gaussian and Gabor pyramids, Center/surround filters. Saliency is assessed by using the Kullback-Leibler divergence between the local position and its neighborhood	None
BM	**Visual dimension**	**Operations**	**Prior knowledge**
Zhang et al. [10]	Luminance, two chromatic channels	Saliency is based on the self-information computation.	Probability distribution estimation

Nevertheless, a new and promising trend seems to appear with models based on a Bayesian framework. The combination of information coming from the low-level visual features is a recurrent issue in VA modeling. Such framework is an elegant and promising approach to tackle this issue. In this category one may add the work of Itti and Baldi concerning the theory of surprise [11]. They proposed a formal Bayesian definition of surprise in order to measure the distance between posterior and prior beliefs of the observers. They proved that this measure, the surprise, has the capability to attract human attention. This work is also closely related to the rareness.
The previous models provide a saliency map, i.e a localized representation of salience. For the sake of completeness, there exists another category of models that are based on a hierarchical selection involving winner-takes-all (WTA) processes. To identify the strongest response of a stimulus, a WTA is applied recursively through the hierarchical decomposition. At each level, irrelevant information are pruned or inhibited. In this category, the most famous models have been proposed by [12, 13].

2.2. Far from being sufficient! The road is still very long...

All the computational models of visual attention described in the previous section are still a very basic description of the human vision, based on assumptions that are subjects to discussion. Firstly, the meaning of saliency map from a biological viewpoint can be raised, secondly the importance of cognitive processing is still not clearly considered neither identified.
The first question concerns the existence of a locus in the brain where a unique saliency map would be located. Numerous evidences suggest that this unique locus does not exist. The concept of saliency map would be more of an abstract representation, updated at each computational level of the brain [14]. This update would take into account information coming from the low-level visual features but also from our knowledge, our memory, our expectation... Notice that Fecteau and Munoz [15] introduced the concept of priority map. such map is a combined representation of bottom-up and top-down salience. The concept of a unique saliency (or priority) map that would control the gaze is however a comfortable situation in a computational modeling point of view. With such condition, the brain is compared to a computer where the inputs come from our different senses and the knowledge is in the memory. This is obviously a limited viewpoint [16].
The second point relies on the idea that visual and cognitive processes are strongly tied. It is currently impossible to evaluate at a given time the extent to which they influence the deployment of the visual attention. The use of eye tracking experiments to assess the performances of computational model of bottom-up visual attention is therefore an issue. Even though eye movements are an overt behavioral manifestation of the allocation of the attention in a given scene, the top-down contribution to this manifestation can not be ruled out. Some heuristics are used during the experiments to attempt to lessen these contributions. For instance, in a free-viewing experiment, no task is given; the only instruction given to observers is to look at the scene as naturally as possible. However, there is no certitude that this type of sentence does not induce or influence the behavior of the observer. To go one step further, an even stronger statement could be made. The eye tracker records eyes movements and the location of the foveal vision when the eyes fixate. This is achieved by tracking our fovea. During such an approach, the role of our peripheral vision is definitively overlooked. Characterizing attention by using such technology might turned out to be a rough measure. It should not be forgotten also that attention does not require eye movements at all. This kind of attention is called covert attention [17], *attending while looking elsewhere*.
Recent advances in technology will probably allow for a deeper understanding of how the visual and cognitive processes interact. For instance, Baccino et al. [18] have quite recently combined eye movements and even-related potentials (ERP). This technique is called eye-fixation-related potentials (EFRP) and is currently used to examine the factors acting on a reading task. In a near future, this approach could be used in order to evaluate the influence of the perceptual, attentional and cognitive factors during a visual scene ex-

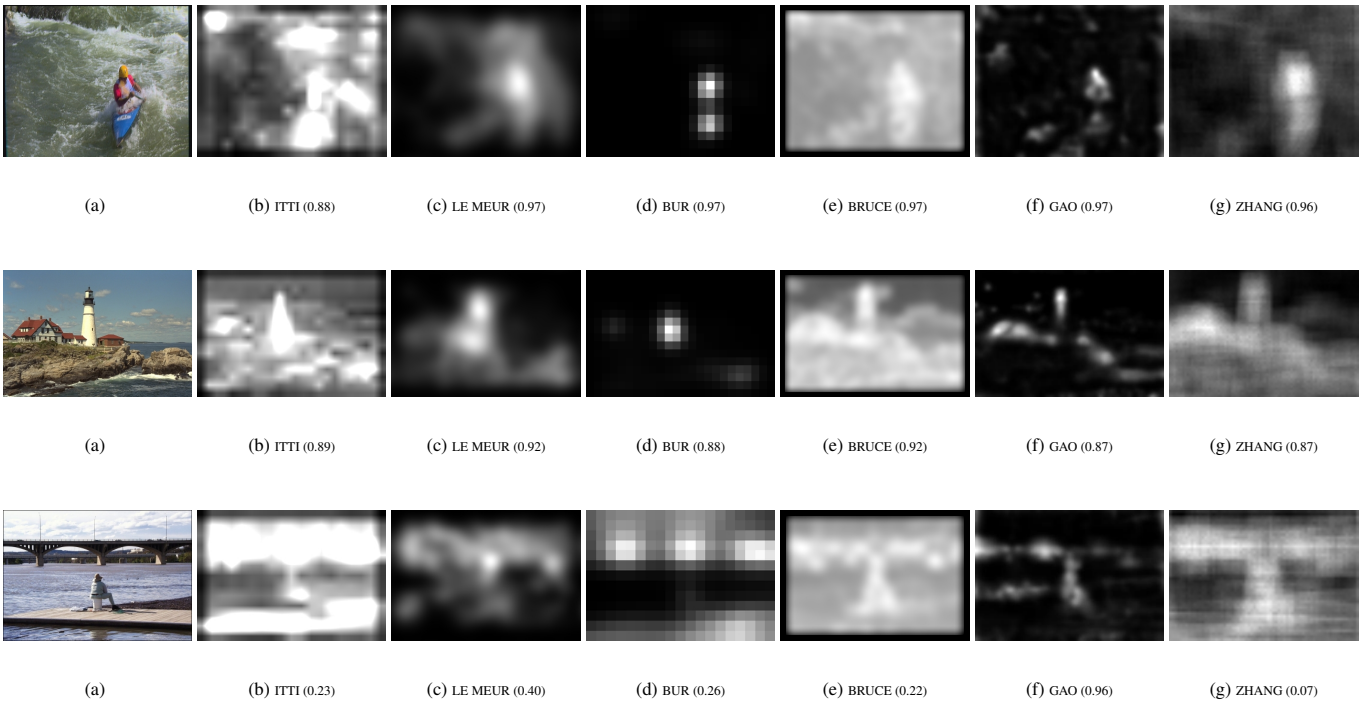

Fig. 1. Saliency maps for different computational models of bottom-up visual attention. (a) original pictures and thier associated saliency maps stemming from the model of Itti (b), Le Meur (c), Bur (d), Bruce (e), Gao (f) and Zhang (g).The number between parenthesis is the area under curve in a ROC Analysis. It involves a predicted saliency map and a priority map deduced from eye tracking experiments (see [4] to have a description of the eye tracking experiment).

ploration.

3. APPLICATIONS

The capability to predict the location onto which an observer will focus his attention presents a strong interest, and this, for numerous video applications. The first attempts to take benefit of VA have been applied to image and video compression based on the idea that non-important areas are subjected to higher compression than more relevant areas [19]. Recently, new applications have been considered:

- Quality assessment: the idea relies on the fact that an artifact appearing on a salient region is more annoying than an artifact appearing in the background. Even though that it seems intuitively correct, first experiments do not succeed in significantly improving the performance of a quality metric [20, 21];
- Re-framing or re-targeting: the goal of a re-framing application is either to enhance the viewing experience when an initial high resolution content has to be displayed on a small screen or to ease the browsing a large collection of picture. It consists in extracting a sub-part of the picture centered on the regions of interest [22, 23]. Figure 2 gives an illustration;
- Channel coding: unequal error protection can be driven by a VA model, applying better resilience in ROI [24].

Looking further, other original recent works could be cited addressing applications related to super resolution [25], user interface design [26], augmented reality [27], automated guidance of robots [28]. All this activity is witnessing the current enthusiasm of the community.

4. CONCLUSION

The first computational models of visual attention rest mainly on the use of low-level visual features. Some of these models are able to mix low and high-level information, however, this is always aimed at tackling specific tasks such as pedestrian detection for instance. We are at the beginning of a new 'generation' of computational models that will use these higher-level information, also called prior knowledge. Mainly based on the perceptual learning, prior knowledge might act on our visual system as a priming effect. Even if there is no consensus concerning the components of the prior knowledge, it is reasonable to think that the understanding of the scene as well as object recognition are two pillars of this knowledge. For instance, in the context of scene recognition, several studies have shown that the categorization of a complex natural scene can be achieved in 100-150 ms [29]. This ability to determine quickly whether an image represents an animal or a person has likely an impact on the deployment of our visual attention. The questions are how can we measure this impact and how can we reproduce it? All these potential improvements will undeniably have an impact on the efficiency of numerous video applications in line with the current trends in image and video processing community. Nevertheless, even if some authors have already demonstrated the potential VA models represent, this is probably just the beginning.

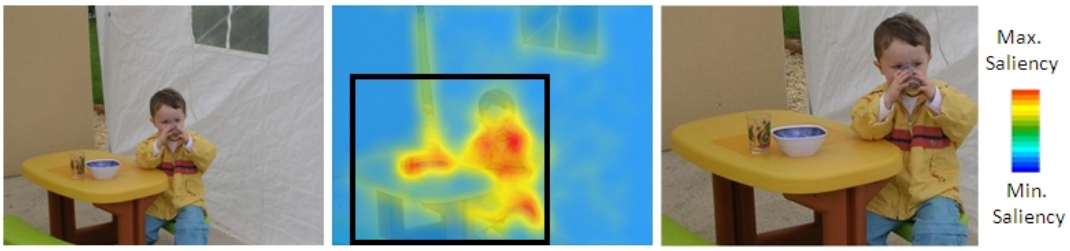

Fig. 2. Example of re-targeting approach: on the left-hand side, the original picture; on the middle, the heat map (computed by Le Meur's model [4]). The red box corresponds to the cropping window; On the right-hand side, the reframed picture.

5. REFERENCES

[1] W. James, "The principles of psychology," *New York: Holt*, 1890.

[2] A. L. Yarbus, "Eye movements and vision," *New York: Plenum Press*, 1967.

[3] L. Itti and C. Koch, "Model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[4] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model the bottom-up visual attention," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 802–817, 2006.

[5] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Research*, vol. 47, no. 19, pp. 2483–2498, 2007.

[6] A. Bur and H. Hugli, "Optimal cue combination for saliency computation: A comparison with human vision," .

[7] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *IEEE International Conference on Image Processing*, 2003, vol. 1, pp. 253–256.

[8] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, pp. 1–24, 3 2009.

[9] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of Vision*, vol. 8, no. 7, pp. 1–18, 6 2008.

[10] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, pp. 1–20, 12 2008.

[11] L. Itti and P.F. Baldi, "Bayesian surprise attracts human attention," in *Advances in Neural Information Processing Systems*, Cambridge, MA, 2006, pp. 547–554, MIT Press.

[12] J.K. Tsotsos, "Analyzing vision at the complexity level," *Behavioral and Brain Sciences*, vol. 13, no. 3, pp. 423–445, 1990.

[13] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annu. Rev. Neurosci.*, vol. 18, pp. 193222, 1995.

[14] R. VanRullen, "Visual saliency and spike timing in the ventral visual pathway," *Journal of Physiology*, vol. 97, no. 13, pp. 365–377, 2003.

[15] J.H. Fecteau and D. P. Munoz, "Salience, relevance, and firing: a priority map for target selection," *Trends in cognitive sciences*, vol. 10, pp. 382–390, 2006.

[16] J. Brockman, *The Third Culture: Beyond the Scientific Revolution*, Simon-Schuster, 1995, Chapter 12: The Emergent Self by F. Varela.

[17] M.I. Posner, "Orienting of attention," *Q. J. Exp. Psychol.*, , no. 32, pp. 3–25, 1980.

[18] T. Baccino and Y. Manunta, "Eye-fixation-related potentials: insight into parafoveal processing," *Journal of Psychophysiology*, vol. 19, no. 3, pp. 204–215, 2005.

[19] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. on Image Processing*, vol. 13, no. 10, pp. 1304–1318, Oct 2004.

[20] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Does where you gaze on an image affect your perception of quality? applying visual attention to image quality," in *IEEE International Conference on Image Processing*, 2007.

[21] E.C. Larson, V. Cuong, and M. Chandler, "Can visual fixation patterns improve image fidelity assessment?," in *IEEE International Conference on Image Processing*, 2008.

[22] X. Fan, X. Xie, W.Y. Ma, H.J. Zhang, and H.J. Zhou, "Visual attention based image browsing on mobile devices," in *IEEE International Conference on Multimedia & Expo.*, 2003, pp. 53–56.

[23] C. Chamaret and O. Le Meur, "Attention-based video reframing: Validation using eye-tracking," in *IEEE International Conference on Pattern Recognition*, 2008.

[24] T. Liu, X. Feng, A. Reibman, and Y. Wang, "Saliency inspired modeling of packet-loss visibility in decoded videos," in *International Workshop VPQM*, 2009.

[25] N.G. Sadaka and L.J. Karam, "Perceptual attentive superresolution," in *International Workshop VPQM*, 2009.

[26] H. Liu, S. Jiang, Q. Huang, and C. Xu, "Ba generic virtual content insertion system based on visual attention analysis," in *ACM Multimedia*, 2008, pp. 379–388.

[27] F. Biocca, A. Tang, C. Owen, and F. Xiao, "Attention funnel: Omnidirectional 3d cursor for mobile augmented reality platforms," in *conference on Human Factors in computing systems*, 2006, pp. 1115–1122.

[28] C. Siagian and L. Itti, "Biologically-inspired robotics vision monte-carlo localization in the outdoor environment," in *Conference on Intelligent Robots and Systems*, Oct 2007.

[29] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, pp. 520–522, 1996.