Modeling visual attention on scenes

Brice Follet (*) (***) — Olivier Le Meur (**) — Thierry Baccino (***)

(*) Thomson, Compression Lab.
1 av. de Belle Fontaine 35576 Cesson-Sevigne France
brice.follet@thomson.net
(**) University of Rennes 1
Campus universitaire de Beaulieu 35042 Rennes France
olemeur@irisa.fr
(***) LUTIN (UMS-CNRS 2809), Cité des sciences et de l'industrie de la Villette
30 av. Corentin Cariou 75930 Paris
baccino@lutin-userlab.fr

ABSTRACT. *Research in the computational modelling of the visual attention has mushroomed in recent years. First generation of computational models, called bottom-up models, allows to calculate a saliency map indicating the degree of interest of each area of a picture. These models are purely based on the low-level visual features. However, it is now well known that the visual perception is not a purely bottom-up process. To improve in a significant manner the quality of the prediction, top-down information (prior knowledge, expectations, contextual guidance) have to be taken into account. We propose in this article to describe some bottom-up models and the metrics used to assess their performances. New generation of models based both on low-level and high-level information is also briefly described. To go one step further in the understanding of the cognitive processes, new computational tools have been recently proposed and are listed in the final section.*

RÉSUMÉ. *La modélisation computationelle de l'attention visuelle connaît actuellement un essor considérable. Les premières modèles, purement basés sur l'attention dite exogène, permettent de calculer une carte de saillance indiquant les zones d'intérêt visuel d'une image. Cependant, afin d'améliorer cette prédiction, il s'avère nécessaire de prendre en compte des informations de plus haut niveaux relatives à l'attention endogène, c'est à dire des informations liées aux processus cognitifs. Afin de rendre compte de cette problématique, le présent article décrit un certain nombre de modèles exogènes ainsi que des modèles intégrant de la*

connaissance a priori. Les méthodes d'évaluation des performances sont également décrites. Afin d'aller plus loin dans la modélisation et dans la compréhension des processus cognitifs, de nouvelles perspectives et direction d'études sont exposées.

KEYWORDS: *Attention, eye movements, saliency map, scanpath, bottom up, top down, computational modeling, EFRP, eye tracking.*

MOTS-CLÉS : *Attention, mouvements oculaires, carte de salience, parcours visuel, mécanismes endogène exogène, modélisation computationnelle, EFRP, suivi du regard.*

1. Intoduction

The computational modeling of visual attention is an important challenge for computer vision researchers. Potential applications are large and various (video compression, surveillance, retargeting...). The targeted goal is to predict in an automatic manner from an input picture or video sequence the locations where an observer would gaze on. For that, most of the computational models rest on the assumption that there exists a single saliency map in the brain. A saliency map is a map indicating the degree of interest of each area. However, it seems that there is no single saliency map in the brain but rather a set of saliency maps that are distributed throughout the different visual areas of the brain. The assumption of unique saliency is then a strong shortcut but a very comfortable view to design a computational model. Indeed, in this condition, the brain can be compared to a computer where the inputs are the sensory information and the output the saliency map.

Based on the assumption that there is a unique saliency map in the brain, Koch and Ullman proposed in 1985 a plausible architecture of a visual attention model [KU85]. From an input picture, several early visual features are extracted in a massively parallel manner, leading to one feature map per channel. A filtering operation is then applied on these maps in order to filter out most of the visually irrelevant information. Then, these maps are mixed together to form a saliency map. From this influential work, a number of computational models have been proposed. They can be grouped into two categories that could be called hierarchical models and probabilistic (or statistical) models. The architecture of these two categories of models is almost the same; they differ in the mathematical operations used to determine the salience.

This first generation of models rests mainly on bottom-up mechanisms. Early visual stages are indeed purely bottom-up. However, a number of other features such as prior knowledge, our expectation, goals we have to perform may have a significant impact on the way we scan our visual field.

In this paper, we propose to tackle these aforementioned points. The first part concerns a brief review of bottom-up models. We also briefly review methods to assess the degree of similarity between predicted and human data. In the second point, we emphasize the factors that a second

generation of computational models of visual attention should take into account. In this same part, some models based on prior knowledge are described. In the last part, new perspectives as well as new computational tools are described.

2. Computational models

2.1. *Hierarchical models*

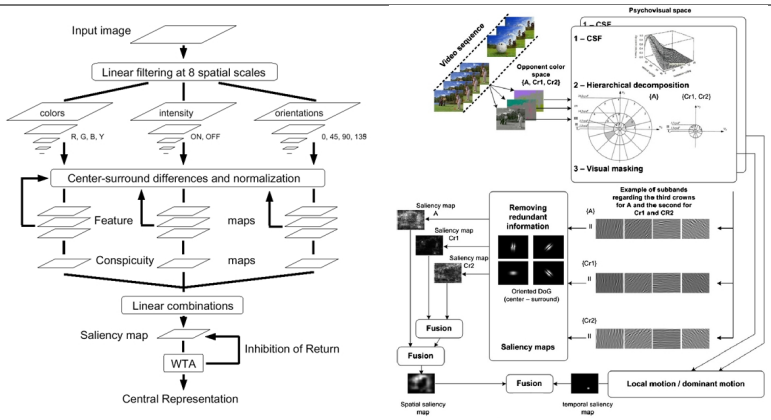
Figure 2.1 (a) gives the architecture of Itti's model [IKN98]. This computational model, proposed in 1998, is the first model able to compute efficiently a saliency map from a given picture. The saliency map is built from 3 separate feature maps coming from 3 channels (color, orientation and luminance).
As depicted in figure 2.1 (a), a Gaussian pyramid is used for each channel to split the input signal into several spatial scales. Center-surround differences are used to yield a feature map. The final saliency map is then computed. A map normalization operator $\mathcal{N}(.)$ is defined. The definition given by Itti et al. is the following operation:

1) Normalizing the values to a fixed range in order to eliminate modality-dependent amplitude differences;

2) Finding the location of the map's global maximum and computing the average of all its other local maxima;

3) Multiplying the map by a factor depending on the previous values.

Figure 2.1 (b) gives an extension of Itti's model. Le Meur's model [MCBT06] improves two aspects of Itti's model: first, the extraction of the early visual features in Itti's model is achieved by different operators that can provide dramatically dynamic range. To cope with these problems, several normalization schemes are used many times at different stages of the model in order to eliminate modality-dependent amplitude differences. Second, the building of the final saliency map is the result of linear combination. The feature maps are first normalized in the same dynamic range and then combined.



(a) Itti's model (b) Le Meur's model

Figure 1: Two examples of hierarchical model of bottom-up visual attention

To deal with the two aforementioned points, Le Meur et al. proposed to extract the early visual features by using psychophysic models. A contrast sensitivity function is used to normalize incoming information to their visibility threshold. Perceptual subband decomposition is then applied on the Fourier spectrum. Each resulting subband can be seen as a population of visual cells that is sensitive to a range of orientations and to a range of spatial frequencies. Finally, a visual masking operation allows modulating the visibility thresholds of a given subband by taking into account information coming from other subbands. The goal of these first steps is to provide homogeneous feature maps that can be directly compared. At this stage, a single normalization is required.
A center-surround operation is then applied on each subband in order to eliminate irrelevant information and to promote contrasted areas. After that, a unique saliency map is computed by using a normalization scheme, called long-term normalization.

2.2. *Statistical models*

Statistical models differ from the previous ones because they rest on a probabilistic approach. The assumption is that a rare event is more

salient than a non rare event. The mathematical tool that can simply simulate this behavior is the self-information. Self-information is a measure of the amount information provided by an event. For a discrete random variable X, defined by $\mathcal{A} = \{x_1, ..., x_N\}$ and by a probability density function, the amount of information of the event $X = x_i$ is given by $I(X = x) = -log_2 p(X = x)$ bit/symbol.
The first model based on this approach has been proposed by Oliva et al. [OTCH03].Bottom-up saliency map is given by: $S = \frac{1}{p(F|G)}$, where F denotes a vector of local visual features observed at a given location while G represents the same visual features but computed over the whole image.

More recently, this approach has been reused by a number of authors. The proposed models differ in the support used to compute the saliency:

– The probability density function is learnt on the whole picture [OTCH03];

– A local surround is used to compute the saliency. The saliency can be either be computed by using the self-information [BT09] or the mutual information [GV09];

– A probability density function has been learned on a number of natural image patches. Features extracted at a given location are compared to this prior knowledge [ZTM$^+$08, BT09];

– A probability density function is learnt from what it happened in the past. This is the theory of surprise proposed by Itti and Baldi [IB06]. This approach allows to compute how incoming information affect our perception. The idea is to compute the difference between posterior and prior beliefs of observers.

2.3. *Performances*

The assessment of computational models is commonly performed by assessing the degree of similarity between predicted saliency maps and data coming from eye tracking experiments. Eye tracking experiments are commonly conducted in a free-viewing task. It means that observers were allowed to examine picture freely without any particular objective.

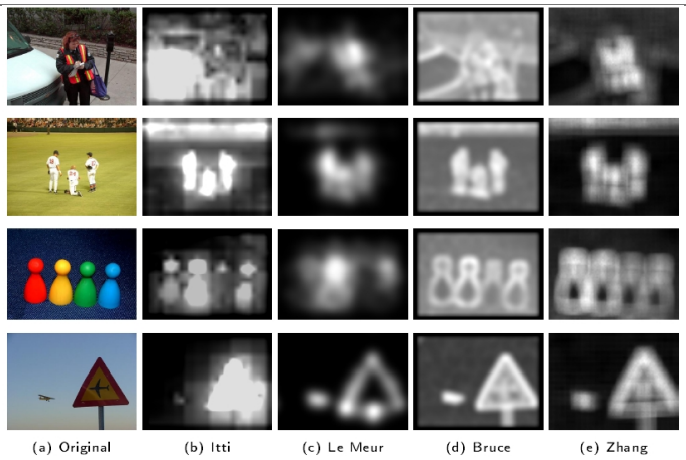


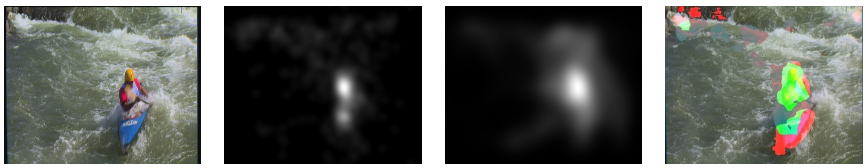
Figure 2: Predicted saliency maps for different original pictures and for different computational models.

The goal is to lessen top-down feedback and to favor a bottom-up gaze deployment.
The collected raw eye tracking data is then parsed into fixations and saccades. A velocity-based algorithm can be used to perform this parsing(see [SG00] to have more details and to have a taxonomy of parsing methods). Several databases exist on internet (see for instance, *http://www.irisa.fr/temics/staff/lemeur/*).
The degree of similarity between the prediction of a computational model and the ground truth can be obtained by two methods. These methods, called saliency-map-based and fixation-point-based method are described in the following subsections.

2.3.1. *Saliency-map-based method*

The saliency-map-based method, as its name suggests, rests on the use of map. Therefore, from the collected eye-tracking data, a human saliency map is computed by taking into account fixation points of all observers. The building of this kind map is described in [Woo02]. Figure 2.3.1 gives an example of human and predicted saliency map.
The degree of similarity between these two maps can be assessed by different methods. However, the most used and probably the most rele-

vant is the ROC (Receiver Operating Characteristic). The idea is to use a binary classifier in order to label each pixel as being fixated or not. Different thresholds are used. For each threshold the true positive rate (TPR) and the false positive rate (FPR) is deduced (an example is given 2.3.1 (d)). A ROC graph depicting the tradeoff between TPR and FPR is plotted. The TPR rate is plotted on the Y axis whereas the FPR rate is plotted on the X axis. On this graph, the point (0,1) represents a perfect similarity. The more the top left-hand corner the curve is close, the better the classification is. The diagonal line (if it is plotted on a linear-linear graph) indicates that the classification is a pure random process. One inter- esting indicator is the area under the curve, called AUC. This indicator is indeed very useful to compare the quality of the prediction. The AUC value is always between 0 and 1. An AUC value equal to 0.5 indicates that there is no similarity between the two sets of data. A value of 1 is obtained for a perfect classification.



(a) Original (b) Exp. SM (c) Predicted SM (d) ROC

Figure 3: Example of human saliency map (b) and predicted saliency map (a). (c) correponds to a ROC classification (example for a given threshold): original pixels stand for true negative, green pixels for true positive, the others for false and true negative.

2.3.2. *Fixation-point-based method*

Rather than considering a saliency map, this method uses the list of the human fixation points [PLN02, PIIK05]. Predicted saliency for each fixation point is extracted from the predicted saliency map at the spatial coordinates of the considered fixation point. At the end of the list of fixation points, a value, called NSS standing for Normalized Scan path Saliency, is obtained. A NSS equal to zero means that there is no similarity between the predicted saliency map and the visual scan paths.

A positive NSS indicates that there is a similarity whereas a negative one indicates an anti-correspondance. Figure 2.3.2 gives an example. Below, the three steps of the NSS method are given:

1) Each saliency map is transformed into a Z-score (normalized to zero mean and one unit standard deviation);

2) Predicted saliency is extracted in a local neighboorhood centered on a human fixation point;

3) Average of the obtained values.

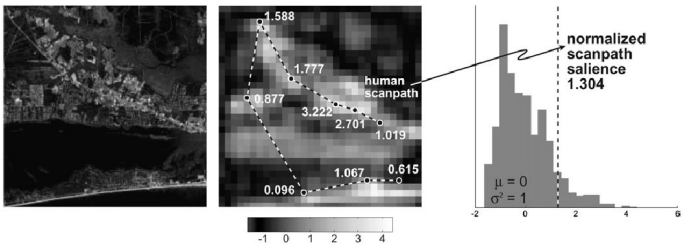


Figure 4: NSS computation (extracted from [PIIK05].)

3. The necessity of cognitive models

A purely contrast guidance-based process is a basic view of how our attention is deployed. A second form of attention controlled by higher areas of the brain can override, or at least can influence, bottom-up process.

3.1. *Top-Down priming*

To improve in a significant manner visual attention models, top-down source guidance of visual attention should be taken into account. Numerous studies indeed show that the attentional deployment is driven not by a single source but rather by a set of guidance sources. The seminal experiment of Yarbus [Yar67] is the perfect illustration of the strong relationship existing between bottom-up and top-down mechanisms. Yarbus recorded eye movements when observers watched

a painting. Observers were given different questions to answer prior viewing the painting. The fact that the recorded visual scan paths were radically different simply demonstrated that the visual attention deployment is not a purely bottom-up process. Since 1967, more and more experiments strengthen and refine Yarbus's conclusions (see [**?**] for a nice and short review).

3.2. *Two visual pathways*

Behavioural researches have proposed many functional splitting like evaluating versus orienting, what and where processes, focal and ambient treatment, examining against noticing computation, figural/spatial and foveal-ambient to conceptualize a dual processing which reconciles opposed classical theoretical approaches of behavioural sciences. Functional neuroanatomy has confirmed this functional dichotomy in showing two distinct visual pathways working in parallel. These findings support idea that cortex computes visual information by separating object identification to scene spatialization. These two ways start from earlier visual areas (V1, V2) with a dorsal pathway directed to posterior parietal cortex to compute spatial attributes of scene. While ventral pathway which is responsible of objects identification is directed to infero-temporal cortex [Bar04].

3.2.1. *A dichotomy in spatial frequencies*

Thorpe et al. [TFM96] have shown that visual system is able to categorize visual scene in 160 ms. This very rapid processing, probably due to a feed-forward mechanism, could be explained by the brain ability to categorize rapidly visual scene from low frequencies information [SO94].
Indeed, many studies claimed that visual processing follows a coarse-to-fine process. This coarse-to-fine process seems to rely on the processing of low versus high spatial frequencies of visual contrast. The coarse process might be based on low frequencies contrary to the fine process based on high frequencies. Thus, it has been shown that fast categorization of visual scene was related to the processing of low frequencies

[SO94] while finer information given by high frequencies corresponded to time-consuming processes for object recognition [Bar04]. Emotional information in face [SO99] and maybe in scene perception is provided by low frequencies but certainly across under cortical pathways. The gist is fundamental for scene recognition [OT06]. It should correspond to the low frequency spatial invariant of each category [Oli05]. These findings brought the idea that a dorsal pathway could be specialized in low frequency coarse processing through the magnocellular dorsal pathway while high frequency fine processing might be related to the parvocellular ventral pathway [Bar04]. Others studies suggest that functional distinction corresponds to an asymmetry between both cerebral hemispheres. However, behavioural data support a dual model based on functional dichotomy and Baddeley and Tatler (2006) have shown that high frequency edges can predict fixation locus. So, the actual matter is to figure out whether this functional parallelism is able to influence attentional guidance.

3.2.2. *A dichotomy in the visual scan path*

Yarbus's experiments [Yar67] indicated that fixation duration increases during scene viewing involving a strong attentional focus. This result has been confirmed by recent studies on eye-movements under various conditions (different tasks like searching, recognition [USP05], memorization [CMH09] or free viewing [PHR$^+$08]. These same experimentations also indicated a decreasing of amplitude saccade according to time viewing.
An important aspect outlined by Unema's study [USP05] concerns the relationship between the fixation duration and the amplitude of the subsequent saccade. Their graph (see figure 3.2.2) shows clearly that larger saccade amplitude corresponded to fixation duration shorter than 200 ms and smaller saccade amplitude to fixation duration greater than 200ms. This relationship across fixation and saccade has allowed them in categorizing ambient and focal fixations. Ambient fixations have been defined by large saccades and short fixation durations and the other way around for focal fixations. These two kinds of fixations/saccade pairs reveal a coarse-to-fine eye movement strategy [OHVE07] in accordance to the cerebral visual processing modes [PV09].
To know whether ambient-to-focal strategy was correlated to a coarse-

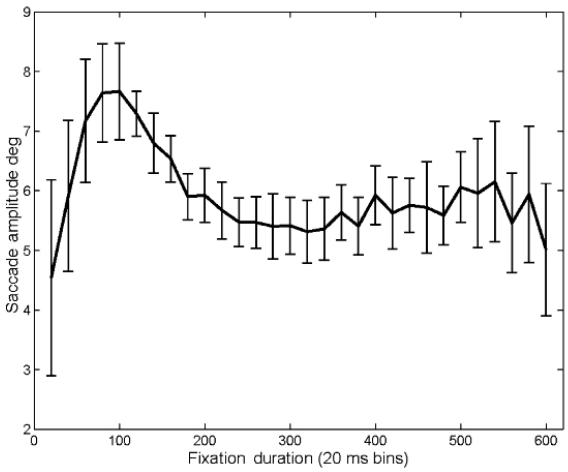


Figure 5: Saccade amplitude as a function of fixation duration (extracted from [USP05]).

to-fine process or not, Follet et al. [FMB09] conducted an eye tracking experiment with hybrid stimuli with four different visual scene categories (Mountain, Street, Coast, Open Country). Hybrid pictures were built with stimuli mixing low frequencies of an image with high frequencies of another. The principle was to compare visual scan pattern (i.e. fixation duration/amplitude saccade relationship). They also measured for each category gaze allocation similarities between original and hybrid pictures. The idea is to assess the extent to which low and high spatial frequencies contribute to the guidance of eye movements. Results indicated firstly that eye movements were more or less driven by low or high frequencies and secondly they were dependent on the visual scene category. These issues suggested a relationship between the processing of low frequencies and the ambient-to-focal strategy.

3.3. *Cognitive models of visual attention*

Most of researchers are now convinced that high-level factors coming from cognitive operations should have a bigger importance in visual guidance [HBCM07]. Saliency map models described in section

2 are very efficient to determine the early fixations during scene viewing but they are weaker to simulate subsequent fixations when semantic information becomes more essential. In fact, fixation positions are less strongly tied to visual saliency when meaningful scenes are involved [OTCH03]. Gaze control becomes knowledge-driven over time which replaces or modulates visual saliency effects. But what kind of knowledge is involved in this visual guidance:

– Memory scene knowledge: This memory can involve either episodic knowledge or semantic knowledge. Episodic memory stores personal events (times, places...) that can be explicitly stated. The use of that memory in visual guidance can describe why we restrict the region of search when we are looking for a specific object in a well known scene. Semantic memory can be related to the concept of schema or frames defended earlier by Minsky (1974) [Min74]. They represent generic semantic and spatial knowledge about a particular type of scene which includes information about the objects likely to be found (a bed in a bedroom), spatial regularities (a painting on a wall) and real world knowledge about scenes (logs can float in the river). These examples show that long-term memory affect the gaze control but not only, short term memory as well. For example, short term knowledge can explain why a viewer can refixate regions of the current scene that are semantically informative or interesting [LM78];

– Task-related knowledge: There is a lot of evidence showing that tasks involve selective fixations since the Yarbus' experiment. Gaze control changes with reading, driving or searching. Even saliency maps are sensitive to the task and simulate rather recognition experiments than search experiments [UF06].

Modeling these types of high-level information is not trivial and the question arises over how it can be combined with low-level factors (saliency maps). One approach is to consider that knowledge structures modify the bottom-up saliency [RZB02]. For example, looking for a car in a street does not need to look at the top of building. Some prior information about a scene maintained in long-term memory serves to constrain the visual field of search and consequently the effect of saliency map. This notion has given the basis of Contextual Guidance model that highlight scene regions likely to contain a specific class of object

[TOCH06]. This model predicts fixation positions during object search significantly better than a saliency map model. However, memory as well as task knowledge is very complex to integrate in a computational model given the lack of accurate definition. While Henderson and others have only built theoretical models avoiding coding the semantic information, Navalpakkam and Itti [NI05] attempted to create a semantic memory by means of ontology. This solution is far from being satisfying since they had first to manually segment the objects and labelize them to be included in the ontology. The ontology is therefore not exhaustive and needs to be updated for each new stimulus. In the future, a challenge would be to associate automatically objects with their visual context as it is used in text processing by Latent Semantic Analysis [LFL98].

4. Perspective regarding the visual attention deployment

All these questions on computational models of visual attention stand on the status of eye fixation and whether it is possible to identify the main factors determining that fixation. Broadly speaking models have to predict where and how long do we look a visual scene (i.e, spatial and temporal determinants of the fixation).
As we have seen in section 2, both scene-statistics and saliency maps approaches have attempted to describe the location of fixations rather than their durations. However, fixation locations can also be related to the meaning of a fixated region since meaningful object may likely differ from scene background in image properties. Human gaze is under the control of the brain that process data not only on available visual input but also according to its mental structures (memory, goals, plans, beliefs...). As an example of that semantic influence, Henderson et al [HP08] shown that fixation locations differed in both their image statistics and their semantic content. One of the future challenges for computational modeling of visual attention is therefore to address that semantic information and investigate how it might be implemented. A promise way would be to consider another characteristic of gaze behavior, namely the variability of fixation durations.
This variable has been largely ignored on gaze control while in other activity, such reading, fixation durations represent the main measure

of cognitive processing. Thus, it has been shown that fixation duration is affected by lexical frequency, predictability and semantic content [Ray98]. Obviously, looking at a visual scene does not involve the same processes than reading which entails that fixation duration on reading reflects a large part of the processing on the fixated region. However, this effect has also been found when viewing a visual scene [HP08]. Average fixation duration during scene viewing is around 330 ms [Hen07]. Visual scenes may contain several meaningful objects and that meaning modulates the duration of fixations [USP05]. There are at least two possibilities for using fixation duration in a computational model of visual attention:

1) By weighting fixation position with fixation duration. However, several open questions are still pending: whether this weighting may be applied linearly or not; what are the main semantic factors that modulate fixation duration on object viewing; how to define semantics for an object, by its own meaning, by the context of use or probably by both factors. This approach is still a challenging direction in research;

2) By separating different components of fixation duration and associating them to specific functions. One way which has been already explored is by combining the eye-fixations with other physiological variables such as EEGs [Bac10]. This technique firstly introduced by Baccino and Manunta [BM05] has been called Eye-Fixation-Related potentials and consisted to analyze EEGs only during the fixation phases. One of the main advantages with this technique is that the interpretation of fixation duration can be enriched by several components coming from the separation of the EEG signal. That separation can be carried out by statistical procedures (PCA, ICA, PARAFAC) in order to detect whether any attentional or semantic components may be associated to the fixation and categorized it.

Modeling visual attention on scenes has received a great interest by cognitive scientists in the last decade. Purely visual models at the beginning, the implementation of neurophysiological and cognitive information has been progressively added over the years. But the way is still long before having a satisfying model that might be cognitively plausible.

References

[Bac10] T. Baccino. Eye movements and concurrent erp's: Efrps investigations in reading. *In S. Liversedge, Ian D. Gilchrist & S. Everling (Eds.), Handbook on Eye Movements. Oxford University Press*, 2010.

[Bar04] M. Bar. Visual objects in context. *Nature Revision Neuroscience*, 5:617–629, 2004.

[BM05] T. Baccino and Y. Manunta. Eye-fixation-related potentials: Insight into parafoveal processing. *Journal of Psychophysiology*, 19(3):204–215, 2005.

[BT09] N.D.B. Bruce and J.K. Tsotsos. Saliency, attention and visual search: an information theoretic approach. *Journal of Vision*, 9:1–24, 2009.

[CMH09] M. Castelhano, M. Mack, and J. M. Henderson. Viewing task influences eye movements control during active scene perception. *Journal of Vision*, 9:1–15, 2009.

[FMB09] B. Follet, O. Le Meur, and T. Baccino. Relationship between coarse-to-fine process and ambient-to-focal visual fixations. In *ECEM, http://www.irisa.fr/temics/staff/lemeur/*, 2009.

[GV09] D. Gao and N. Vasconcelos. Bottom-up saliency is a discrimant processe. In *IEEE ICCV*, 2009.

[HBCM07] J. M. Henderson, J.R. Brockmole, M.S. Castelhano, and M. Mack. Visual saliency does not account for eye movements during visual search in real-world scenes. In *In R. P. G. van Gompel, M. H. Fischer, W. S. Murray & R. L. Hill (Eds.), Eye movements: A window on mind and brain. Amsterdam, Netherlands: Elsevier*, 2007.

[Hen07] J.M. Henderson. Regarding scenes. *Current Directions in Psychological Science*, 16(4):219–222, 2007.

[HP08] J. M. Henderson and G. Pierce. Eye movements during scene viewing: Evidence for mixed control of fixation durations. *Psychonomic Bulletin and Review*, 15(3):566, 2008.

[IB06] L. Itti and P. Baldi. Bayesian surprise attracts human attention. In *Advances in neural information processing systems*, 2006.

[IKN98] L. Ittti, C. Koch, and E. Niebur. A model for saliency-based visual attention for rapid scene analysis. *IEEE Trans. on PAMI*, 20:1254–1259, 1998.

[KU85] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[LFL98] T.K. Landauer, P.W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284, 1998.

[LM78] G.R. Loftus and N. H. Mackworth. Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 4:565–572, 1978.

[MCBT06] 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 PAMI*, 28, 2006.

[Min74] M. Minsky. A framework for representing knowledge. In *The Psychology of Computer Vision, P. H. Winston (ed.), McGraw-Hill 1975*, 1974.

[NI05] V. Navalpakkam and L. Itti. Modeling the influence of task on attention. *Vision Research*, 45:205–231, 2005.

[OHVE07] E.A.B. Over, I.T.C. Hooge, B.N.S. Vlaskamp, and C.J. Erkelens. Coarse-to-fine eye movement strategy in visual search. *Vision Research*, 47:2272–2280, 2007.

[Oli05] A. Oliva. Gist of the scene. *In the Encyclopedia of Neurobiology of Attention. L. Itti, G. Rees, and J.K. Tsotsos (Eds.), Elsevier*, pages 251–256, 2005.

[OT06] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research: Visual Perception*, pages 23–26, 2006.

[OTCH03]	A. Oliva, A. Torralba, M.S. Castelhano, and J.M. Henderson. Top-down control of visual attention in object detection. In *IEEE ICIP*, 2003.
[PHR+08]	S. Pannasch, J.R. Helmert, K. Roth, A.K. herbold, and H. Walter. Visual fixation durations and saccade amplitudes: Shifting relationship in a variety of conditions. *Journal of Eye Movement Research*, 2:1–19, 2008.
[PIIK05]	R. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 2005.
[PLN02]	D. Parkhurst, K. Law, and E. Niebur. Modelling the role of salience in the allocation of overt visual attention. *Vision Research*, 42:107–123, 2002.
[PV09]	S. Pannasch and B. M. Velichkovsky. Does the distractor effect allow identifying different modes of processing? In *ECEM*, 2009.
[Ray98]	K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422, 1998.
[RZB02]	R. P. N. Rao, G.J. Zelinsky, and M.M. Hayhoe D.H. Ballard. Eye movements in iconic visual search. *Vision Research*, pages 1447–1463, 2002.
[SG00]	D.D. Salvucci and J.H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *ETRA*, 2000.
[SO94]	P.G. Schyns and A. Oliva. From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, 5, 1994.
[SO99]	P.G. Schyns and A. Oliva. Dr. angry and mr. smile: when categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognition*, 69:243–265, 1999.
[TFM96]	S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 2:381–350, 1996.
[TOCH06]	A. Torralba, A. Oliva, M.S Castelhano, and J.M. Henderson. Contextual guidance of eye movements and attention

in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766–786, 2006.

[UF06] G. Underwood and T. Foulsham. Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology*, 59(11):1931–1949, 2006.

[USP05] P. Unema and B.M. Velichkovsky S. Pannasch, M. Joos. Time-course of information processing during scene perception: the relationship between saccade aplitude and fixation duration. *Visual Cognition*, 12:473–494, 2005.

[Woo02] D. Wooding. Fixation maps: quantifying eye-movement traces. In *Eye Tracking Research and Applications*, 2002.

[Yar67] A. Yarbus. Eye movements and vision. In *New-York, Plenum*, 1967.

[ZTM+08] 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*, 8:1–20, 2008.

ANNEXE POUR LA FABRICATION
A FOURNIR PAR LES AUTEURS AVEC UN EXEMPLAIRE
PAPIER
DE LEUR ARTICLE

1. ARTICLE POUR LA REVUE :
   Studia Informatica Universalis.

2. AUTEURS :
   Brice Follet (*)(***) — Olivier Le Meur (**) — Thierry Baccino (***)

3. TITRE DE L'ARTICLE :
   Modeling visual attention on scenes

4. TITRE ABRÉGÉ POUR LE HAUT DE PAGE MOINS DE 40 SIGNES :
   studia-Hermann

5. DATE DE CETTE VERSION :
   November 27, 2009

6. COORDONNÉES DES AUTEURS :

   – adresse postale :
   (*) Thomson, Compression Lab.
   1 av. de Belle Fontaine 35576 Cesson-Sevigne France
   brice.follet@thomson.net
   (**) University of Rennes 1
   Campus universitaire de Beaulieu 35042 Rennes France
   olemeur@irisa.fr
   (***) LUTIN (UMS-CNRS 2809), Cité des sciences et de l'industrie de la Villette
   30 av. Corentin Cariou 75930 Paris
   baccino@lutin-userlab.fr
   – téléphone : 01 44 10 84
   – télécopie : 00 00 00 00
   – e-mail : ivan.lavallee@gmail.com

7. LOGICIEL UTILISÉ POUR LA PRÉPARATION DE CET ARTICLE :
   LaTeX, avec le fichier de style `studia-Hermann.cls`,
   version 1.2 du 03/12/2007.

SERVICE ÉDITORIAL – STUDIA UNIVERSALIS