OBJECTIVE VALIDATION OF A DYNAMICAL AND PLAUSIBLE COMPUTATIONAL MODEL OF VISUAL ATTENTION

Matthieu Perreira da Silva, Vincent Courboulay, Pascal Estraillier

▶ To cite this version:

Matthieu Perreira da Silva, Vincent Courboulay, Pascal Estraillier. OBJECTIVE VALIDATION OF A DYNAMICAL AND PLAUSIBLE COMPUTATIONAL MODEL OF VISUAL ATTENTION. IEEE European workshop on visual information processing, Jul 2011, Paris, France. pp.223-228, 10.1109/EuVIP.2011.6045511 . hal-00617730

HAL Id: hal-00617730
https://hal.archives-ouvertes.fr/hal-00617730

Submitted on 18 Oct 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

OBJECTIVE VALIDATION OF A DYNAMICAL AND PLAUSIBLE COMPUTATIONAL MODEL OF VISUAL ATTENTION

Matthieu Perreira Da Silva, Vincent Courboulay, Pascal Estraillier

L3i - University of La Rochelle
Avenue M. Crépeau
17042 La Rochelle Cedex 01
France

ABSTRACT

Building a generic and highly capable vision system is still an open research problem. Actually, real-world vision systems need to face the challenge of dimensionality and ambiguity of data. To tackle this problem we introduced, in [1], a dynamic computational model of visual attention. This latter selects the most salient scene information while being able to adapt its behavior to the needs of a generic vision system. In this article, we focus on the objective validation of the plausibility of this original model. To check this property we compare (through three classical measures) the results obtained by several algorithms to an eye-tracking ground truth. Additionally, we study the influence of the model parameters on plausibility.

Index Terms— Visual attention, evaluation, objective measures, plausibility.

1. INTRODUCTION

Despite 40 years of research on computer vision, strong increase of computer processing power and huge improvements of image sensors quality (now 3D), building smart and generic computer vision system is still unreachable.

> «There are today numerous sophisticated methods for extracting visual information, but they seldom work consistently and robustly in the real, dynamically changing world» Eklundh et Christensen [2]

Yet, this general statement can be put into perspective when considering more specific application fields like face detection and tracking, industrial defect detection, scene reconstruction, motion analysis, content based information retrieval and so on. Actually, when some hypothesis related to the context can be stated, fine tuned algorithms, outperforming humans, can be built. However, as soon as a more general problem is considered (e.g. complex or changing environment) numerous problems arise : ambiguous data interpretation, combinatorial explosion due to data exploration, etc. Finding efficient mechanisms for selecting relevant features is one possible way that can be explored in order to tackle this problem.

The attentional system is one example of such a mechanism. It has been deeply studied by psychologists and neuro-psychologists since the 1950's. More recently, thanks to neurosciences progresses, computational attention models have emerged. These models help the study and understanding of specific aspects of human attention (see [3] for more details).

Real time execution constraints must also be taken into account when trying to adapt a computational attention model to computer vision. These facts lead to some questions:

- How can we make the attention model fast enough so that it remains a pre-processing ?
- To which extent should we imitate biological mechanism ?
- How can we integrate contextual information to attentional mechanisms ?

Concerning the last question, neuropsychologists have identified two complementary attentional pathways. The bottom-up pathway modulates attention according to the saliency of perceived stimuli, whereas the top-down pathway modulates attention according to the current context and objective. In [1] we introduced a dynamic and real-time computational bottom-up attention model which does not include a top-down pathway but is nevertheless highly tunable. In this paper, we focus on the objective validation of the plausibility of the model and the study of the influence of its main parameters.

In section 2, we propose a set of constraints for evaluating the capacity of an attentional model to be integrated in an intelligent computer vision system. In section 3, we briefly present the bottom-up visual attention model proposed in [1]. In section 4, we evaluate the model objectively through three quantified measures.

2. COMPUTATIONAL ATTENTION AND COMPUTER VISION

2.1. Computer vision needs

As previously mentioned, it is necessary to get some criteria in order to classify or evaluate algorithms. Usually, processing time or robustness are used. Nevertheless, we can closely adapt a set of constraints to computer vision applications. We have named this set PAIRED, it is composed of the following elements:

- Plausible when compared to human behavior;
- Adaptable to varying contexts;
- Invariant through different transforms (rotation, translation and scale);
- Rapid to compute the focus of attention;
- Extensible concerning its ability to take into account new characteristics;
- Dynamic and capable of producing results at any time.

Once these criteria presented, we illustrate their importance in several vision applications in the following Table 1. One dot means a weak constraint opposed to a strong one represented by three dots.

In the next sub-section, we briefly present two classical families of attention model in computer vision and put them in regards to the PAIRED criteria.

2.2. Classical models in computer vision

2.2.1. Central map base models

In the 80's, Anne Treisman proposed the *Feature-Integration Theory* [4]. She stated that attention is encoded by an internal central map (named master map of locations, saliency map, etc.). Many works inherits this works, like hierarchical or algorithmic models.

Hierarchical models

These models compute several feature maps from an image and progressively combine them until providing a unique map usually named saliency map. One of the most important models belonging to this class is the one of Laurent Itti [5]. Many reasons explain its popularity among which we found :

- it is one of the first model of attention;
- it is based on an influential theory [4];
- its biologically inspired architecture is efficient and simple to understand;
- thanks to the *Neuromorphic Visionl Toolkit*[1] (*iNVT*), researchers have source code and software to either improve the model or compare with their own
- since its presentation in 1998, it is regularly improved.

We describe this model in more details in section 3 since it is the base of the hierarchical part of Perreira's model.

Algorithmic models.

These models are usually built for a specific application. For example, in the field of computer vision, an interesting approach is proposed by Aziz et Mertsching [6]. They use the notion of proto-objects: images are segmented into regions of uniforms colors on which different features are computed (contrast, color, size, symmetry orientation). This results in a very fast model which may however be very sensitive to quality of the initial color segmentation process.

2.2.2. Distributed models

Another attention modeling way can be found in distributed models. These are based on works done by neuroscientists and connectionists and are usually neuromimetic [7]. Experimental studies may then be as close as possible of *in vivo* experiments through specific signals measures.

Most of related works are inspired by biased competition model proposed by Desimone and Duncan [8] in the middle of 90's. For instance, Deco, Stringer and Rolls [9] and Ji and Weng [10] propose to deeply mimic pathways information processing through two channels; the "what "pathway, responsible for recognition, and the "where" pathway that takes into account spatial attention.

2.2.3. Conclusion

Considering interesting properties of fidelity, invariance dynamic and adaptability of the distributed models and rapidity and extensibility of hierarchical models, an hybrid approach between these two model families seems to be promising. This solution has been partially developed through connectionist centralized models [11, 12], actually they combined a hierarchical approach to produce a saliency map and a distributed model to generate the focus of attention. One of the main limitations is that competition between features (color, intensity and orientation) is done by the hierarchical part of the model. As a consequence, saliency sources are not competed. It is the main reason why we think that an "ideal" model (if it exists) do not have to represent saliency through a centralized map. We exploited this idea in [1] where we proposed that conspicuity maps should compete via a dynamical system (prey predator system for instance) and thus, be able to provide a dynamical focus of attention.

[1] Available at : `http://ilab.usc.edu/toolkit/`.

	Plausible	Adaptable	Invariant	Rapid	Extensible	Dynamic
Attention modelling	●●●	●	●	●	●	●●●
ergonomics/advert	●●●	●	●	●	●●	●●
Vision	🟢🟢	🟢🟢🟢	🟢🟢	🟢🟢🟢	🟢🟢🟢	🟢🟢🟢
CBIR	●	●	●●●	●●	●●●	●
Image processing	●●	●●	●●	●●	●●●	●●

Table 1. *PAIRED* constraints balance for some vision applications.

	Plausible	Adaptable	Invariant	Rapid	Extensible	Dynamic
Target	⋆⋆	⋆⋆⋆	⋆⋆	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆
Distributed	🟢🟢🟢	🟢🟢🟢	🟢🟢🟢	●	●	🟢🟢🟢
Hierarchical	●●	●●	●●	●●	🟢🟢🟢	●
Algorithmic	●	●●	🟢🟢	🟢🟢🟢	●	●

Table 2. Set of constraints put in regards to computer vision applications. First row corresponds to our objective. Criteria reached or exceeded by models families are represented in green.

3. BOTTOM-UP COMPUTATIONNAL ATTENTION AND DYNAMICAL SYSTEMS

In [1], we proposed a new method which allows studying the temporal evolution of the visual focus of attention. We modified the classical algorithm proposed by Itti [5], in which the first part of his architecture relies on the extraction of three conspicuity maps based on low level computation. These three conspicuity maps are representative of the three main human perceptual channels: color, intensity and orientation. These low level computations are optimized following works presented in [13]. Actually the way to accelerate computation, (*i.e.* the use of integral images [14]) is reused and extended to all maps. The second part of Itti's architecture proposes a medium level system which allows merging conspicuity maps and then simulates a visual attention path on the observed scene. The focus is determined by a winner takes all and an inhibition of return algorithms. We have substituted this second part by a competitive dynamical system, in order to introduce a temporal parameter, which allows generating saccades, fixations and more realistic paths (figure 1).

Preys-predators based dynamical systems are particularly well adapted for such a task. The main reasons are:

- preys-predators systems are dynamic, they intrinsically include time evolution of their activities. Thus, visual attention focus, seen as a predator, can evolve dynamically;
- without any objective (top down information or pregnancy), choosing a method for conspicuity maps fusion is hard. A solution consists in developing a competition between conspicuity maps and waiting for a natural balance in the preys / predators system, reflecting the competition between emergence and inhibition of elements that engage or do not engage our attention;
- discrete dynamic systems can have a chaotic behavior. Despite the fact that this property is not often interesting, it is an important one for an attention model. Actually, it allows the emergence of original paths and an exploration of the visual scene, even in non salient areas, reflecting something like *curiosity*.

In our original article, we proposed a subjective validation, with different improvements, of Laurent Itti's visual attention model. In this article, we present an objective evaluation using three different measures that confirm our first conclusions about the plausibility of the method.

4. ORIGINAL MODEL PLAUSIBILITY

4.1. Comparison to existing models

We have performed an objective validation, related to plausibility, of the model introduced in [1]. This validation consists in checking the plausibility of the system, *i.e.* checking if it is apparently reasonably valid, and truthful.

All measures were done on two image databases. The first one is proposed by Bruce[2] [15]. It is made up of 120 color images which represent streets, gardens, vehicles or buildings, more or less salients. The second one, proposed by Le Meur[3] [16], contains 26 color images. They represent sport

[2] Bruce database is available at `http://www-sop.inria.fr/members/Neil.Bruce/`

[3] Le Meur database is available at `http://www.irisa.fr/temics/staff/lemeur/visualAttention/`

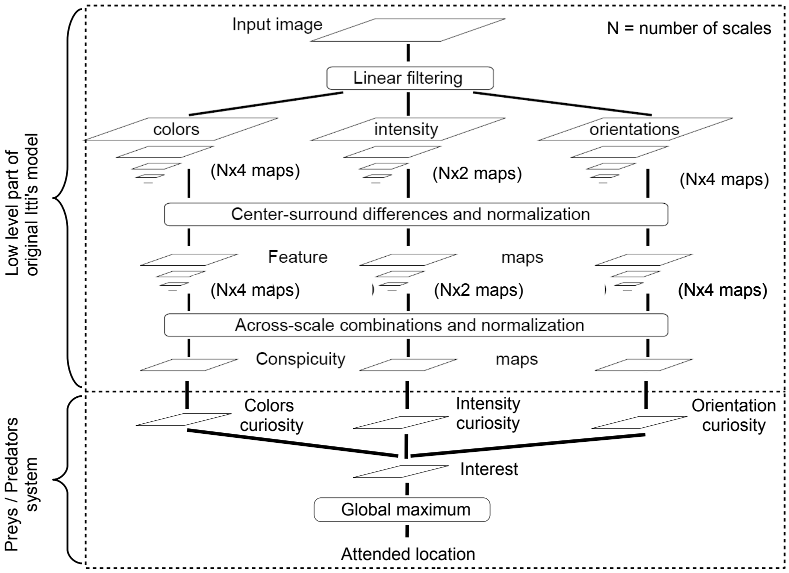


Fig. 1. Architecture of our bottom-up visual attention model. This diagram has been adapted from [5].

scenes, animals, building, indoor scenes or landscapes. For both databases, eye movements recordings were performed during a free viewing task.

In [1], we have presented a subjective evaluation of our model. We confirm this evaluation by a more ”classical” objective evaluation. Cross-correlation, Kullback-Leibler divergence and normalized scanpath saliency were used to compare 6 algorithms to an eye-tracking ground-truth. The models evaluated were :

- two naïve models. ”AllEqual” correspond to a constant saliency map, consider all points as equally salient. ”Gaussian” model considers the central part of the image as the most salient area. Saliency is distributed using a centered gaussian distribution, scaled in order to cover all the image;
- Le Meur model [16], in its ”coherent normalization” flavor;
- the AIM model of Bruce and Tsotsos[15];
- the NVT model of Itti [5].
- Perreira's model (with fast retinal blur, see [1] for details).

All models were tested using their default parameters.

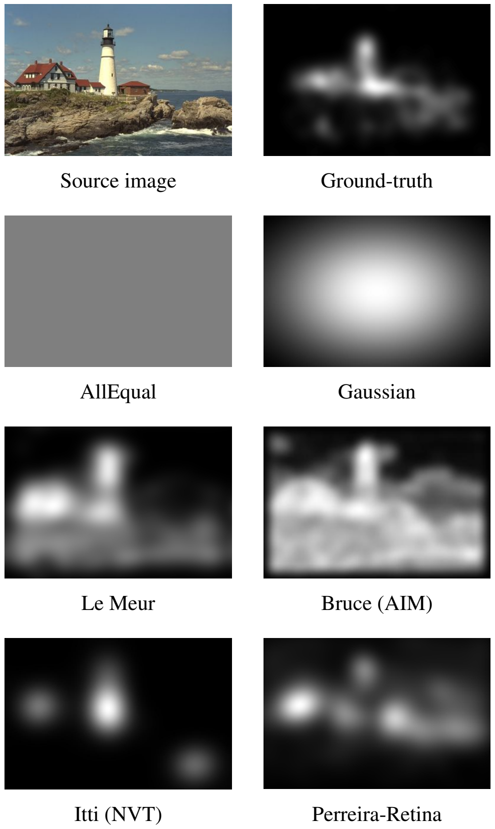
Figure 2 shows some sample saliency maps and heatmaps whereas table 4 is a summary of the performance of each algorithm over all the images of the two test databases. The analysis of the latter table leads to the following remarks :

- Kullback-Leibler divergence is sensitive to maps normalization : the ”AllEqual” model seems to perform better than Itti's model whereas it obtains a null score with the two other measures;
- the ”AllEqual” model is (quite unsurprisingly) the worst performer;
- despite its simplicity, the ”Gaussian” model is quite a plausible model. This central preference is well known bias when evaluating computational attention models over eye-tracking data. It may be due to the type of images included in the databases, the experimental protocol, the photographer bias (which tends to center it's subject in the picture), or a real attentional bias against the central position;
- Model's performances are comparable to other state of the art models and even outperform them on Bruce database (NSS measure).

4.2. Influence of parameters

In the previous section, we have shown that our dynamical attention model is as plausible as other state of the art models. However, this model (and in particular its dynamical system) depends on some parameters. Table 3 summarizes the influences of some of these parameters on the plausibility of the model. The following conclusions can be drawn :

- using a retinal filter during the generation of feature and conspicuity maps improves plausibility significantly. This tends to prove that each new attentional focus depends on the location of the previous attentional focus;



Source image

Ground-truth

AllEqual

Gaussian

Le Meur

Bruce (AIM)

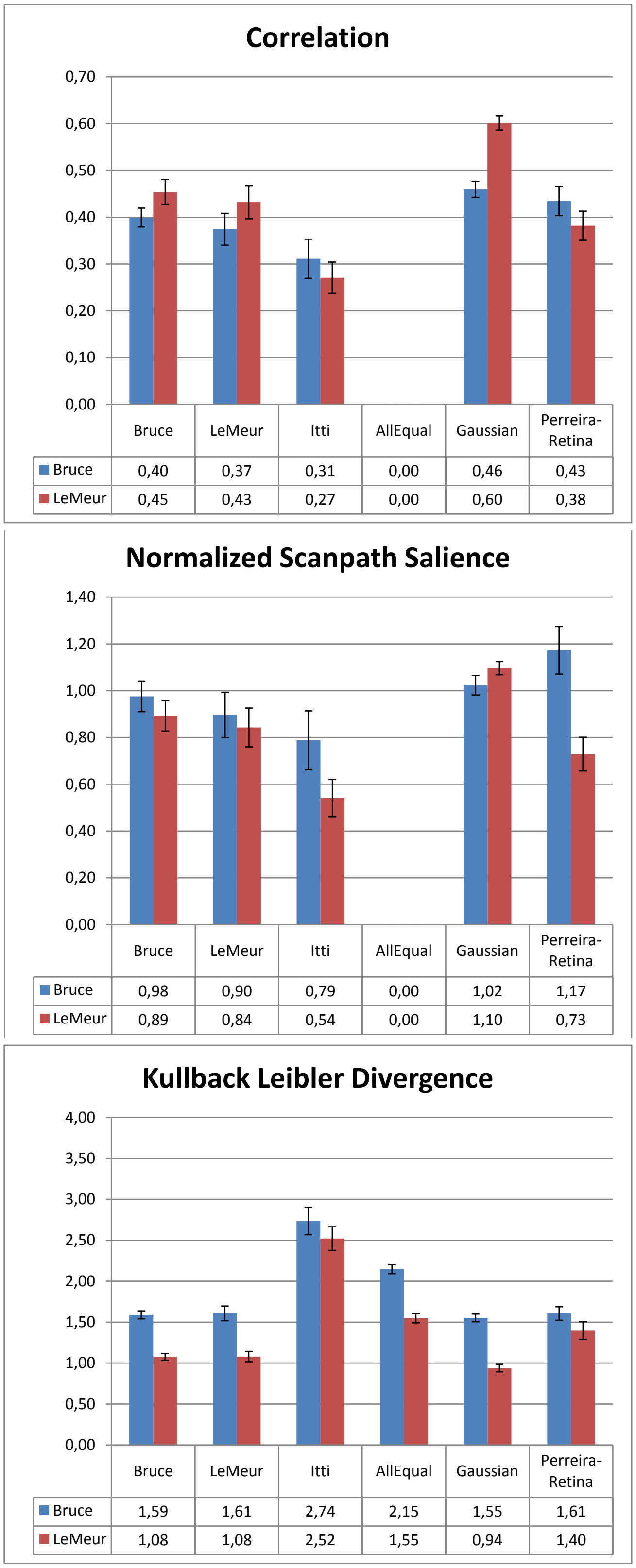
Itti (NVT)

Perreira-Retina

Fig. 2. Example of heatmaps and saliency maps generated by different computational attention models.

	CC		KLD		NSS		Mean
	Bruce	LeMeur	Bruce	LeMeur	Bruce	LeMeur	gain
Default	0,35	0,30	1,80	1,76	0,95	0,56	0%
RetinalFilter	0,43	0,38	1,61	1,40	1,17	0,73	22%
CentralBias=0.00	0,20	0,14	2,33	2,29	0,57	0,27	-41%
CentralBias=0.25	0,48	0,44	1,57	1,49	1,29	0,82	32%
CentralBias=0.50	0,55	0,53	1,92	1,66	1,49	1,01	45%
Diffusion=0.00	0,33	0,23	2,06	2,26	0,96	0,47	-14%
Diffusion=0.125	0,35	0,29	1,77	1,70	0,95	0,55	0%
Diffusion=0.50	0,35	0,31	1,83	1,72	0,94	0,59	1%
Noise=0.00	0,17	0,06	4,32	4,77	0,49	0,13	-95%
Noise=0.25	0,16	0,07	4,21	4,49	0,48	0,15	-90%
Noise=0.75	0,46	0,44	1,61	1,17	1,25	0,83	33%
Noise=1.00	0,27	0,35	1,89	1,30	0,68	0,64	0%

Fig. 3. Influence on plausibility of the main model's parameters.



	Bruce	LeMeur	Itti	AllEqual	Gaussian	Perreira-Retina
🟦 Bruce	0,40	0,37	0,31	0,00	0,46	0,43
🟥 LeMeur	0,45	0,43	0,27	0,00	0,60	0,38

	Bruce	LeMeur	Itti	AllEqual	Gaussian	Perreira-Retina
🟦 Bruce	0,98	0,90	0,79	0,00	1,02	1,17
🟥 LeMeur	0,89	0,84	0,54	0,00	1,10	0,73

	Bruce	LeMeur	Itti	AllEqual	Gaussian	Perreira-Retina
🟦 Bruce	1,59	1,61	2,74	2,15	1,55	1,61
🟥 LeMeur	1,08	1,08	2,52	1,55	0,94	1,40

Fig. 4. Comparison to ground-truth : comparision of different algorithms. (Please note that for KLD, lower values mean more plausibility),

- using central biasing in an attention model can improve significantly its plausibility, but this bias is partly due to the experimental protocol [17];
- the dynamical system used in our attention model needs some diffusion in order to work correctly, but adding more diffusion does not improve plausibility;
- similarly, noise is an important factor for the plausibility of the model. However, the influence of noise on the repeatability of the system (variation in behavior between different runs) is still an open question.

5. CONCLUSION

In this article, we have presented an objective and quantified validation of a new computational attention model dedicated to intelligent vision. In a first part, we propose a new set of constraints (named *PAIRED*) for the evaluation of an attention model to a specific application. Considering the evaluation of classical models of visual attention, in regards of these constrains, we propose to take advantage of the qualities of both most popular model types (hierarchical and distributed). This initial bottom-up hybrid model, subjectively validated in [1] is now objectively validated through three measures. These measures are used on two set of images. Each of them provides oculometric measures that represent ground truth. We show that our dynamic model is plausible and that some of its parameters have a great influence on its plausibility. An extension of this work will be to study how feedback mechanisms can be exploited in order to reuse the outputs of a real computer vision system.

6. REFERENCES

[1] M. Perreira Da Silva, V. Courboulay, A. Prigent, and P. Estraillier, "Evaluation of preys / predators systems for visual attention simulation," in *VISAPP 2010 - International Conference on Computer Vision Theory and Applications*, Angers, 2010, pp. 275–282, INSTICC.

[2] J.O. Eklundh and H. Christensen, *Computer vision: Past and future*, pp. 328–340, Springer-Verlag, Berlin, 2001.

[3] L. Itti, G. Rees, and J.K. Tsotsos, Eds., *Neurobiology of attention*, Academic Press, 1st edition, 2005.

[4] A. Treisman and G. Gelade, "A Feature-Integration Theory of Attention," *Cognitive Psychology*, vol. 136, no. 12, pp. 97–136, 1980.

[5] L. Itti, C. Koch, E. Niebur, and Others, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[6] M.Z. Aziz and B. Mertsching, "Fast and robust generation of feature maps for region-based visual attention.," *IEEE transactions on image processing*, vol. 17, no. 5, pp. 633–44, 2008.

[7] E.T. Rolls and S.M. Stringer, "Invariant visual object recognition: a model, with lighting invariance.," *Journal of physiology*, vol. 100, no. 1-3, pp. 43–62, 2006.

[8] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual review of neuroscience*, vol. 18, pp. 193–222, 1995.

[9] G. Deco, "A Neurodynamical cortical model of visual attention and invariant object recognition," *Vision Research*, vol. 44, no. 6, pp. 621–642, 2004.

[10] Z. Ji and J. Weng, "Where-What Network 1 : Where and What Assist Each Other Through Top-down Connections," in *IEEE 7th International Conference on Development and Learning*, 2008, pp. 61–66.

[11] S. Ahmad, *VISIT: An efficient computational model of human visual attention*, Ph.D. thesis, University of Illinois, Urbana-Champaign, Champaign, IL, 1992.

[12] Julien Vitay, N.P. Rougier, and F. Alexandre, *Biomimetic Neural Learning for Intelligent Robots*, vol. 3575, chapter A distributed model of spatial visual attention, pp. 54–72, Springer, 2005.

[13] S. Frintrop, M. Klodt, and E. Rome, "A real-time visual attention system using integral images," in *5th International Conference on Computer Vision Systems (ICVS)*, Bielefeld, Germany, 2007, Applied Computer Science Group.

[14] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2002.

[15] 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. 5, 2009.

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

[17] B.W. Tatler, "The central fixation bias in scene viewing : Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of Vision*, vol. 7, pp. 1–17, 2007.