Real-time computational attention model for dynamic scenes analysis: from implementation to evaluation.

Vincent Courboulay, Matthieu Perreira da Silva

Real-time computational attention model for dynamic scenes analysis: from implementation to evaluation.

Vincent Courboulay[a] and Matthieu Perreira Da Silva[b]

[a]L3i lab, University of La Rochelle, France;
[b]IRCCyN lab, University of Nantes, France

ABSTRACT

Providing real time analysis of the huge amount of data generated by computer vision algorithms in interactive applications is still an open problem. It promises great advances across a wide variety of fields. When using dynamics scene analysis algorithms for computer vision, a trade-off must be found between the quality of the results expected, and the amount of computer resources allocated for each task. It is usually a design time decision, implemented through the choice of pre-defined algorithms and parameters. However, this way of doing limits the generality of the system. Using an adaptive vision system provides a more flexible solution as its analysis strategy can be changed according to the new information available. As a consequence, such a system requires some kind of guiding mechanism to explore the scene faster and more efficiently. We propose a visual attention system that it adapts its processing according to the interest (or salience) of each element of the dynamic scene. Somewhere in between hierarchical salience based and competitive distributed, we propose a hierarchical yet competitive and non salience based model. Our original approach allows the generation of attentional focus points without the need of neither saliency map nor explicit inhibition of return mechanism. This new real-time computational model is based on a preys / predators system. The use of this kind of dynamical system is justified by an adjustable trade-off between nondeterministic attentional behavior and properties of stability, reproducibility and reactiveness.

1. INTRODUCTION

While machine vision systems are becoming increasingly powerful, in most regards they are still far inferior to their biological counterparts. In human, the mechanisms of evolution have generated the visual attention system which selects the most important information in order to reduce both cognitive load and scene understanding ambiguity. Thus, studying the biological systems and applying the findings to the construction of computational vision models and artificial vision systems are a promising way of advancing the field of machine vision.

In the field of scene analysis for computer vision, a trade-off must be found between the quality of the results expected, and the amount of computer resources allocated for each task. It is usually a design time decision, implemented through the choice of pre-defined algorithms and parameters. However, this way of doing it limits the generality of the system. Using an adaptive vision system provides a more flexible solution as its analysis strategy can be changed according to the information available concerning the execution context. As a consequence, such a system requires some kind of guiding mechanism to explore the scene faster and more efficiently.

In this article, we propose a first step to building a bridge between computer vision algorithms and visual attention. In particular, we will describe how to create and evaluate a visual attention system tailored for interacting with a computer vision system so that it adapts its processing according to the interest (or salience) of each element of the scene. Somewhere in between hierarchical salience based and competitive distributed models, we propose a hierarchical yet competitive model. Our original approach allows us to generate the evolution of attentional focus points without the need of either saliency map or explicit inhibition of return mechanism. This new real-time computational model is based on a dynamical system. The use of such a complex system is justified by an adjustable trade-off between nondeterministic attentional behavior and properties of stability, reproducibility and reactiveness.

We justify why dynamical systems are a good choice for visual attention simulation, and we show that preys / predators models provide good properties for simulating the dynamic competition between different kinds of

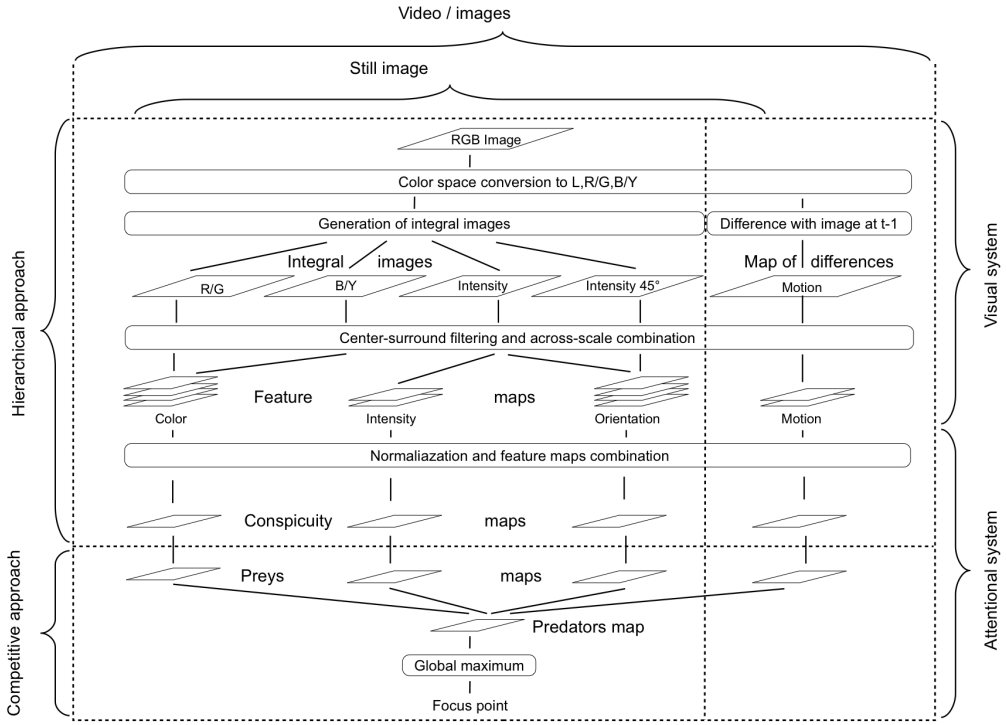


Figure 1. Architecture of the computational model of attention.

information. This dynamical system is also used to generate a focus point at each time step of the simulation. In order to show that our model can be integrated in an adaptable computer vision system, we show that this architecture is fast and allows a flexible real time visual attention simulation. In particular, we present a feedback mechanism used to change the scene exploration behavior of the model. This mechanism can be used to maximize the scene coverage (explore each and every part) or maximize focalization on a particular salient area (tracking). In a last section we present the evaluation results of our model. Since the model is highly configurable, its evaluation will cover not only its plausibility (compared to human eye fixations), but also the influence of each parameter on a set of properties (stability, reproducibility, scene exploration, dynamic behavior).

2. APPLICATION TO A REAL TIME VISUAL ATTENTION MODEL

Many reaserchers have worked on visual attention, we classiquely used Laurent Itti's work.[1] The first part of its architecture relies on the extraction of three conspicuity maps based on low level characteristics computation, that's correspond to the production of information on retina. These three conspicuity maps are representative of the three main human perceptual channels: color, intensity and orientation.

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 (Figure 1).

We propose to substitute this second part by our optimal competitive theory conclusion: a preys / predators system. This optimal criteria, preys / predators equations are particularly well adapted for such a task:

• preys / predators systems are dynamic, they include intrinsically time evolution of their activities. Thus, the visual focus of attention , 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

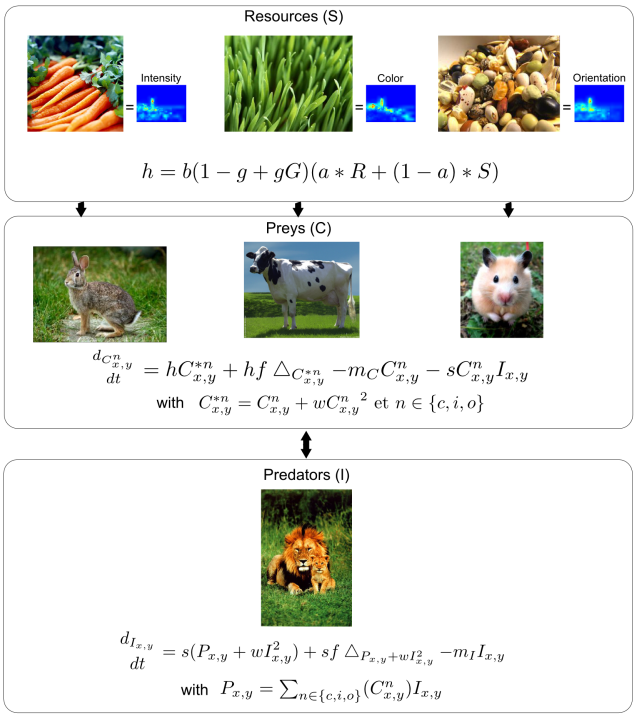


Figure 2. Competitive preys / predators attention model. Singularity maps are the resources that feed a set of preys which are themselves eat by predators. The maximum of the predators map represents the location of the current focus of attention.

balance in the preys / predators system, reflecting the competition between emergence and inhibition of elements that engage or not 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 us. Actually, it allows the emergence of original paths and exploration of visual scene, even in non salient areas, reflecting something like curiosity.

2.1 How to modelise visual attention with a 2D preys / predators system

As previously mentionned, we have demonstrated that model visual attention with a competitive dynamical system biologically inspired is an optimal way of extracting information. General architecture is represented in Figure 2

Starting from this "basic" version of preys / predators equations, we can enriche processing in several ways:

- the number of parameters can be reduced by replacing s' by s. Indeed, mortality rates differences between preys and predators can be modeled by an adjustment of factors b and m_I
- the original model represents the evolution of a single quantity of preys and predators over time. It can be spatially extended in order to be applied to 2D maps where each point represents the amount of preys or predators at a given place and time. Preys and predators can then "move" on this map using a classical diffusion rule, proportional to their Laplacian $\triangle_C$ and a diffusion factor f.
- natural mortality of preys in the absence of predation is not taken into account. If the model only changes temporally, mortality is negligible when compared to predation. However, when the model is applied to a 2D map (which is the case in our system), some areas of the map may not contain any predator. Natural mortality of prey can no longer be considered negligible. A new mortality term $-m_c$ need to be added to the model.

This yield to the following set of equations, modeling the evolution of preys and predators populations on a two dimensional map:

$$\begin{cases} \frac{d_{C_{x,y}}}{dt} & = bC_{x,y} + f \triangle_{C_{x,y}} - m_C C_{x,y} - sC_{x,y}I_{x,y} \\ \frac{d_{I_{x,y}}}{dt} & = sC_{x,y}I_{x,y} + sf \triangle_{P_{x,y}} - m_I I_{x,y} \end{cases} \quad (1)$$

A last phenomenon can be added to this model: a positive feedback, proportional to C^2 or I_2 and controlled by a factor w. This feedback models the fact that (provided unlimited resources) the more numerous a population is, the better it is able to grow (more efficient hunting, higher encounter rater favoring reproduction, etc.). The new preys / predators system is now:

$$\begin{cases} \frac{d_{C_{x,y}}}{dt} & = b(C_{x,y} + w(C_{x,y})^2) + f \triangle_{C_{x,y}} - m_C C_{x,y} - sC_{x,y}I_{x,y} \\ \frac{d_{I_{x,y}}}{dt} & = s(C_{x,y}I_{x,y} + w(I_{x,y})^2) + sf \triangle_{P_{x,y}} - m_I I_{x,y} \end{cases} \quad (2)$$

In order to simulate the evolution of the focus of attention, we propose a preys / predators system (as described above) with the following features:

- the system is comprised of four types of preys and one type of predators;
- these four types of preys represent the spatial distribution of the curiosity generated by our four types of conspicuity maps (intensity, color, orientation and motion);
- the predators represent the interest generated by the consumption of curiosity (preys) associated to the different conspicuity maps;
- the global maximum of the predators maps (interest) represents the focus of attention at time t.

The equations described in the next sub-section are obtained by building a preys / predators system which integrates the above cited features.

2.2 Simulating the evolution of the attentional focus with a Preys / predators system

For each of the three conspicuity maps (color, intensity and orientation) extended with another one, motion, the preys population C evolution is governed by the following equation:

$$\frac{d_{C^n_{x,y}}}{dt} = hC^{*n}_{x,y} + hf \triangle_{C^{*n}_{x,y}} - m_C C^n_{x,y} - sC^n_{x,y}I_{x,y} \quad (3)$$

with $C^{*n}_{x,y} = C^n_{x,y} + w{C^n_{x,y}}^2$ and $n \in \{c, i, o, m\}$, which mean that this equation is valid for C^c, C^i,C^o and C^m which represent respectively color, intensity, orientation and motion populations.

C represents the curiosity generated by the image's intrinsic conspicuity. It is produced by a sum $h$of four factors:

$$h = b(1 - g + gG)(a * R + (1 - a) * SM_n)(1 - e) \quad (4)$$

- the image's conspicuity SM_n (with $n \in \{c, i, o, m\}$) is generated using our real time visual system, previously described in this article. Its contribution is inversely proportional to a;
- a source of random noise R simulates the high level of noise that can be measured when monitoring our brain activity ([2]). Its importance is proportional to a. Equations that model the evolution of our system become stochastic differential equations. A high value for a gives some "freedom" to the attentional system, so it can explore less salient areas. On the contrary, a lower value for a will constraint the system to only visit high conspicuity areas;

a	b	g	w	m_C	m_I	s	f	$T_{Hysteresis}$
0.5	0.007	0.1	0.001	0.3	0.5	0.025	0.25	0.0

Table 1. . Default parameters of the preys / predators model

- a Gaussian map G which simulates the central bias generally observed during psycho-visual experiments ([3,4]). The importance of this map is modulated by g
- the entropy e of the conspicuity map (color, intensity, orientation or motion). This map is normalized between 0 and 1. C is modulated by $1-e$ in order to favor maps with a small number of local minimums. Explained in terms of preys / predators system, we favor the growth of the most organized populations (grouped in a small number of sites). This mechanism is the preys / predators equivalent to the feature maps normalization presented above.

The population of predators I, which consume the 4 kinds of preys, is governed by the following equation:

$$\frac{d_{I_{x,y}}}{dt} = s(P_{x,y} + wI_{x,y}^2) + sf \triangle_{P_{x,y}+wI_{x,y}^2} - m_I I_{x,y} \tag{5}$$

with $P_{x,y} = \sum_{n\in\{c,i,o\}}(C_{x,y}^n)I_{x,y}$.

As already mentioned the positive feedback factor w enforces the system dynamics and facilitates the emergence of chaotic behaviors by speeding up saturation in some areas of the maps. Lastly, please note that curiosity C is consumed by interest I, and that the maximum of the interest map I at time t is the location of the focus of attention.

To allow less frequent changes of the position of the focus of attention, we added an optional hysteresis mechanism. This latter changes the focus of attention only if the new maximum of the predators map exceeds its previous value by more than a certain threshold:

$$Focus(t) = \begin{cases} (x_{max}, y_{max}) & if\ \max_{x,y}(P_{x,y}(t)) > \\ & (1 + Seuil_{Hysteresis}) \times \max_{x,y}(P_{x,y}(t-1)) \\ Focus(t-1) & otherwise \end{cases}$$

with (x_{max}, y_{max}) the coordinates of the maximum of $P_{x,t}(t)$, $Seuil_{Hysteresis}$is the hysteresis threshold and $Focus(t)$are the coordinates of the current focus of attention.

This system has been implemented in real time, see.[?,5,6]

2.3 Default parameters of the preys / predators system

During the experiments presented at the end of this article, the following (empirically determined) parameters were used (Table 1):

These parameters represent reasonable values that can be used to obtain a system at equilibrium. This equilibrium is obtained when the system is run without any input image. Other parameters combinations are possible. In particular, experiments have shown that these values can be varied within a wide range without compromising the system's stability (see further for details). The system is thus quite robust to its parameters variation.

Please note that our implementation of the model evolves according to Euler method using a step size of 0.33 and that 3 sub-iterations are run before computing each simulated focus of attention.

2.4 Top-Down feedback

The attention model presented in this article is computationally efficient and plausible. It provides many tuning possibilities (adjustment of curiosity, central preferences, etc.) that can be exploited in order to adapt the behavior of the system to a particular context. This adaptation is however somewhat limited. In this sub-section, we propose to extend our bottom-up model so that it can take into account more information concerning its objectives.

This top-down influence can be expressed as a simple modification of the model parameters, but it can also reuse information generated by the system itself in order to modify its behavior. In the latter case, a feedback loop is created (auto-adaptation).

In the following, we define the adaptation mechanisms used in our model. We will also explore how previously visited locations can be used as inputs to an attentional feedback mechanism aimed at controlling the scene exploration capabilities of the model.

2.4.1 Adaptation mechanisms

In this sub-section, we describe the different mechanisms that can be used in order to adapt the model behavior to external constraints (e.g. top-down information).

Top down map Usually, top-down information is included in hierarchical computational attention model in either of these two ways:

- global weighting of feature maps which allows a bias of the attentional system in favor of the distinctive features of a target object. This mechanism is used, for example, in ([7]) in order to learn which features are salient, depending on the context.
- local weighting of feature maps. This approach is an extension of the global weighting scheme which allows specifying prior knowledge about the target localization. This mechanism is exploited by ([8]) where it is called task-relevance maps.

Other extensions are also possible, for example using prior knowledge about the intensity of some expected features ([9]).

Even if the conspicuity maps fusion part of our model of attention is competitive (and thus non hierarchical), it can be biased using top-down maps. This can be done using a map (different for each king of prey) which will favor the growth of one kind of preys against others (eventually at preferred locations) :

$$\frac{d_{C^n_{x,y}}}{dt} = T^n_{x,y}\left(1 - \frac{C^n_{x,y}}{Max_{population}}\right)\left(C^{*n}_{x,y} + hf\triangle_{C^{*n}_{x,y}}\right) - m_C C^n_{x,y} - sC^n_{x,y}I_{x,y} \quad (6)$$

where $T^n_{x,y}$ is the top-down map associated to a prey type, $n \in \{i, c, o, m\}$ and $max_{x,y}(T^n_{x,y}) = 1.0$.

If $T^n_{x,y} = W_n \forall (x, y)$ then the evolution of prey n is constrained by a global weight (Figure 3).

Otherwise, saliency boosting is local ([10]). It can be used, for example, to favor colored targets located in the right part of the scene.

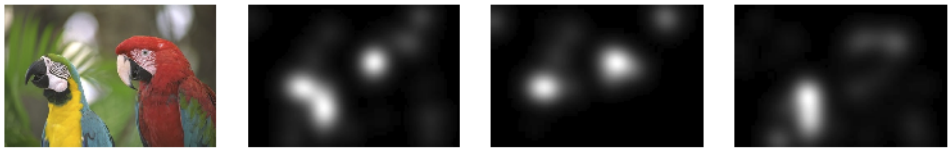


Figure 3. Effects of global weighting. a) heatmap generated with default parameters ($Wc = Wi = Wo = 1.0), b$) heatmap generated with lower color weights ($Wc = 0.5), c$) heatmap generated with high color weight ($Wi = Wo = 0.5$).

Feedback maps Top-down maps described in the previous paragraphs help modifying the attentional system behavior using contextual prior-knowledge (external to the model of attention). But the system can also be biased using information generated by the model itself or the computer vision it is connected to.

This mechanism can be implemented using a global feedback map R which will be used as a facilitation or inhibition mechanism. Preys growth equation now becomes:

$$\frac{d_{C^n_{x,y}}}{dt} = R_{x,y} T^n_{x,y} \left(1 - \frac{C^n_{x,y}}{Max_{population}}\right) \left(hC + hf \triangle_{C^{*n}_{x,y}}\right) - m_C C^n_{x,y} - s C^n_{x,y} I_{x,y} \tag{7}$$

where $R_{x,y}$ is built according to one or more feedback criteria. An example criterion, based on scene exploration is given in the next sub-section.

2.4.2 A feedback criterion: scene exploration

In this sub-section, we describe how we can use a scene exploration feedback criterion based on the history of attentional focus points generated by our system. If we build a map of previously visited locations and modulate (negatively of positively) its influence in the preys growing equation, we can define two complementary attentional strategies (and all intermediate states):

- scene exploration maximization : the attentional system will favor unvisited areas;
- focalizations stability : the attentional system will favor already visited areas.

We now describe how this visited areas maps is constructed, and how it can be used as a feedback map.

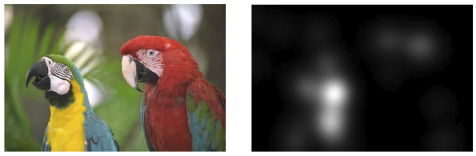
Visited areas map The visited areas map construction is based on the following hypothesis. During an attentional focus, most of the information is acquired at the center of a circular area (equivalent to the fovea in the retina). In the rest of this circular area, information linearly looses importance as we move away from the center.

The visited areas map is constructed incrementally in order to keep a memory of all the information acquired in the scene:

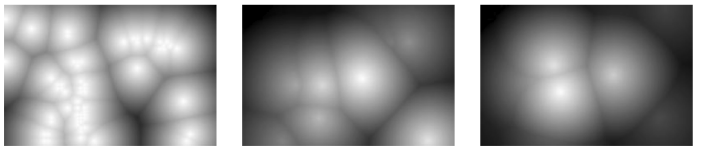
$$M_{visit}(x, y, t) = \max(M_{visit}(x, y, t-1), \frac{N_{Levels} - \min\left(\frac{dist(x,y,x_f,y_f)}{BlurSize}, N_{Levels}\right)}{N_{Levels}} \tag{8}$$

where (x_f, y_f) are the coordinates of the focus of attention at time t; $dist(x_1, y_1, x_2, y_2)$ is the Euclidian distance between (x_1, y_1) and (x_2, y_2); $BlurSize$ the size of the retinal area (fixed to 10 of the largest image dimension; this value may be associated with human fovea size (about 2 degrees of visal field)); $NbLevels = ceiling(log_2(min(W, H))$ and $(W, H)$the size of the input image. It guarantees that $M_{visites}(x, y) \in [0, 1] \, \forall x, y$.

Human memory is however limited, so attentional focus is most probably influenced by only the most recent focus points. To improve the plausibility of M_{visit}, we should take into account this fact and update M_{visit}



(a) Original image (b) *Heatmap.*



(c) Visited areas, $F_{forget} = 1.0$. (d) Visited areas, $F_{forget} = 0.95$ (e) Visited areas, $F_{forget} = 0.90$.

Figure 4. . Influence of F_{forget} on the visited area map, after 100 attention simulation iterations.

equation by introducing a "forgetting" factor $F_{forget} \in [0, 1]$ which will iteratively attenuate the role of the oldest focus points:

$$M_{visites}(x, y, t) = \max(F_{forget} \times M_{visites}(x, y, t-1), \frac{N_{Levels} - \min\left(\frac{dist(x,y,x_f,y_f)}{BlurSize}, N_{Levels}\right)}{N_{Levels}} \tag{9}$$

Figure 4 shows the influence of F_{forget} on the visited areas map $Mvisit$.

Feedback map processing The feedback map R is built upon M_{visit}. Its parameter $F_{feedback}$ allows modulating the influence of the visited areas map in intensity and feedback type (positive or negative):

$$R(x, y) = \begin{cases} \frac{1+|F_{feedback}| \times M_{visites}(x,y)}{1+|F_{feedback}|} & si\ F_{feedback} \geq 0 \\ \frac{1+|F_{feedback}| \times (1-M_{visites}(x,y))}{1+|F_{feedback}|} & sinon \end{cases} \tag{10}$$

with$R(x, y) \in [0, 1] \forall x, y$.

A positive feedback value will lead to a focalization or tracking behavior since already visited objects / locations are preferred. A negative feedback value will lead to an exploration behavior since unknown (unvisited) objects or locations will be favored.

The computational model of attention described in this chapter provides many tuning parameters and adaptation mechanisms. In order to validate this model we need to evaluate its plausibility by comparing its prediction with human fixations; but we also need to study the way it reacts when its parameters are adjusted. Indeed, the model is dedicated to computer vision and as such we should provide some clues concerning its general behavior (plausibility, reproducibility, etc.). This is the purpose of the next section.

2.5 Model properties

In order to conduct the study of our model, it is necessary to define one or more observation levels (microscopic or macroscopic) as well as a set of properties. In this article we study macroscopic properties, since we are interested in the overall behavior of the model (competition between different sources of attention). The properties studied are derived from classical constraint usually defined:

- stability: do the values of the dynamical system stay within their nominal range when the different parameters of the model are changed ?

Paramètre	Default value	Min stable value	Max stable value
preys natality b	0.007	0.006	0.013
Preys mortality m_C	0.3	0.3	0.36
Predation s	0.025	0.017	0.05
Predators natality m_I	0.5	0.1	1.5
Positive feedback w	0.001	0	0.003

Table 2. Stability range of the main parameters of the preys / predators system.

- reproducibility: as discrete dynamical system can have a chaotic behavior, what is the influence of the various parameters of the model (in particular noise) on the variability of the focus paths generated during different simulations on the same data ?
- scene exploration: which parameters do influence the scene exploration strategy of our model?
- system dynamics: how can we influence the reactivity of the system? In particular how do we deal with mean fixation time?

For all of these properties we have also studied the influence of top-down feedback.

All the measures presented in this section were done on two image databases. The first one is proposed by Bruce.[11] It is made up of 120 color images which contexts are streets, gardens, vehicles or buildings, more or less salient. The second one, proposed by Le Meur,[3] contains 26 color images. They represent sport scenes, animals, building, indoor scenes or landscapes. Unless otherwise stated, the system is run using the parameters define in Table 1.

2.5.1 Stability

Volterra-Lotka equations are only stable in a predefined range of parameters values.[?] This statement is also true for our attention model. For example, if preys birth rate is too small compared to predation rate and natural mortality of predators is high, neither preys or predators will see their populations grow.

We have studied the stability of our system by monitoring the mean value of the preys maps C_n and of the predator map I. If these values stay within a finite range, the system is stable. Table 2 gives an overview of the system behavior for different of values of natality, mortality and predator parameters b, m_C, m_I, s. Outside of the stability ranges defined in this table preys and / or predators population gradually saturate.

2.5.2 Reproducibility

Since we are using a discrete dynamical system and because we have added a random map when computing the growth factor of the attentional system, our model is nondeterministic. This behavior is interesting because it simulates the natural variability observed when performing multiple eye-tracking experiments on the same person and the same data set. It is also a way to adjust the curiosity of our attentional system by encouraging the exploration of relatively low saliency areas of an image.

However, giving more "curiosity" to our system also leads to less reproducibility. In order to study this phenomenon, we have used the same measures as when studying attentional models plausibility. We have

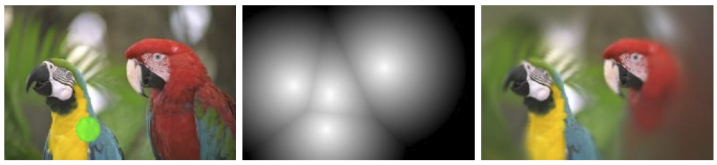
t=11 

Figure 5. Example of image "reconstruction". Left, source image and current focus point; middle, blur mask; right, "reconstructed" image.

compared heatmaps generated from eye-tracking measures, with heatmap generated from various simulations with our model. We have used classical similarity / dissimilarity measures: cross-correlation[3]), Kullback-Leibler divergence ([12]) and normalized scanpath salience ([13]).

During our experiments, we used ground truth heatmaps included in Bruce and LeMeur datasets. Simulated heatmaps were generated using the same method as for ground truth maps : integrating all focalizations on a single maps, and then filtering this maps with a Gaussian filter $G_{\sigma x,\sigma y}$ where:

$\sigma x = \sigma y = 0.3 \times foveaSize \times max(W, H)$

With W and H the width and height of the source image, and $foveaSize = 0.15$ (which correspond to a Gaussian width of approximately 15 of the source image).

As the number of parameters studied is important (retinal filter, central bias, diffusion, hysteresis, noise, positive feedback and top-down feedback) we have decided not to include detailed results of our measures in this article. A summary of the analysis of these results is however available in Table 4.

2.5.3 Scene exploration

Scene exploration validation is based on a measure of the quantity of information lost between the original image and a "reconstruction" of this image through the evolution of the attentional focus. To "reconstruct" the image during the dynamic simulation, we start from a completely blurred image and "add" details from the image in areas which get the focus. The "reconstruction" becomes sharper and sharper through time evolution.

Actually, we update a blurring mask Mb which maximum values represents non blurred areas and minimum values represent highly blurred area. Its iterative construction is similar to the one used for M_{visit} equation:

$$M_b(x, y, t) = \max(M_b(x, y, t-1), N_{Levels} - \min\left(\frac{dist(x, y, x_f, y_f)}{BlurSize}, N_{Levels}\right) \tag{11}$$

Reconstructed image I_R is then generated through a convolution between the source image I_S and a mean filter B_s :

$$I_R(x, y) = I_S(x, y) * B_{s(x,y)}$$

with s the size of the filter, and $s(x, y) = 2^{N_{Levels} - M_b(x,y)}$

Examples of image "reconstruction" are presented Figure 5.

After generating these "reconstructed images", we have used the minimum description length (MDL) principle inspired by ([14]). Following this principle, the simpler a data is, the easier it is to compress. We have decided to adapt this principle to images using two compression technics: JPEG (lossy compression) and PNG (lossless compression). An estimator between 0 and 1 is obtained at any t :

$$InformationRatio_{JPEG} = \frac{size\left(compress_{JPEG}(I_S)\right)}{size\left(compress_{JPEG}(I_R)\right)} \tag{12}$$

$$InformationRatio_{PNG} = \frac{size\left(compress_{PNG}(I_S)\right)}{size\left(compress_{PNG}(I_R)\right)} \tag{13}$$

Feedback	PNG			JPG		
	t=50	t=150	t=300	t=50	t=150	t=300
-1.0	0,899	0,963	0,978	0,714	0,860	0,911
-0.8	0,896	0,968	0,983	0,704	0,872	0,928
-0.6	0,899	0,970	0,984	0,707	0,879	0,928
-0.4	0,865	0,919	0,937	0,661	0,772	0,812
-0.2	0,820	0,869	0,886	0,600	0,671	0,703
0.0	0,876	0,921	0,940	0,694	0,793	0,836
0.2	0,894	0,959	0,977	0,694	0,850	0,907
0.4	0,901	0,962	0,977	0,717	0,861	0,909
0.6	0,896	0,958	0,975	0,704	0,846	0,901
0.8	0,869	0,943	0,967	0,658	0,810	0,873
1.0	0,882	0,954	0,972	0,682	0,835	0,891

Figure 6. Influence of feedback on scene exploration .

Parameters	Defaut	CentralBias=0.5	FeedBack=-1.0	FeedBack=1.0	Step = 0.1 Iterations=1	Hyster
Fixation time (ms)	70	143	53	417	807	1

Table 3. Influence of a few parameters on simulated fixations time

with I_S the source image and I_R the reconstructed image.

The feedback mechanism aims at controlling the way visual scene is explored. Results obtained from measures described above (JPEG and PNG ratio) confirm this expected behavior (Table 6):

- positive feedback leads to a faster but not necessarily a more exhaustive exploration. Without any feedback, the scene is already almost covered after 300 simulation steps);
- negative feedback can greatly reduce the explored area. For a feedback value $F_{feedback} = -1$, even after 300 simulation step, the scene exploration ratio is still inferior to the one obtained without any feedback.

Other parameters also play a role in scene exploration, in particular noise and central bias. Their influence is summarized in Table 4.

2.5.4 Dynamics

Even if our system does not generate saccades and fixations which are directly comparable to human eye fixations (we do not take into account eye movements constraints), we can estimate the average time $FixationTime$ between two changes of position of the simulated focus of attention. New fixations detection can be done:

- when the position of the focus of attention changes, regardless of the distance to the next position;
- if the distance between the current focus and the next exceeds a threshold S_{Fixing}.

We have chosen the second method because it allows canceling the effect of small movements that would otherwise bias the estimation of the average time between two fixations. The value of S_{Fixing} (15% of the longest side of the source image) was determined so as to be consistent with the $foveaSize$ parameter used to generate heatmaps from the focus of attention output by our attention model.

We measured the effect of the different parameters of our model on the mean fixation time for Bruce and Le Meur image databases. The results of this study are summarized in Table 4. Table 3 gives a few examples for some representative parameters. These results should be compared to mean human fixation time: 300 ms.

Dynamics can be fine tuned using many parameters. But the most efficient ones are differential equation evolution parameters (simulation step and number of sub-iterations), feedback, and central bias. However, these parameters don't have the same side-effects on other properties (plausibility, scene exploration, etc.).

Parameters	Default value	Fidélity	Reproducibility	Exploration	Dynamics
Retinal blur	no	$\nearrow$	$\searrow$	$\nearrow$	$\rightarrow$
Central bias (g)	0.1	$\uparrow$	$\rightarrow$	$\uparrow$	$\downarrow$
Diffusion (f)	0.25	$\rightarrow$	$\searrow$	$\nearrow$	$\searrow$
$T_{Hysteresis}$	0	$\rightarrow$	$\rightarrow$ / $\searrow$	$\searrow$	$\searrow$
Noise (a)	16	$\times$	$\times$	$\times$	$\rightarrow$
Positive feedback (w)	0.5	$\uparrow$ / $\downarrow$	$\downarrow$	$\uparrow$	$\nearrow$
Simulation step	0.001	$\nearrow$ / $\downarrow$	$\searrow$ / $\downarrow$	$\nearrow$ / $\searrow$	$\nearrow$ / $\downarrow$
# of sub-iterations	1/3	$\times$	$\times$	$\times$	$\uparrow$
Top-down feedback	3	$\times$	$\times$	$\times$	$\uparrow$

Table 4. Summary of the influence of each parameter of the model

2.5.5 Summary

Table 4 summarizes the influence of each parameter one the system behavior. We have not mentioned the influence of birth and death factors b, s, M_C and M_I since the only affect stability.

The arrows used have the following meanings:

$\uparrow$ strong positive influence.

$\downarrow$ strong negative influence.

$\nearrow$ weak positive influence.

$\searrow$ weak negative influence.

$\rightarrow$ no significant influence.

$\times$ non tested / theoritically non influent.

In the case of the retinal filter, arrows correspond to the influence of activating the filter. Arrows separated by a slash (for example: $\rightarrow$ / $\searrow$) represent a first type of influence for small increases of the parameters, followed by a second type of influence for higher increases.

2.6 Model Plausibility

2.6.1 Comparison to existing models

In ([5]) we have presented a subjective validation of the plausibility of our model. In this article, we confirm the latter by a more classical objective evaluation. This validation consists in checking the plausibility of the system, i.e. checking if it is apparently reasonably valid, and truthful.

Cross-correlation, Kullback-Leibler divergence and normalized scanpath saliency were used to compare 6 algorithms to an eye-tracking ground-truth (Figure 7). 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 ([3]), in its "coherent normalization" version;
- the AIM model of Bruce and Tsotsos([11]);
- the NVT model of Itti ([1]).

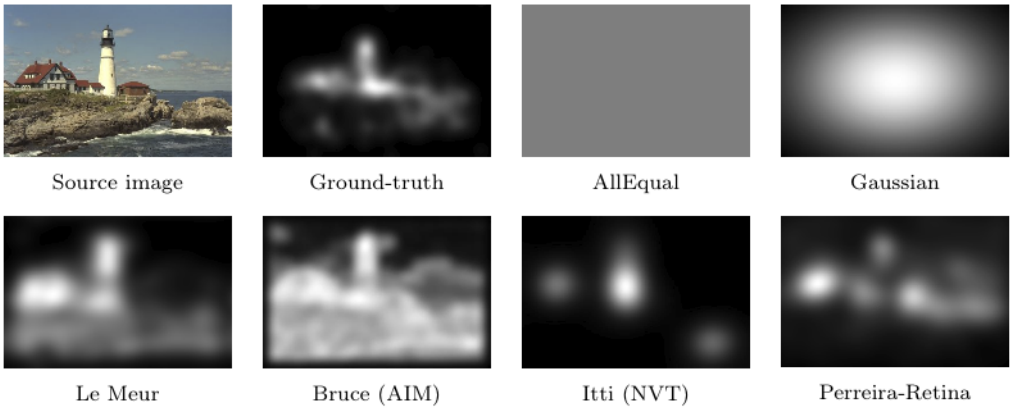


Figure 7. Plausibility of computational models. Comparison of the saliency maps and heatmaps generated by the 6 algorithms tested.

	CC		KLD		NSS	
	Bruce	LeMeur	Bruce	LeMeur	Bruce	LeMeur
Bruce	0.4	0.45	1.59	1.08	0.98	0.89
LeMeur	0.37	0.43	1.61	1.08	0.90	0.84
Itti	0.31	0.27	2.74	2.52	0.79	0.53
AllEqual	0.00	0.00	2.15	1.55	0.00	0.00
Gaussian	0.46	0.60	1.55	0.94	1.02	1.10
Perreira Retina	0.43	0.38	1.61	1.4	1.17	0.73

Table 5. Comparison of different algorithms to ground-truth. (Please note that for KLD, lower values mean more plausibility)

- our model (with fast retinal blur).

All models were tested using their default parameters.

Table 5 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. This effect can however be attenuated by using an alternative metric: the area under ordinal dominant score ([15]);
- Model's performances are comparable to other state of the art models and even outperform them on Bruce database (NSS measure).

2.7 Influence of parameters

We have shown that our model is as plausible as other state of the art models. However, this model (and in particular its dynamical system) depends on numerous parameters. Table 6 summarizes the influences of some of these parameters on the plausibility of the model. The following conclusions can be drawn:

Parameters	CC		KLD		NSS		Mean Gain
	Bruce	LeMeur	Bruce	LeMeur	Bruce	LeMeur	
Defaut	0.35	0.30	1.80	1.76	0.95	0.56	0%
Retinal filter	0.43	0.38	1.61	1.40	1.17	0.73	22%
CentralBias=0.00	0.2	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	1.7	0.95	0.55	0%
Diffusion=0.5	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%

Table 6. Influence of model parameters on plausibility. Gains are relative to default parameters defined in Table 1.

Gain / Feedback	-1.0	-0.8	-0.6	-0.4	-0.2	0	0.2	0.4	0.6	0.8	1
CC+NSS	-19%	-18%	-16%	-13%	-7%	0%	9%	10%	12%	14%	20%
KLD	-1%	0%	1%	1%	3%	0%	-10%	-26%	-40%	-55%	-64%

Table 7. Influence of top-down feedback on plausibility. Gains are relative to the bottom-up only version of the model.

- 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 attentionnal focus;
- using central biasing in an attention model can improve significantly its plausibility, but this bias is partly due to the experimental protocol;
- 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.

2.8 Influence of Feedback

The influence of feedback on the plausibility of our model is quite tricky to explain. Indeed, as can be seen in Table 7, the mean changes observed seem contradictory: • for cross correlation and normalized scanpath salience, the use of top-down feedback appears to improve the plausibility of our model; • for the Kullback Leibler divergence, it seems rather to reduce it.

Our explanation is the following: NSS and correlation are similarity measures while Kullback-Leibler divergence is a dissimilarity measure, as a consequence they react differently to a change in the exploration strategy of our model.

It is therefore difficult to judge the influence of feedback, as it is twofold. However, we can conclude that feedback slightly improves correlation of our model with ground truth in the most salient areas, the price of increasing the difference with ground truth in the less salient areas.

3. CONCLUSION

In this article, we have presented a complete implementation and evaluation system of a computational model of attention for computer vision. Concerning implementation, we have shown that preys / predators models provide good properties for simulating the dynamic competition between different kinds of information. We have described the architecture of our model which can be divided in two parts. The first one is hierarchical, it

improves the model of L. Itti by providing much faster processing times while allowing the computation of more scales during its multi-resolution analysis of the scene. The second part is our major contribution: it makes use of a dynamical system (inspired from a preys / predators competition analogy) to handle the fusion of conspicuity maps generated by the first part of the model. This dynamical system is also used to generate a focus point at each time step of the simulation. Concerning evaluation, we have presented different results (cross-correlation, Kullback-Leibler divergence, normalized scanpath salience) that demonstrate that, in spites of being fast and highly configurable, our results are as plausible as existing models designed for high biological fidelity.

REFERENCES

1. Itti, L., Koch, C., Niebur, E., and Others, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence* **20**(11), 1254–1259 (1998).
2. Fox, M. D., Snyder, A. Z., Vincent, J. L., and Raichle, M. E., "Intrinsic Fluctuations within Cortical Systems Account for Intertrial Variability in Human Behavior," *Neuron* **56**, 171–184 (Oct. 2007).
3. Le Meur, O., Le Callet, P., Barba, D., and Thoreau, D., "A coherent computational approach to model bottom-up visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(5), 802–817 (2006).
4. Tatler, B. W., "The central fixation bias in scene viewing : Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of Vision* **7**, 1–17 (2007).
5. Perreira Da Silva, M., Courboulay, V., Prigent, A., and Estraillier, P., "Evaluation of preys / predators systems for visual attention simulation," in [*VISAPP 2010 - International Conference on Computer Vision Theory and Applications*], 275–282, INSTICC, Angers (2010).
6. Perreira Da Silva, M., Courboulay, V., and Estraillier, P., "IMAGE COMPLEXITY MEASURE BASED ON VISUAL ATTENTION," in [*Ieee International Conference On Image Processing*], 3342–3345 (2011).
7. Frintrop, S., *VOCUS:AVisual Attention System for Object Detection and Goal-Directed Search*, phd, University of Bonn (2005).
8. Navalpakkam, V., Arbib, M., and Itti, L., "Attention and scene understanding," in [*Neurobiology of Attention*], Itti, L., Rees, G., and Tsotsos, J., eds., (December 2004), 197–203, ACADEMIC PRESS (2005).
9. Navalpakkam, V. and Itti, L., "Top-down attention selection is fine grained," *Journal of Vision* **6**(11), 4 (2006).
10. Torralba, A., Oliva, A., Castelhano, M. S., and Henderson, J. M., "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search.," *Psychological review* **113**, 766–86 (Oct. 2006).
11. Bruce, N. D. B. and Tsotsos, J. K., "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision* **9**(3), 5 (2009).
12. Tatler, B. W., Baddeley, R. J., and Gilchrist, I. D., "Visual correlates of fixation selection: effects of scale and time.," *Vision research* **45**, 643–59 (Mar. 2005).
13. Peters, R., Iyer, A., Itti, L., and Koch, C., "Components of bottom-up gaze allocation in natural images," *Vision Research* **45**, 2397–2416 (2005).
14. Rissanen, J., "Modeling by shortest data description," *Automatica* **14**, 465–471 (1978).
15. Berg, D. J., Boehnke, S. E., Marino, R. A., Munoz, D. P., and Itti, L., "Free viewing of dynamic stimuli by humans and monkeys," *Journal of Vision* **9**, 1–15 (2009).