Implémentation et évaluation d'un modèle d'attention pour la vision adaptative

Matthieu Perreira Da Silva [1], **Vincent Courboulay** [2]

1. IRCCyN, Université de Nantes
Rue Christian Pauc - BP 50609, F-44306 Nantes cedex 03
matthieu.perreiradasilva@univ-nantes.fr

2. L3I, Université de La Rochelle
Avenue M. Crépeau, F-17042 La Rochelle cedex 01
vincent.courboulay@univ-lr.fr

RÉSUMÉ. *Dans le domaine de l'analyse de scène en vision par ordinateur, un compromis doit être trouvé entre la qualité des résultats attendus et les ressources allouées pour effectuer les traitements. Une solution flexible consiste à utiliser un système de vision adaptatif capable de moduler sa stratégie d'analyse en fonction de l'information disponible et du contexte. Dans cet article, nous décrivons comment concevoir et évaluer un système d'attention visuelle conçu pour interagir avec un système de vision de façon à ce que ce dernier adapte ses traitements en fonction de l'intérêt (de la saillance) de chaque élément de la scène. Nous proposons également un nouvel ensemble de contraintes nommé PAIRED, permettant d'évaluer l'adéquation du modèle à différentes applications. Nous justifions le choix des systèmes dynamiques par leurs propriétés intéressantes pour simuler la compétition entre différentes sources d'informations. Nous présentons enfin une validation à travers différentes métriques montrant que nos résultats sont rapides, hautement configurables et pertinents.*

ABSTRACT. *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. 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. We 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. We propose a new set of constraints named PAIRED to evaluate the adequacy of a model with respect to its different applications. We justify why dynamical systems provide good properties for simulating the dynamic competition between different kinds of information. We present different results that demonstrate that our results are fast and highly configurable and plausible.*

MOTS-CLÉS : *modèle dynamique d'attention, vision adaptative, implémentation, évaluation.*

KEYWORDS: *attention model, dynamical model, adaptive vision, implementation, evaluation.*

DOI:10.3166/TS.28.611-641

Extended abstract

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 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.

In the first two sections, we start by giving a brief overview of the main theories and concepts of human visual attention and we provide the forces and weaknesses of state of the art attention models. This analysis is based on their potential of integration into adaptable computer vision system. We propose a new set of constraints called 'PAIRED' to evaluate the adequacy of a model with respect to its different applications.

In a third section, we provide an in-depth description of our model and its implementation. 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 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 maxi-

mize 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 not cover not its plausibility compared to human eye fixations (already studied in (Perreira Da Silva *et al.*, 2011)), but the influence of each parameter on a set of properties:

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

– 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 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.

1. Introduction

Alors que les systèmes de vision par ordinateur deviennent de plus en plus puissants, ils restent toujours loin des systèmes mis en place par la Nature. Chez les Hommes, les mécanismes de l'évolution ont fait émerger les mécanismes attentionels. Ils sélectionnent les informations les plus pertinentes afin de réduire la charge mentale et les ambiguïtés de la scène observée. Une voie prometteuse d'avancée de la vision par ordinateur consiste donc à étudier et appliquer les découvertes faites chez l'Homme.

Par exemple, dans le domaine de l'analyse de scène, il est souvent nécessaire de trouver un compromis entre la qualité des résultats attendus et les ressources allouées pour effectuer les traitements. Ce choix est habituellement fait au moment de la conception du système au moyen d'algorithmes et de paramètres prédéfinis. Cependant cette façon de procéder limite la généricité du système. Une solution flexible consiste à utiliser un système de vision adaptatif capable de moduler sa stratégie d'analyse en fonction de l'information disponible et du contexte. Un tel système requiert néanmoins un mécanisme permettant d'explorer la scène rapidement et efficacement.

Dans cet article, nous proposons un modèle computationnel d'attention visuelle permettant de mettre en œuvre ce mécanisme de façon rapide et adaptable. Dans la section 2, nous présentons un rapide résumé des principaux modèles computationnels d'attention, ainsi que leurs forces et faiblesses sous le filtre d'un ensemble de contraintes nommé PAIRED. En section 3, nous décrivons le modèle et son implémentation. Enfin, en section 4, nous présentons les résultats de l'évaluation de notre modèle. Comme celui-ci est hautement configurable, son évaluation porte sur l'in-

fluence de chaque paramètre sur un ensemble de propriétés (stabilité, reproductibilité, exploration, comportement dynamique).

2. Modèles computationnels d'attention visuelle

2.1. L'attention visuelle

Ce que nous percevons est-il vraiment le reflet de ce que nous voyons ? Notre vision du monde semble précise, continue et cohérente. Pourtant, l'étude des différents composants de notre système visuel laisse apparaître une situation bien différente (Rensink, 2000). L'œil et la rétine sont loin de capturer une image fidèle du monde. Ainsi, pour pouvoir interpréter ce que nous voyons, notre système visuel a mis en place des mécanismes de sélection et d'optimisation du traitement de l'information. C'est le cas de l'attention, qui voit l'explication de son origine disputée par deux théories duales.

La première, et la plus largement répandue (Treisman, 1969 ; Tsotsos, 1990), suppose que comparativement à la quantité de données qu'il a à traiter, notre cerveau a une capacité de traitement limitée. Ses partisans soutiennent l'idée que si notre cerveau était plus gros, nous n'aurions pas besoin de mécanismes attentionnels. Dans ce cadre, l'attention permet de sélectionner une partie de l'information afin de ne pas surcharger notre système cognitif.

La seconde théorie (Allport, 1987 ; Heijden, Bem, 1997) consiste à considérer que nos capacités de traitement ne sont en rien limitées. La perception seule ne nous sert à rien, nous construisons une représentation du monde afin de pouvoir interagir avec celui-ci par l'intermédiaire de nos actions. Or, nos yeux ne perçoivent précisément qu'une petite partie de l'environnement, et nos mains ne peuvent manipuler qu'un (voire deux) objets simultanément. Ce sont donc nos capacités d'action qui sont limitées et imposent une sélection de l'information perçue afin de pouvoir la traiter correctement.

Quelle que soit la façon de justifier le processus attentionnel, il est indiscutablement nécessaire à notre perception visuelle car il permet de lever les ambiguïtés (Van Rullen, Koch, 2005). Il est également indispensable à la construction d'une représentation cohérente de notre environnement (Rensink, 2000) et à la détection de changements dans celui-ci.

La complexité du phénomène attentionnel ouvre la porte à une multiplicité d'interprétations théoriques. A la fin des années 1990, la puissance de calcul des ordinateurs est devenue suffisante pour envisager une mise en œuvre computationnelle de ces multiples modèles. De nouvelles théories continuent aussi à voir le jour, et sont presque systématiquement accompagnées d'une implémentation sur ordinateur permettant la simulation et donc une nouvelle forme de validation. Dans la prochaine section, nous explorons une large sélection de ces modèles afin de déterminer l'adéquation entre les différents types d'approches proposés et le domaine d'application visé : la vision par ordinateur.

2.2. Un modèle idéal...

Le monde que nous percevons est complexe : les informations à analyser sont nombreuses, souvent ambiguës et sans cesse changeantes. Pour gérer cette complexité, le cerveau doit trouver des principes simplificateurs, permettant de gérer rapidement et efficacement les différentes situations auxquelles il est confronté. Alain Berthoz (2009) propose de regrouper les différents mécanismes permettant la gestion de cette complexité sous une théorie unique : la *simplexité*. D'après cette théorie un système simplexe doit : séparer les différentes fonctions/être modulaire ; être rapide ; être fiable ; être flexible ; posséder de la mémoire et avoir des propriétés de généralisation. Compte tenu de ses propriétés, l'attention visuelle répond parfaitement à cette caractérisation.

D'un point de vue computationnel, il est classiquement nécessaire d'utiliser certains critères pour classer ou évaluer des modèles d'attention. Habituellement, le temps de traitement, la robustesse ou la vraisemblance sont utilisés. Néanmoins, nous avons décidé d'utiliser les propriétés d'un système simplexe afin d'en tirer un ensemble de contraintes qui permettra d'évaluer l'adéquation d'un modèle à l'égard de ses différentes applications. Nous avons nommé cet ensemble PAIRED. Il est composé des éléments suivants :

- plausible comparé au modèle humain ;
- adaptable à différents contextes ;
- invariant à travers différentes transformations (rotation, translation et échelle) ;
- rapide pour calculer le focus d'attention ;
- extensible concernant sa capacité à intégrer de nouvelles caractéristiques ;
- dynamique et fournissant des résultats à tout instant.

Une fois ces critères présentés, on peut définir leur importance pour différentes applications (cf. tableau 1). Un point signifie une contrainte faible ; une forte est représentée par trois points. Nous utilisons cet ensemble de contraintes pour évaluer l'adéquation entre les différentes familles de modèles d'attention et notre application cible : la vision par ordinateur, le modèle *idéal* devant satisfaire à toutes les contraintes.

Tableau 1. Contraintes PAIRED liées à différents types d'application

	Plausible	Adaptable	Invariant	Rapide	Extensible	Dynamique
Ergonomie, publicité	●●●	●	●	●	●●	●●
Vision	●●	●●●	●●	●●●	●●●	●●●
CBIR	●	●	●●●	●●	●●●	●
Traitement d'images	●●	●●	●●	●●	●●●	●●

2.3. Une taxonomie des modèles existants

Nous avons choisi de séparer la présentation des différents modèles computationnels d'attention visuelle en deux familles, basées sur des concepts duaux. Dans cette section, nous effectuons un panorama (non exhaustif) des différents modèles existants dans ces deux familles.

2.3.1. Modèles d'attention distribués

Ces modèles ont été initiés par les neuroscientifiques connexionnistes. Ils sont généralement développés en utilisant une approche neuromimétique. Leur niveau de granularité peut descendre jusqu'à celui des neurones (Deco, 2004). La majorité de ces modèles sont inspirés par la théorie de la compétition biaisée proposée par (Desimone, Duncan, 1995). Pour ces modèles, l'attention n'est pas un faisceau mental (Treisman, Gelade, 1980) traversant la scène visuelle à grande vitesse : les objets dans le champ visuel sont en concurrence pour l'attribution de ressources cognitives limitées.

L'attention peut également être modélisée de façon moins compétitive et plus centralisée. Nous discutons des modèles utilisant ce paradigme dans le reste de cette section.

2.3.2. Modèles à représentation centrale

Ces modèles s'inscrivent dans la continuité des travaux initiaux de (Treisman, Gelade, 1980). Selon cette théorie, l'attention est encodée dans une carte centrale qui représente le champ de vision dans son intégralité. Bien que des études plus récentes n'aient toujours pas prouvé l'unicité de la représentation de la saillance dans notre cerveau, ce modèle est très populaire, computationnellement efficace et a un pouvoir explicatif prouvé.

Le nombre de modèles basés sur ce paradigme étant important, nous proposons une scission taxinomique en cinq sous-familles :

– modèles hiérarchiques construits à partir de cartes fonctionnelles progressivement combinées pour fournir une seule carte d'attention. De nombreux modèles influents sont basés sur cette approche (Itti *et al.*, 1998 ; Koch, Ullman, 1985 ; Walther, Koch, 2006 ; Le Meur *et al.*, 2006 ; Frintrop *et al.*, 2007 ; Belardinelli *et al.*, 2009) ;

– modèles statistiques et probabilistes qui considèrent comme saillant le moins fréquent ou le moins probable des évènements caractéristiques ou des objets dans une scène (Hamker, 2005 ; Torralba *et al.*, 2006 ; Baldi, Itti, 2005 ; Avraham, Lindenbaum, 2010) ;

– modèles basés sur la théorie de l'information, animés par des modèles probabilistes et maximisant la quantité d'information acquise (Gilles, 1996 ; Kadir, Brady, 2001 ; Bruce, Tsotsos, 2009 ; Park *et al.*, 2002 ; Mancas, 2007) ;

– modèles connexionnistes généralement basés sur des réseaux de neurones et fournissant un focus d'attention dynamique (Ahmad, 1992 ; Mozer, Sitton, 1998 ; Vitay *et al.*, 2005) ;

– modèles algorithmiques qui proposent des méthodes qui sont généralement liées à une application spécifique (Lopez *et al.*, 2006 ; Orabona *et al.*, 2008 ; Aziz, Mertsching, 2009).

Il existe d'autres classifications (Tsotsos, 2007 ; Le Meur, Le Callet, 2009), et certains algorithmes sont difficiles à placer dans une seule catégorie. Le tableau 2 résume les avantages et les inconvénients de chaque type de modèle d'attention. Malgré la diversité des approches utilisées dans chacune des familles, il est possible d'identifier certaines caractéristiques communes.

2.4. Étude comparative

Le tableau 2 résume les avantages et les inconvénients des deux grandes approches citées précédemment. L'approche distribuée est proche de la réalité biologique, gère efficacement le problème de la concurrence, mais elle est plus lourde à mettre en œuvre et à étendre. L'approche centralisée est généralement plus facile à calculer, mais elle ne tient pas compte de la dynamique (évolution du focus de l'attention dans le temps) et nécessite l'ajout de méthodes connexionnistes lourdes (*winner takes all* + inhibition de retour) pour générer des fixations.

Tableau 2. Avantages et inconvénients des modèles centralisés et distribués

Type de modèle	Avantage(s)	Inconvénient(s)
Distribués	Gestion de la compétition entre les sources d'information Gestion de la dynamique	Complexité Ajout de nouvelles caractéristiques plus délicat
Centralisés	Efficacité computationnelle Facilement extensible	Gestion de la dynamique

Le tableau 3 présente la façon dont chacune des familles de modèles d'attention répondent aux exigences des contraintes PAIRED. On peut remarquer qu'aucune approche n'est idéalement adaptée à l'application cible : vision par ordinateur. Compte tenu des propriétés de fidélité, d'invariance, d'adaptation dynamique des modèles distribués, et des propriétés de vitesse et d'évolutivité des modèles hiérarchiques, nous pouvons conclure qu'une approche hybride entre ces deux alternatives permettrait d'obtenir le modèle désiré.

Une telle approche a déjà partiellement été explorée par certains modèles connexionnistes : ils combinent généralement un modèle hiérarchique de génération de carte de saillance et une approche distribuée pour gérer la dynamique de l'attention. Toutefois, dans ces modèles, c'est le système hiérarchique qui est responsable de la concurrence entre les différentes caractéristiques (intensité, couleur, orientation, etc). Par conséquent, nous ne bénéficions pas du principe de la concurrence entre ces différentes sources.

Pour surmonter ce problème, nous proposons d'éviter d'utiliser une représentation centrale de la saillance. Au lieu de cela, notre principale contribution consiste à

mettre en concurrence les cartes de caractéristiques. Dans la suite de cet article, nous présentons cette solution et étudions ses propriétés au vue des critères PAIRED.

Tableau 3. Adaptation des différents modèles aux contraintes d'un système de vision. La première ligne du tableau correspond aux objectifs que nous avons définis. En gris : les critères atteints ou dépassés par les différentes familles de modèles

	Plausible	Adaptable	Invariant	Rapide	Extensible	Dynamique
Objectif	⋆⋆	⋆⋆⋆	⋆⋆	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆
Distribués	●●●	●●●	●●●	●	●	●●●
Hiérarchiques	●●	●●	●●	●●	●●●	●
Statistiques	●●	●	●●	●●	●●	●
Théorie de l'information	●●	●	●●	●●	●●	●
Connexionnistes	●●	●●●	●●	●●	●●	●●●
Algorithmiques	●	●●	●●	●●●	●	●

3. Un modèle d'attention visuelle hiérarchique compétitif

3.1. Modélisation proies/prédateurs et attention visuelle

Comme mentionné à la fin de la section précédente, nous proposons un nouveau modèle hybride qui permet d'étudier l'évolution temporelle du focus attentionel. Le système visuel est inspiré des modèles d'attention centralisés hiérarchiques, et en particulier des modèles de (Itti *et al.*, 1998) et (Frintrop, 2005). La scène visuelle est décomposée en différentes caractéristiques selon une approche multirésolution. Dans un souci de performance, nous avons privilégié l'efficacité de calcul à la fidélité biologique. L'approche suivie est donc computationnaliste : les différentes caractéristiques sont calculées à partir de filtres numériques.

Le système génère, pour chacune des caractéristiques prises en compte (intensité, couleur, orientation et mouvement), un certain nombre de cartes représentant les éléments les plus saillants. Les attributs calculés par ce système peuvent être utilisés par le système attentionnel et/ou par un système de vision de plus haut niveau (reconnaissance/suivi d'objets par exemple). Le système attentionnel est principalement compétitif et d'inspiration connexionniste : la compétition entre les différentes cartes de caractéristiques est effectuée par l'interaction de différentes *proies* et *prédateurs* au sein d'un même « écosystème ». On s'éloigne ici des systèmes d'attention hiérarchique pour se rapprocher des systèmes distribués de compétition biaisée.

La figure 1 donne un aperçu de l'enchaînement des différents traitements appliqués par notre modèle d'attention. Le vocabulaire utilisé dans ce schéma reprend les termes proposés par (Itti *et al.*, 1998) dans leurs travaux.

Les équations proies/prédateurs sont particulièrement bien adaptées pour une telle tâche. Les principales raisons de ce choix sont les suivante :

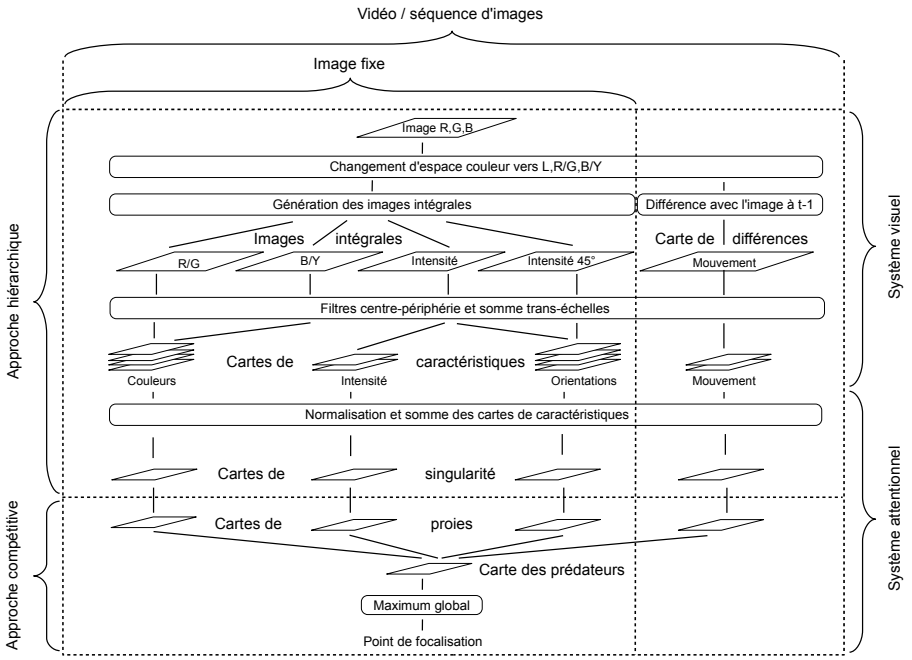

Figure 1. Schéma général du système d'attention visuelle

– les modèles proies/prédateurs sont dynamiques, ils intègrent intrinsèquement l'évolution temporelle de leurs activités. Ainsi, le focus d'attention, vu comme le *lieu* de densité maximale de prédateurs, peut évoluer de manière dynamique ;

– sans aucun objectif (information de haut niveau ou de *prégnance*), le choix d'un procédé de fusion des cartes de caractéristiques est difficile. Une solution consiste à développer la concurrence entre ces cartes dans l'attente d'un équilibre naturel entre les proies et les prédateurs, ce qui reflète la compétition entre l'émergence et l'inhibition des éléments qui engagent ou pas notre attention ;

– les systèmes dynamiques discrets peuvent avoir un comportement chaotique. Bien que cette propriété ne soit pas souvent intéressante, elle l'est pour nous. En effet, elle permet l'émergence de voies originales d'exploration de la scène visuelle, même dans les zones non saillantes, ce qui pourrait se traduire par de la curiosité. Dans la suite, nous allons décrire et présenter l'implémentation de notre modèle.

3.2. Description du modèle et implémentation

La partie « système visuel » de notre modèle d'attention permet de générer les cartes de caractéristiques de deux façons :

– de manière classique, en traitant de la même façon l'ensemble du champ visuel ;

– en utilisant un pseudo-flou rétinien, permettant de « simuler » la résolution variable de la rétine sans faire appel à la représentation *log polaire* généralement utilisée dans ce cas (Sun *et al.*, 2008).

3.2.1. Prétraitements

Chez l'homme, l'information visuelle est traitée dans la rétine par des cellules achromatiques et chromatiques, sensibles respectivement à l'intensité et aux contrastes rouge/vert et bleu/jaune. Nous devons donc convertir les images RGB $I_{R,G,B}$ que notre système attentionnel reçoit en entrée dans un espace couleur plus adapté. Certains modèles computationnels (Frintrop, 2005) utilisent des espaces couleur perceptuels de type Lab afin d'obtenir un système le plus fidèle possible au système visuel humain. Notre objectif étant de créer un modèle rapide mais plausible, nous avons privilégié l'espace $L, R/G, B/Y$ utilisé dans (Itti *et al.*, 1998). Une fois la conversion des couleurs effectuée, nous disposons d'un canal achromatique I_L et de deux canaux chromatiques $I_{R/G}$ et $I_{B/Y}$. Ceux-ci vont servir de base au calcul des cartes de caractéristiques d'intensité, couleur, orientation et mouvement *via* une approche multirésolution. Pour accélérer les calculs, nous avons choisi une approche alternative, inspirée de (Frintrop *et al.*, 2007), qui permet de remplacer le calcul des pyramides multirésolutions par des images intégrales [1] (Viola, Jones, 2002). A la différence de (Frintrop *et al.*, 2007) qui avait d'abord bâti son modèle à partir de pyramides et ensuite optimisé certaines parties, nous utilisons cette solution dès la conception et généralisons son utilisation à l'ensemble des cartes, sauf celle de mouvement (justification fournie par la suite).

Compte tenu des différentes caractéristiques à prendre en compte, nous devons précalculer quatre images intégrales :

– une image d'intensité II_L, issue de I_L pour le calcul des cartes d'intensité et des cartes d'orientation à 0° et 90° ;

– deux images couleur $II_{R/G}$ et $II_{B/Y}$ issues de $I_{R/G}$ et $I_{B/Y}$ pour le calcul des cartes couleur ;

– une image d'intensité avec rotation de 45° IIR_L issue de I_L permettant de calculer les cartes d'orientation à 45° et 135°. Les images intégrales « basiques » ne permettant pas de calculer ce type d'information, nous avons utilisé une version étendue proposée par (Lienhart, Maydt, 2002).

Les traitements effectués ensuite dépendent de la prise en compte de la résolution spatiale variable de la rétine humaine. Dans la section suivante, nous décrivons le cas classique, sans ce flou rétinien.

1. Les images intégrales permettent de calculer à coût constant la somme ou la moyenne d'une zone rectangulaire quelconque d'une image.

3.2.2. Sans pseudo-flou rétinien

3.2.2.1. Pyramides d'intensité et de couleur

Les modèles d'attention hiérarchiques, dont nous nous sommes inspirés pour la partie « système visuel » de notre modèle, calculent généralement leurs cartes de caractéristiques (couleur et intensité) à l'aide de différences de gaussiennes (*DoG*), permettant d'approximer les champs récepteurs *centre-périphérie* des cellules de la rétine. Ces calculs sont relativement lourds et nous avons choisi de les remplacer par un calcul approximatif combinant images intégrales et différences de zones rectangulaires également appelées différences de boîtes (*DoB*). En effet, les images intégrales ne permettant de calculer « que » des sommes de zones rectangulaires, il nous faut adapter l'ensemble des filtres à ce mode de calcul. En utilisant le procédé décrit ci-avant, on calcule les différences de boîtes à partir de II_L, $II_{R/G}$ et $II_{B/Y}$ pour différents niveaux de résolution r_0 à r_{N-1}. r_0 est le niveau de résolution maximale (image d'origine) et r_{N-1}la résolution la plus grossière. Cette dernière dépend de la taille de l'image d'origine. Pour une image I de taille $W * H$, on aura au plus $N = 1 + log_2(min(W, H)$, arrondi à l'entier inférieur.

Théoriquement, nous devrions utiliser deux filtres : un pour la réponse des cellules centre-périphérie on/off, un autre pour la réponse des cellules Off/On. Cependant contrairement aux cellules humaines, notre représentation informatique peut aisément représenter les nombres négatifs. On peut donc filtrer les images avec un unique filtre (on/off par exemple) : les réponses positives correspondront à la réponse de celui-ci (on/off), les négatives à la réponse de son filtre complémentaire (off/on).

Afin de ne pas perdre d'informations, on stocke les résultats des filtres on/off et off/on dans des pyramides différentes. Le choix de l'une ou l'autre des pyramides s'effectue en fonction du signe de la réponse du filtre centre-périphérie. On génère les pyramides de caractéristiques$P_{L_{on}}$, $P_{L_{off}}$ de la manière suivante :

$$P_{L_{on}}(x, y, r) = \begin{cases} CS_{L,r}(x, y) & si\ CS_{L,r}(x, y) > 0 \\ 0 & sinon \end{cases} \quad (1)$$

$$P_{L_{off}}(x, y, r) = \begin{cases} -CS_{L,r}(x, y) & si\ CS_{L,r}(x, y) \leq 0 \\ 0 & sinon \end{cases} \quad (2)$$

avec $CS_{L,r} = I_L * (B_{1,r} - B_{2,r}) = (I_L * B_{1,r}) - (I_l * B_{2,r})$ le résultat de la convolution (réalisée *via* les images intégrales) entre le canal d'intensité et le filtre « différence de boîte »[2]. On procède de même avec les couples de pyramides P_R, P_G et P_B ,P_Y .

3.2.2.2. Pyramides d'orientation

Comme dans les modèles de (Itti *et al.*, 1998) et (Frintrop, 2005), nous calculons quatre pyramides orientées P_{Ori0}, P_{Ori45}, P_{Ori90} et P_{Ori135} correspondant aux

2. Le produit de convolution étant distributif sur l'addition (et la soustraction).

orientations 0°, 45°, 90° et 135°. Nous avons privilégié la réutilisation des images intégrales calculées lors des prétraitements afin de générer plus efficacement les pyramides orientées. On utilise alors des filtres de type *Harr like*, dont la réponse est sélective en orientation.

3.2.2.3. Pyramides de mouvements

Nous avons pour l'instant décrit les caractéristiques statiques calculées par le modèle. Cependant, lorsqu'une séquence vidéo est présentée à notre système, il est nécessaire qu'il prenne également en compte le mouvement. Celui-ci est généralement estimé à partir du flot optique de l'image. Aussi, bien qu'il existe maintenant des méthodes permettant d'estimer celui-ci en utilisant par exemple la puissance de calcul des cartes graphiques, le calcul du flot optique est une opération relativement complexe, peu compatible avec le traitement temps réel de l'ensemble du système attentionnel.

Du fait des mécanismes *centre-périphérie* mis en œuvre par le système visuel humain, ce sont principalement les différences de vitesse ou de direction entre objets qui sont saillantes. Ainsi, il n'est pas nécessaire de connaître précisément la vitesse et la direction des objets pour y porter attention. En conséquence, nous avons choisi d'utiliser une méthode simple et rapide basée sur les différences entre deux trames successives qui, bien que peu justifiée biologiquement, est un moyen efficace d'obtenir une estimation acceptable du mouvement dans l'image.

Pour obtenir cette estimation, on génère tout d'abord une image I_{Diff_L} représentant la valeur absolue des différences entre le canal intensité de la trame courante et de la trame précédente :

$$I_{Diff_L} = \mid I_{L_t} - I_{L_{t-1}} \mid \tag{3}$$

On calcule ensuite les classiques pyramides centre-périphérie on/off $P_{M_{on}}$ et off/on $P_{M_{off}}$. Il ne serait pas efficace de calculer une image intégrale pour le traitement du mouvement, car son coût de calcul ne serait pas amorti par son utilisation unique. On calcule alors la pyramide multirésolution plus classiquement, en remplaçant toutefois les filtres gaussiens par des filtres boîte.

Les cartes de caractéristiques *(feature maps)* d'intensité $FM_{L_{on}}$, $FM_{L_{off}}$, de couleur FM_R, FM_G, FM_B, FM_Y, d'orientation FM_{Ori0}, FM_{Ori45}, FM_{Ori90} et FM_{Ori135}, et de mouvement $FM_{M_{on}}$, $FM_{M_{off}}$ sont calculées par simple somme de toutes les résolutions de leurs pyramides respectives. Contrairement à (Itti *et al.*, 1998) et de manière similaire à (Frintrop, 2005), la somme n'est pas effectuée sur la résolution la plus basse, mais à une résolution intermédiaire (généralement r_2) afin de garder un maximum d'informations.

On pourrait directement fournir ces cartes de caractéristiques en entrée de notre système de fusion de cartes proies/prédateurs. Cependant, en prenant en compte seulement des caractéristiques simples (intensité, couleur, orientation, mouvement), nous avons déjà généré 12 cartes. La fusion de ces 12 cartes par le système proies/prédateurs est possible, mais alourdirait inutilement notre modèle attentionnel. Pour simplifier les calculs, nous procédons à un regroupement des cartes de caractéristiques en 4 cartes

de singularité *(conspicuity maps)* d'intensité SM_L, de couleur $SM_{\mathbf{C}}$, d'orientation SM_O et de mouvement SM_M (Itti *et al.*, 1998), (Frintrop, 2005).

Comme l'ont montré (Bruce, Jernigan, 2003), les cartes de singularité ne peuvent pas se résumer à une simple somme (ou combinaison linéaire) des cartes de caractéristiques. Une normalisation est nécessaire afin de favoriser les éléments saillants de chaque carte et/ou les cartes les plus saillantes. Différentes solutions sont proposées par (Itti *et al.*, 1998) ou (Itti, Koch, 2001).

Nous proposons un nouvel opérateur de normalisation inspiré de (Mancas, 2007) et de la théorie de l'information. On considère que l'information saillante est par nature rare (une chose courante peut difficilement attirer l'attention), on choisit alors de favoriser les données peu fréquentes. En termes de théorie de l'information, la rareté correspond à une information propre *(self-information)* élevée. On va donc normaliser chaque pixel en fonction de son information propre $SI(x, y)$. Celle-ci est évaluée relativement à la répartition des niveaux de gris de chaque carte.

A l'aide d'un histogramme, on calcule la probabilité $p(x)$ d'un pixel d'appartenir au niveau de gris n donné. L'espace des niveaux de gris est ici séparé en 16 intervalles : $n \in [0, 15]$. On calcule alors SI avec la formule classique :

$$SI_i(x, y) = -\log(p(FM_i^{'}(x, y))) \tag{4}$$

$FM^{'}$ étant la version quantifiée sur 16 valeurs de la carte de singularités à normaliser et $i \in \{L_{on}, L_{off}, R, G, B, Y, Ori0, Ori45, Ori90, Ori135, M_{on}, M_{off}\}$. On obtient la version normalisée $FM_i^{"}$ de chacune des cartes de caractéristiques originales par simple multiplication avec son information propre.

$$FM_i^{"}(x, y) = \frac{FM_i(x, y) \times SI_i(x, y)}{\log(Card(FM_i))} \tag{5}$$

Le facteur de normalisation globale $\log(Card(FM_i)$ permet de garder le rapport $\frac{SI_i(x,y)}{\log(Card(FM_i)}$ dans l'intervalle $[0, 1]$. En effet SI_i atteint sa valeur maximale lorsque $SI_i(x, y) = -\log(p(FM_i^{'}(x, y))) = -\log(\frac{1}{Card(FM_i^{'})}) = log(Card(FM_i^{'}))$.

Cette normalisation permet de favoriser les pics rares, sans avoir recours à un seuil. On peut alors moyenner les différentes cartes de caractéristiques normalisées afin d'obtenir les cartes de singularité.

3.2.3. Pseudo-flou rétinien

La grande majorité des modèles computationnels d'attention effectue les mêmes traitements sur tous les pixels des images qui leur sont fournis. Certains néanmoins, utilisent la transformation *log polaire* afin de simuler la structure interne de la rétine, dont la résolution est variable en fonction de l'éloignement par rapport au centre de projection de l'image. Cette représentation est intéressante car il est possible que cette

perte graduelle d'information ait un impact sur l'allocation de l'attention. Cependant, la transformation *log polaire* n'est pas très adaptée à la structure de notre algorithme d'attention visuelle, le calcul des filtres *centre-périphérie* étant alors beaucoup plus complexe. Une solution efficace et adaptée à notre architecture est l'utilisation non plus de pyramides multirésolutions, mais de colonnes multirésolutions. Nous décrivons le principe de calcul des colonnes multirésolutions en prenant pour exemple les cartes de caractéristiques d'intensité. Le principe est également appliqué aux cartes de couleur et d'orientation. Dans le cas des colonnes multirésolutions, on limite les calculs à une zone de taille constante (par exemple 16×16 pixels), quelle que soit la résolution. Cette zone de taille constante est centrée sur la position du dernier point de focalisation calculée par le système attentionnel. Cette façon de procéder permet de créer un effet de flou lors du calcul des cartes de caractéristiques car lors de l'addition trans-échelles *(across-scale addition)* la quantité d'information disponible à chaque échelle est variable. Ainsi au niveau du point de focalisation, l'ensemble des données des différentes échelles est disponible. Dès que l'on s'éloigne du centre, les données des résolutions les plus fines n'existent plus (elles n'ont pas été calculées) : l'addition est effectuée avec les informations des résolutions les plus basses.

3.3. Le système attentionnel

L'attention visuelle peut être vue comme une compétition entre différentes sources d'information. Ce parti pris est notamment celui des modèles distribués de compétition biaisée. Pour résoudre ce problème de compétition, de nombreuses solutions sont possibles. La plus courante est une approche neuromimétique, basée sur les réseaux de neurones (Spratling, Johnson, 2004 ; Deco, 2004 ; Tsotsos *et al.*, 2005). Cette approche implique cependant une fidélité au modèle biologique dont nous n'avons pas besoin (notre modèle doit avoir un comportement plausible, mais n'est pas destiné à être une réplique du système humain) et qui peut s'avérer être un handicap pour les performances de notre système (de par leur complexité). Une autre approche consiste à considérer le cerveau comme un système dynamique dont le comportement peut être modélisé plus globalement (Eliasmith, 1995 ; Lesser, Dinah, 1998). On remplace alors les réseaux de neurones des modèles distribués par un système d'équations différentielles représentatives du comportement à reproduire (Vitay *et al.*, 2005). Dans ce cadre, nous proposons de modéliser le phénomène de compétition attentionnelle par un système dynamique compétitif, inspiré de la modélisation de la chaîne alimentaire animale : le système proies/prédateurs. Son architecture est représentée en figure 2.

Les prochaines sections auront pour but de justifier cette analogie faisant le lien entre attention et système proies/prédateurs, ainsi que de décrire les équations régissant l'évolution du système.

3.3.1. Construction d'un système proies/prédateurs

Les systèmes proies/prédateurs sont des systèmes d'équations habituellement utilisés pour simuler l'évolution et l'interaction de différentes colonies de proies et de prédateurs ainsi que d'autres phénomènes biologiques (Murray, 2003). Pour notre mo-

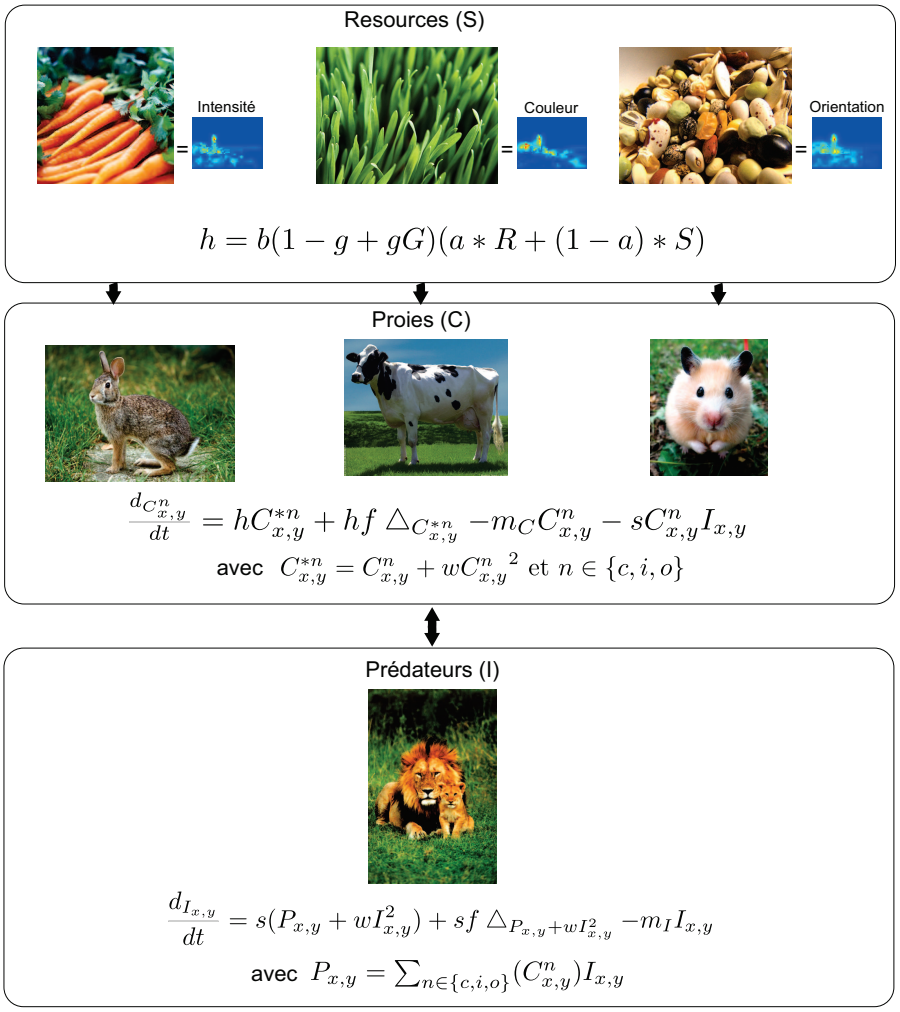

Figure 2. Architecture du système proies/prédateurs de fusion de cartes de singularité

dèle, nous nous sommes inspirés de (Lesser, Dinah, 1998) afin de représenter l'évolution temporelle du focus d'attention.

Classiquement, l'évolution d'un système proies/prédateurs est régie par quelques règles simples, initialement développées dans les années 1920 par Vito Volterra pour modéliser l'évolution de populations de poissons dans différents ports italiens :

1. les proies C ont un taux de croissance proportionnel à leur population et un facteur de croissance b ;

2. les prédateurs I ont un taux de croissance proportionnel au taux de rencontre entre les proies et les prédateurs CI et un facteur de prédation s ;

3. les prédateurs ont un taux de mortalité naturelle proportionnel à leur population et un facteur de mortalité m_I ;

4. les proies ont un taux de mortalité proportionnel au taux de prédation CI et un facteur de mortalité s'.

En appliquant ces quatre règles, on obtient les équations de Volterra-Lotka :

$$\begin{cases} \frac{dC}{dt} & = bC - s^{'}CI \\ \frac{dI}{dt} & = sCI - m_I I \end{cases} \tag{6}$$

Cette version « de base » des équations proies/prédateurs peut être enrichie de différentes façons :

1. on peut diminuer le nombre de paramètres en remplaçant $s^{'}$ par s. En effet, la différence de dynamique entre la mortalité des proies et la croissance des prédateurs peut être modélisée par un ajustement des facteurs b et m_I ;

2. le modèle original ne prend pas en compte la mortalité naturelle des proies, en l'absence de toute prédation. Cela n'est pas très important lorsque le modèle n'évolue que de façon temporelle, la mortalité naturelle est alors négligeable par rapport à la prédation. Cependant, lorsque le modèle est appliqué sur une carte 2D, certaines zones de la carte peuvent ne pas contenir de prédateur. La mortalité naturelle des proies ne peut alors plus être considérée comme négligeable. On peut donc ajouter un terme $-m_c C$ représentant cette mortalité additionnelle ;

3. on peut appliquer le modèle, non plus à une quantité générale de proies et de prédateurs, mais à une des cartes 2D où chaque point représente la quantité de proies ou de prédateurs à un instant et un lieu donnés. Les proies et les prédateurs peuvent alors se déplacer grâce à une règle de diffusion classique proportionnelle à leur laplacien $\triangle_C$ et à un facteur de diffusion f ;

On obtient alors le jeu d'équations suivant, permettant la modélisation de l'évolution des quantités de proies et prédateurs sur un espace à deux dimensions :

$$\begin{cases} \frac{d_{C_{x,y}}}{dt} & = bC_{x,y} + f \triangle_{C_{x,y}} - m_C C_{x,y} - sC_{x,y}I_{x,y} \\ \frac{d_{I_{x,y}}}{dt} & = sC_{x,y}I_{x,y} + sf \triangle_{P_{x,y}} - m_I I_{x,y} \end{cases} \tag{7}$$

Un dernier phénomène peut alors être ajouté : un *feedback* positif, proportionnel à C^2ou I^2 et contrôlé par le facteur w. Celui-ci modélise le fait que (en présence de ressources illimitées) plus une population est nombreuse, mieux elle est à même de croître (chasse plus efficace, taux de rencontre plus élevé favorisant la reproduction, etc.). On obtient alors :

$$\begin{cases} \frac{d_{C_{x,y}}}{dt} & = b(C_{x,y} + w(C_{x,y})^2) + f \triangle_{C_{x,y}} - m_C C_{x,y} - sC_{x,y}I_{x,y} \\ \frac{d_{I_{x,y}}}{dt} & = s(C_{x,y}I_{x,y} + w(I_{x,y})^2) + sf \triangle_{P_{x,y}} - m_I I_{x,y} \end{cases} \tag{8}$$

Pour que ce modèle puisse être utilisé pour simuler l'évolution temporelle du focus d'attention nous avons développé l'analogie suivante :

– il existe quatre populations de proies et une seule population de prédateurs ;

– les quatre populations de proies représentent la répartition spatiale de la curiosité engendrée par les quatre cartes de singularité (intensité, couleur, orientation et mouvement) décrites dans la section précédente ;

– la population de prédateurs représente l'intérêt généré suite à la consommation de la curiosité liée aux différentes cartes ;

– le maximum global d'intérêt (maximum de la carte des prédateurs) représente le focus d'attention à l'instant t.

L'application de ces caractéristiques au système d'équations (8) permet d'obtenir les équations décrites dans le paragraphe suivant.

3.3.2. Simulation de l'évolution du focus d'attention par système proies/prédateurs

Pour chacune des quatre cartes de caractéristiques (couleur, intensité, orientation et mouvement) calculées, l'équation de la matrice des proies C est régie par l'équation (9), directement dérivée de l'équation (8) :

$$\frac{d_{C^n_{x,y}}}{dt} = hC^{*n}_{x,y} + hf \triangle_{C^{*n}_{x,y}} - m_C C^n_{x,y} - sC^n_{x,y} I_{x,y} \quad (9)$$

avec $C^{*n}_{x,y} = C^n_{x,y} + w{C^n_{x,y}}^2$et $n \in \{c, i, o, m\}$, ce qui signifie que cette équation est valable pour les trois matrices C^c, C^i,C^o et C^m représentant respectivement la couleur, l'intensité, l'orientation et le mouvement. C représente la curiosité générée à partir de la singularité intrinsèque de l'image. Elle est créée à partir d'une combinaison h de quatre facteurs :

$$h = b(1 - g + gG)(a * R + (1 - a) * SM_n)(1 - e) \quad (10)$$

– *la singularité SM_n de l'image* (avec $n \in \{c, i, o, m\}$), dont la contribution est inversement proportionnelle au facteur a ;

– *une source R de bruit aléatoire*, simulant le haut niveau de bruit de l'activité de notre cerveau (Fox *et al.*, 2007) et dont a définit l'intensité par rapport à S. Les équations différentielles modélisant l'évolution de notre système proies/prédateurs deviennent alors des équations différentielles stochastiques. On pourra, en faisant varier a, donner un peu de liberté au système attentionnel et lui faire explorer des zones moins saillantes de l'image ou au contraire, contraindre le système à ne visiter que les zones de forte singularité ;

– *une carte gaussienne G* permettant de simuler la préférence centrale observée lors des expérimentations psycho-visuelles (Le Meur, 2005 ; Tatler, 2007). L'importance de cette carte est modulée par le facteur g ;

– *l'entropie e* de la carte de singularité (couleur, intensité ou orientation) normalisée entre 0 et 1. La modulation par $(1 - e)$ permet de favoriser les cartes possédant un nombre limité de maximum locaux. Traduit en termes de proies/prédateurs,

on favorise la croissance des populations de proies les plus organisées (regroupées en un petit nombre de sites). Ce mécanisme est l'équivalent au niveau du système proies/prédateurs de la normalisation des cartes de caractéristiques présentée dans l'équation (5).

L'évolution de la matrice I des prédateurs consommant ces 3 types de proies est régie par l'équation (11) :

$$\frac{d_{I_{x,y}}}{dt} = s(P_{x,y} + wI_{x,y}^2) + sf \triangle_{P_{x,y}+wI_{x,y}^2} - m_I I_{x,y} \tag{11}$$

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

Comme nous l'avons déjà évoqué, le terme quadratique modulé par le facteur w permet de renforcer la dynamique du système et facilite l'émergence d'un comportement chaotique en favorisant la saturation de certaines valeurs des matrices. Enfin, nous rappelons que la curiosité C est consommée par l'intérêt I et que le point de fixation à un instant t est le maximum de la carte d'intérêt. Pour permettre un changement moins fréquent de la position du focus d'attention, nous avons ajouté un mécanisme optionnel d'hystérésis permettant de ne changer le focus d'attention que si le nouveau maximum de la carte des prédateurs dépasse l'ancien de plus d'un certain seuil.

Quelques contraintes ont également été ajoutées lors de l'implémentation du modèle afin de renforcer sa stabilité :

– les proies et les prédateurs ne peuvent pas voir leur population chuter en dessous d'un seuil minimal (ici fixé à 1). Cela permet au système de fonctionner, même si localement ou pendant un court instant l'une des populations vient à baisser fortement. En contrepartie, on peut assister à des phénomènes d'écrêtage, modifiant la dynamique du système ;

– les proies et les prédateurs ne peuvent pas voir leur population augmenter au-delà d'un seuil maximal $Max_{population}$ (ici fixé à 65535). Les équations deviennent alors :

$$\begin{aligned}
\frac{d_{C_{x,y}^n}}{dt} &= \left(1 - \frac{C_{x,y}^n}{Max_{population}}\right)\left(hC_{x,y}^{*n} + hf\triangle_{C_{x,y}^{*n}}\right) - m_C C_{x,y}^n - sC_{x,y}^n I_{x,y} \\
\frac{d_{I_{x,y}}}{dt} &= \left(1 - \frac{I_{x,y}}{Max_{population}}\right)\left(s(P_{x,y} + wI_{x,y}^2) + sf\triangle_{P_{x,y}+wI_{x,y}^2}\right) - m_I I_{x,y}
\end{aligned}$$

3.3.3. Valeur par défaut des paramètres du système

Nous avons dans un premier temps déterminé empiriquement un jeu de paramètres par défaut (tableau 4), permettant un équilibre du système en l'absence de toute image.

Ces paramètres représentent des valeurs permettant d'obtenir un système à l'équilibre. Cet équilibre est obtenu lorsque le système fonctionne sans aucune image d'entrée et reste stable. Nous montrons dans la suite que ces valeurs peuvent varier dans

Tableau 4. Paramètres par défaut du système proies/prédateurs

a	b	g	w	m_C	m_I	s	f
0.5	0.007	0.1	0.001	0.3	0.5	0.025	0.25

une large gamme sans compromettre la stabilité du système. Il est à noter que la mise en œuvre du modèle évolue suivant la méthode d'Euler avec un pas de 0.33 et que 3 sous-itérations sont exécutées avant le calcul de chaque focus d'attention.

3.4. Le bouclage top-down

Le modèle d'attention présenté dans cet article est computationnellement efficace et plausible. Il offre de nombreuses possibilités de réglages (curiosité, préférences centrales, etc.) qui peuvent être exploitées afin d'adapter le comportement du système à un contexte particulier (voir sections suivantes). Cette adaptation est cependant quelque peu limitée. Dans la présente section, nous proposons d'étendre notre modèle bottom-up afin qu'il puisse prendre en compte des informations concernant ses objectifs. Cette influence top-down peut être exprimée comme une simple modification des paramètres du modèle, mais elle peut aussi réutiliser des informations générées par le système lui-même pour modifier son comportement. Dans ce dernier cas, une boucle de rétroaction est créée (auto-adaptation). Dans ce qui suit, nous définissons les mécanismes d'adaptation de notre modèle. Nous examinerons également la façon dont les lieux visités précédemment peuvent être utilisés comme entrées à un mécanisme de rétroaction d'attention visant à contrôler les capacités d'exploration de la scène du modèle.

3.4.1. Mécanismes d'adaptation

Dans le cadre d'une recherche guidée, deux mécanismes sont couramment utilisés pour introduire de l'information *top-down* dans les modèles computationnels hiérarchiques :

– appliquer des poids différents aux cartes de caractéristiques. Cela permet de biaiser le système en faveur de la connaissance *a priori* sur le/les objet(s) recherché(s) dans la scène. C'est ce type de mécanisme qui est utilisé dans (Frintrop *et al.*, 2005) afin d'apprendre au système attentionnel à reconnaître ce qui est important en fonction du contexte ;

– appliquer des poids différents à chaque pixel des cartes de caractéristiques (soit globalement, soit indépendamment pour chacune des cartes). Cette approche reprend et étend la précédente en permettant de spécifier au modèle attentionnel un *a priori* sur la localisation des objets recherchés. Ce principe est proposé par exemple dans (Navalpakkam *et al.*, 2005) *via* l'utilisation d'une *task-relevance maps*. Dans cette approche la carte *top-down* permet de fournir des informations concernant la possibilité de trouver des éléments intéressants en fonction du type de scène observée.

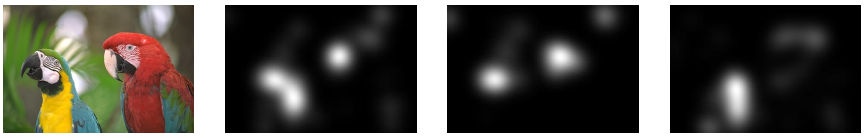

Image originale (a) (b) (c)

Figure 3. Exemples de heatmaps *obtenues après modification du comportement attentionnel par pondération des caractéristiques. a) paramètres par défaut du système ($W^i = 1.0$, $W^c = 1.0$, $W^o = 1.0$). b) diminution de l'importance de la couleur ($W^i = 1.0$, $W^c = 0.5$, $W^o = 1.0$). c) augmentation de l'importance de la couleur ($W^i = 0.5$, $W^c = 1.0$, $W^o = 0.5$)*

D'autres raffinements sont encore possibles, puisque l'on peut également fournir un *a priori* sur l'intensité des caractéristiques attendues (Navalpakkam, Itti, 2006). Cependant, la mise en œuvre d'un tel système devient alors assez difficile.

Bien que non hiérarchique, notre système attentionnel compétitif peut être biaisé à l'aide de cartes *top-dow*n. Il suffit de modifier l'équation de mise à jour des proies afin d'utiliser une carte (différente pour chaque type de proie) favorisant la croissance d'un type de proie (éventuellement en un lieu particulier) plutôt qu'un autre :

$$\frac{d_{C^n_{x,y}}}{dt} = T^n_{x,y}\left(1 - \frac{C^n_{x,y}}{Max_{pop}}\right)\left(hC^{*n}_{x,y} + hf\triangle_{C^{*n}_{x,y}}\right) - m_C C^n_{x,y} - sC^n_{x,y}I_{x,y} \quad (12)$$

avec $T^n_{x,y}$ la carte *top-down* associée au type de proie $n \in \{i, c, o, m\}$ et $\max_{x,y}(T^c_{x,y}) = 1.0$ (la carte *top-down* est normalisée entre 0 et 1).

Si $T^n_{x,y} = W^n \; \forall (x, y)$ alors on contraint l'évolution globale par un poids constant. (figure 3).

Les cartes *top-down* décrites précédemment permettent de modifier le comportement du système attentionnel grâce à un *a priori* fourni par le contexte (donc de manière externe au système attentionnel, voire visuel). On peut également souhaiter biaiser le fonctionnement du système de manière interne afin de le diriger vers une ou des zones à (ne pas) visiter. Ceci peut être effectué en fonction des informations fournies par le système de vision auquel est rattaché le système attentionnel (dans notre cas ce système est théorique) ou directement selon des informations calculées à partir du système d'attention lui même.

On peut alors biaiser la croissance de toutes les proies par une carte de rétroaction R commune. Celle-ci agit comme un mécanisme de facilitation ou d'inhibition, définit

par (Berthoz, 2009), comme un des dispositifs-clés permettant la compétition et la sélection, amenant à la simplexité. On aura alors :

$$\frac{d_{C^n_{x,y}}}{dt} = R_{x,y}T^n_{x,y}\left(1 - \frac{C^n_{x,y}}{Max_{pop}}\right)\left(hC^{*n}_{x,y} + hf\triangle_{C^{*n}_{x,y}}\right) - m_C C^n_{x,y} - sC^n_{x,y}I_{x,y} \quad (13)$$

$R_{x,y}$ étant calculé en fonction d'un ou plusieurs critères de rebouclage, dont nous fournissons un exemple dans la prochaine section.

3.4.2. Un critère de bouclage : l'exploration de la scène

Notre système attentionnel n'étant pour l'instant relié à aucun système de vision, son comportement ne peut pas être adapté en fonction de critères externes (par exemple : l'estimation de la position ou des attributs des objets à reconnaître, fournie par la mémoire du système de vision hôte). Pour tester nos mécanismes d'adaptation, il nous faut alors définir un ou plusieurs critères calculés à partir des données disponibles dans le système attentionnel.

Nous avons choisi un critère d'exploration de l'espace : à partir des données de focalisation fournies par le système attentionnel, nous calculons une carte représentant les parties déjà visitées par le système. En utilisant cette carte comme carte de rétroaction, et en modulant son influence (négativement ou positivement) on peut définir deux stratégies attentionnelles opposées (ainsi que tous leurs intermédiaires) :

– maximisation de l'exploration de l'espace : le système attentionnel va privilégier les zones qu'il n'a pas encore visitées ;

– stabilité des focalisations : le système attentionnel va privilégier les zones qu'il a déjà visitées (focalisation) ;

Le calcul de la carte des zones visitées est basé sur le principe suivant : lors d'une focalisation, on considère qu'un maximum d'information est acquis au centre de la zone de focalisation ; l'information est ensuite moins précise au fur et à mesure que l'on s'éloigne du centre (du fait par exemple de la résolution variable de la rétine). On peut étendre cette notion en prenant en compte le fait que notre système (tout comme l'homme) doit avoir une mémoire limitée : il oublie graduellement les informations acquises. La figure 4 montre l'influence du facteur d'oubli sur la carte des zones visitées.

A partir de la carte des zones visitées, nous pouvons construire la carte de rétroaction R. Nous devons alors pouvoir moduler la rétroaction en fonction de l'intensité voulue, et du type d'influence (positive ou négative). Ceci est contrôlé par un seul paramètre $F_{feedback}$. Une valeur positive de $F_{feedback}$ engendrera un comportement ayant tendance à explorer les zones déjà visitées (focalisation/suivi) ; une valeur négative de $F_{feedback}$ engendrera un comportement privilégiant au contraire les zones non visitées (exploration). Nous allons maintenant aborder l'évaluation de notre modèle.

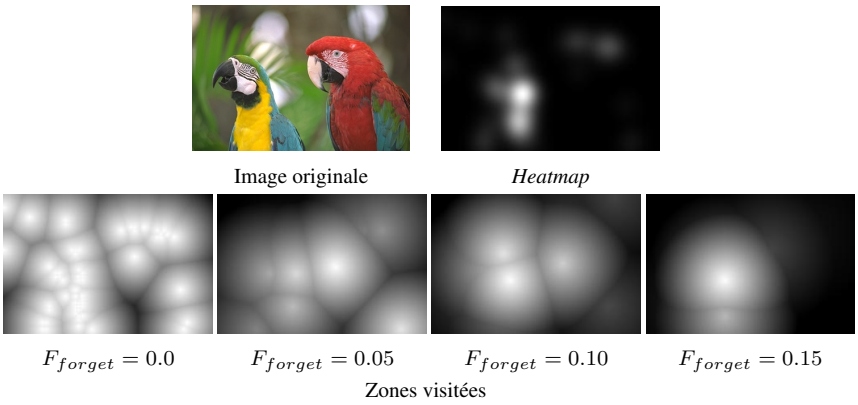

Figure 4. Influence du facteur d'oubli sur la carte des zones visitées après 100 itérations

4. Évaluation du modèle

Les performances de notre modèle, comparativement aux principales solutions proposées dans la littérature, ont été présentées dans (Perreira Da Silva *et al.*, 2011). Dans cet article, nous nous concentrons plutôt sur l'étude de l'influence des différents paramètres du système sur son comportement. Afin de réaliser cette étude, il est nécessaire de définir un ou plusieurs niveaux d'observation (microscopique ou macroscopique) vis-à-vis d'un ensemble de propriétés. Ces propriétés étudiées sont tirées de l'ensemble PAIRED : stabilité, reproductibilité, exploration de la scène, dynamique du système. Pour l'ensemble de ces propriétés, nous avons également étudié l'influence de la rétroaction *top-down*.

Toutes les mesures présentées dans cette section ont été effectuées sur deux bases d'images. La première est proposée par (Bruce, Tsotsos, 2009). Elle est composée de 120 images couleur présentant des rues, des jardins, des véhicules ou des bâtiments. La seconde, proposée par (Le Meur *et al.*, 2006), contient 26 images couleur. Elles représentent des scènes de sport, animaux, bâtiments, scènes d'intérieur ou paysages. Sauf indication contraire, le système fonctionne en utilisant les paramètres définis dans le tableau 4.

4.1. Stabilité

Les équations de Volterra-Lotka ne produisent un modèle stable [3] que dans une plage de paramètres déterminée (Idema, 2005). C'est également le cas pour notre système. Par exemple si le taux de natalité a des proies est trop faible par rapport au taux

3. Nous définissons ici la stabilité comme un régime oscillant (approximativement) stationnaire pour toutes les cartes du système (proies et prédateurs).

de prédation s, et que le taux de mortalité naturelle des prédateurs est élevé, ni les proies ni les prédateurs n'auront l'occasion de voir leurs populations croître.

Nous avons étudié la stabilité de notre système en considérant la valeur moyenne des cartes de proies et prédateurs. Si ces valeurs restent dans une fourchette limitée, le système est considéré stable. Le tableau 5 donne un aperçu du comportement du système pour différentes valeurs de la natalité et de mortalité des proies et de prédateurs. En dehors de la plage de stabilité les populations saturent progressivement.

Tableau 5. Plage de stabilité des principaux paramètres

Paramètres	Valeur par défaut	valeur minimale de stabilité	valeur maximale de stabilité
Natalité des proies b	0.007	0.006	0.013
Mortalité des proies m_c	0.3	0.3	0.36
Prédation s	0.025	0.017	0.05
Natalité des prédateurs m_I	0.5	0.1	1.5
Feedback w	0.001	0	0.003

4.2. Reproductibilité

De par sa discrétisation et l'introduction d'une carte aléatoire lors du calcul du facteur de croissance globale h, notre système attentionnel est non déterministe. Ce mode de fonctionnement est intéressant car il permet de simuler la variabilité naturelle observée lorsque l'on mesure plusieurs fois les focalisations attentionnelles d'une même personne sur la même image. C'est également un moyen d'inciter notre système attentionnel à explorer des zones de l'image moins saillantes. On peut alors ajuster la « curiosité » du système. Toutefois, en donnant plus de « curiosité », notre système conduit aussi à moins de reproductibilité. Afin d'étudier le phénomène, nous avons comparé les *heatmaps* générées par les mesures de suivi du regard, avec les *heatmaps* générées par les différentes simulations fournies par notre modèle. Nous avons utilisé des mesures de similarité/dissimilarité classiques : la corrélation croisée (*cross-correlation*) (Le Meur *et al.*, 2006), la divergence de Kullback-Leibler (Tatler *et al.*, 2005) et la *normalized scanpath saliency* (Peters *et al.*, 2005). Pour ces expérimentations, nous avons utilisé les *heatmaps* fournies par Bruce et LeMeur. Comme le nombre de paramètres étudiés est important (la diffusion, l'hystérésis, le bruit, la rétroaction positive et la rétroaction *top-down*...), nous avons décidé de ne pas inclure les résultats détaillés de nos mesures dans cet article, néanmoins, les résultats synthétiques sont présentés dans le tableau 8.

4.3. Exploration de l'espace

Comme nous l'avons vu, le rôle du système d'attention visuelle humain est entre autres d'optimiser l'exploration de la scène afin d'acquérir l'information le plus efficacement possible. Un modèle computationnel d'attention doit également remplir ce

rôle. Mais comment vérifier que l'information est acquise efficacement en l'absence de tâche de vision de plus haut niveau à évaluer ? Même en présence d'une tâche à effectuer, la mesure de performance serait biaisée par le type de tâche et son implémentation. Pour estimer la capacité de notre système à explorer efficacement la scène en l'absence de système de vision, nous proposons une nouvelle méthodologie. A chaque pas de temps de la simulation, lorsque le focus d'attention change, nous « reconstruisons » incrémentalement l'image initiale à partir des informations disponibles à travers une rétine simulée par un flou variable, centré sur la zone de focalisation. L'image reconstruite devient ainsi plus précise au fur et à mesure des différentes focalisations. Pour cela, nous mettons à jour à chaque pas de temps, un masque de flou M_{flou} dont la valeur maximale représente les zones non floutées et la valeur minimale les zones de flou maximum. L'image « reconstruite » I_R est ensuite générée à partir de l'image source I_S par convolution avec un filtre boîte (moyenneur) de taille inversement proportionnelle à M_{flou} :

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

avec B_s un filtre boîte (moyenneur) carré de taille s, et $s(x,y) = 2^{N-M_{flou}(x,y)}$ la fonction calculant la taille du filtre en fonction du masque de flou.

Nous mesurons alors le rapport entre la quantité d'information contenue dans l'image initiale, et celle contenue dans l'image reconstituée. Pour cela nous appliquons les principes de la théorie de l'information. La quantité d'information contenue dans une image peut être évaluée par le principe du *minimum description length (MDL)* qui est une version calculable de la complexité de Kolmogorov (Rissanen, 1978). D'après ce principe, plus une donnée est simple, plus elle sera facile à compresser (du fait de la redondance de certaines de ses données). A l'opposé, si une donnée est complexe, elle sera difficile à compresser efficacement. Nous appliquons ce principe aux images dont nous souhaitons évaluer la quantité d'information en les compressant *via* deux algorithmes de compression : JPEG (compression avec pertes) et PNG (compression sans perte).

En effectuant le rapport entre la taille de l'image initiale compressée et la taille de l'image reconstituée, on obtient un estimateur (compris entre 0 et 1) de la performance d'exploration de l'espace de notre algorithme à un instant t.

$$InformationRatio_{JPEG} = \frac{size\left(compress_{JPEG}(I_S)\right)}{size\left(compress_{JPEG}(I_R)\right)} \quad (14)$$

$$InformationRatio_{PNG} = \frac{size\left(compress_{PNG}(I_S)\right)}{size\left(compress_{PNG}(I_R)\right)} \quad (15)$$

avec I_S l'image source et I_R l'image « reconstruite ».

Le mécanisme de rétroaction vise à contrôler la façon dont la scène visuelle est explorée. Les résultats obtenus par les mesures décrites ci-dessus (JPEG et PNG ratio) confirment ce comportement attendu (tableau 6) :

– une rétroaction négative conduit à une exploration plus rapide, mais pas nécessairement plus exhaustive (la scène est déjà presque totalement couverte après 300 étapes de la simulation) ;

– une rétroaction positive peut réduire considérablement la zone explorée. Pour une valeur de $F_{feedback} = -1$, même après 300 étapes de simulation, le taux de couverture de la scène est encore inférieur à celui obtenu sans aucun *feedback* après 50 iterations.

Tableau 6. Influence du feedback *dans l'exploration de la scène*

Feedback	JPG			PNG		
itérations	50	150	300	50	150	300
-1.0	0,779	0,915	0,955	0,940	0,983	0,992
-0.6	0,772	0,913	0,955	0,930	0,978	0,987
-0.2	0,736	0,894	0,939	0,914	0,977	0,988
0	0,714	0,860	0,911	0,899	0,963	0,978
0.2	0,609	0,751	0,817	0,810	0,900	0,935
0.6	0,522	0,648	0,721	0,739	0,830	0,871
1.0	0,481	0,594	0,645	0,703	0,792	0,826

4.4. La dynamique

Même si notre système ne génère pas de saccades ni de fixations directement comparables à celles de l'œil humain (nous ne tenons pas compte, par exemple, des contraintes sur les mouvements des yeux), on peut estimer la durée moyenne de fixation entre deux changements du focus d'attention. Ce changement est détecté si la distance entre la position actuelle et la prochaine dépasse un seuil S_{fixing}. La valeur de S_{fixing} (15 % du côté le plus long de l'image source) est déterminée de manière à être compatibles avec le paramètre $fovea_{size}$ utilisé pour générer les *heatmaps*.

Tableau 7. Influence de certains paramètres sur le temps de fixation

Paramètres	Par défaut	CentralBias =0.5	Feedback= -1.0	Feedback =1.0	Step=0.1 Iterations=1	Hysteresis =0.5
Temps de fixation	70 ms	143 ms	53 ms	416 ms	806 ms	102 ms

Nous avons mesuré l'effet des différents paramètres de notre modèle sur le temps moyen de fixation sur les bases de Bruce et Le Meur. Les résultats de cette étude sont résumés dans le tableau 8. Le tableau 7 donne quelques exemples de certains paramètres représentatifs. Ces résultats devraient être comparés au temps moyen de fixation de l'homme d'environ 300 ms (Dorr *et al.*, 2010).

La dynamique peut être affinée à l'aide de nombreux paramètres. Mais les plus efficaces sont les paramètres différentiels de l'équation d'évolution (pas de simulation et nombre de sous-itérations), le *feedback*, et la préférence centrale. Ceux-ci n'ont néanmoins pas tous les mêmes effets secondaires sur d'autres propriétés (plausibilité, l'exploration scène, etc.)

4.5. Bilan

Le tableau 8 résume l'influence des différents paramètres sur le fonctionnement du système. Nous ne rappelons pas ici l'influence des facteurs de natalité et mortalité b, s, M_C et M_I, uniquement utilisés pour ajuster la stabilité du système. Les flèches utilisées ont la signification suivante :

$\uparrow$ forte influence positive. $\downarrow$ forte influence négative. $\nearrow$ faible influence positive.

$\searrow$ faible influence négative. $\rightarrow$ pas d'influence significative. $\times$ non testé car théoriquement non influent.

Tableau 8. Résumé de l'influence des différents paramètres

Paramètres	Valeur par défaut	Fidélité	Reproductibilité	Exploration	Dynamique
Filtre rétinien	non	$\nearrow$	$\searrow$	$\nearrow$	$\rightarrow$
Biais central (g)	0.1	$\uparrow$	$\rightarrow$	$\uparrow$	$\downarrow$
Diffusion (f)	0.25	$\rightarrow$	$\searrow$	$\nearrow$	$\searrow$
Hystéresis ($Seuil_{Hysteresis}$)	0	$\rightarrow$	$\rightarrow/\searrow$	$\searrow$	$\searrow$
Valeur initiale	16	$\times$	$\times$	$\times$	$\rightarrow$
Bruit (a)	0.5	$\uparrow/\downarrow$	$\downarrow$	$\uparrow$	$\nearrow$
Feedback positif (w)	0.001	$\nearrow/\downarrow$	$\searrow/\downarrow$	$\nearrow/\searrow$	$\nearrow/\downarrow$
Pas de simulation	$1/3$	$\times$	$\times$	$\times$	$\uparrow$
Nombre de sous-itérations	3	$\times$	$\times$	$\times$	$\uparrow$

Dans le cas du filtre rétinien, les flèches correspondent à l'influence de son activation. Les flèches séparées par une barre oblique (par exemple : $\rightarrow/\searrow$) représentent un premier type d'influence pour de faibles augmentations du paramètre, puis un second pour des augmentations plus fortes.

5. Conclusion

Dans cet article, nous avons présenté une mise en œuvre complète et une évaluation d'un modèle computationnel de l'attention pour la vision par ordinateur. En ce qui concerne la mise en œuvre, nous avons montré que le modèle proies/prédateurs offre de bonnes propriétés pour la simulation dynamique de la compétition entre les

différents types d'information fournie en entrée. Nous avons décrit l'architecture de notre modèle qui peut être divisé en deux parties. La première est hiérarchique, elle améliore le modèle de L. Itti en proposant un traitement plus rapide tout en permettant le calcul sur un plus grand nombre d'échelles au cours de son analyse multirésolution de la scène. La seconde partie est notre contribution majeure : elle présente l'usage d'un système dynamique (inspiré d'un système proies/prédateurs) pour gérer la fusion de cartes de singularités générées dans la première partie du modèle. Ce système dynamique est également utilisé pour générer un focus d'attention à chaque pas de temps de la simulation. Concernant l'évaluation, nous avons présenté des résultats différents (corrélation croisée, la divergence de Kullback-Leibler, *normalized scanpath saliency*) qui démontrent que nos résultats sont rapides, hautement configurables et particulièrement pertinents. Les prochaines étapes de ce travaille consistent à étendre le modèle à la vidéo (prise en compte du mouvement) et intégrer celui-ci dans un système de vision plus global.

Bibliographie

Ahmad S. (1992). *Visit: An efficient computational model of human visual attention*. Phd, University of Illinois, Champaign, IL. http://ftp.icsi.berkeley.edu/ftp/pub/techreports/1991/tr-91-049.pdf

Allport D. A. (1987). Selection for action: Some behavioral and neurophysiological considerations of attention and action. In H. Heuer, S. A.F. (Eds.), *Perspectives on perception and action*, p. 395–419. Hillsdale, NJ, Lawrence Erlbaum Associates.

Avraham T., Lindenbaum M. (2010). Esaliency (extended saliency): meaningful attention using stochastic image modeling. *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, n° 4, p. 693–708. http://www.ncbi.nlm.nih.gov/pubmed/20224124

Aziz M., Mertsching B. (2009). Towards Standardization of Evaluation Metrics and Methods for Visual Attention Models. In *Attention in cognitive systems*, p. 227–241. Springer. http://www.springerlink.com/index/v713433834617727.pdf

Baldi P., Itti L. (2005). Attention: Bits versus Wows. In *2005 international conference on neural networks and brain*, p. 56–61. Ieee. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1614548

Belardinelli A., Pirri F., Carbone A. (2009). Motion Saliency Maps from Spatiotemporal Filtering. In *Lecture notes in artificial intelligence*, p. 112–123. Springer. http://www.springerlink.com/index/425j618q84762l43.pdf

Berthoz A. (2009). *La simplexité*. Paris, Odile Jacob.

Bruce B., Jernigan E. (2003). Evolutionary design of context-free attentional operators. In *Proc. icip'03*, p. 0–3. Citeseer. http://www.cse.yorku.ca/~neil/ICIPnbruce.pdf

Bruce N. D. B., Tsotsos J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, vol. 9, n° 3, p. 5. http://www.journalofvision.org/content/9/3/5.full.pdf

Deco G. (2004). A Neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, vol. 44, n° 6, p. 621–642. http://linkinghub.elsevier.com/retrieve/pii/S0042698903006928

Desimone R., Duncan J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, vol. 18, p. 193–222. http://www.ncbi.nlm.nih.gov/pubmed/7605061

Dorr M., Gegenfurtner K. R., Barth E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, vol. 10, p. 1–17. http://www.journalofvision.org/content/10/10/28.full.pdf

Eliasmith C. (1995). *Mind as a dynamical system*. Thèse de master, University of Waterloo. http://www.arts.uwaterloo.ca/~celiasmi/Papers/eliasmith.1995.dynamic%20mind.masters.pdf

Fox M. D., Snyder A. Z., Vincent J. L., Raichle M. E. (2007). Intrinsic Fluctuations within Cortical Systems Account for Intertrial Variability in Human Behavior. *Neuron*, vol. 56, n° 1, p. 171–184. http://linkinghub.elsevier.com/retrieve/pii/S0896627307006666

Frintrop S. (2005). *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*. Phd, University of Bonn. http://www.iai.uni-bonn.de/~frintrop/paper/frintrop_phd06.pdf

Frintrop S., Backer G., Rome E. (2005). Selecting what is important: Training visual attention. In *28th annual german conference on ai (ki)*, p. 351–366. Koblenz, Germany, Springer Verlag. http://www.iai.uni-bonn.de/~frintrop/paper/frintrop_etal_ki05.pdf

Frintrop S., Klodt M., Rome E. (2007). A real-time visual attention system using integral images. In *5th international conference on computer vision systems (icvs).* Bielefeld, Germany, Applied Computer Science Group. http://biecoll.ub.uni-bielefeld.de/volltexte/2007/36/pdf/ICVS2007-66.pdf

Gilles S. (1996). *Description and experimentation of image matching using mutual information*. Rapport technique. Oxford University, Robotics Research Group, Department of Engineering Science. http://www.robots.ox.ac.uk/~cvrg/trinity2002/seb/mutual_info.ps.gz

Hamker F. (2005). The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. *Computer Vision and Image Understanding*, vol. 100, n° 1-2, p. 64–106. http://linkinghub.elsevier.com/retrieve/pii/S1077314205000767

Heijden A. H. C. van der, Bem S. (1997). Successive approximations to an adequate model of attention. *Consciousness and cognition*, vol. 6, n° 2-3, p. 413–28. http://www.ncbi.nlm.nih.gov/pubmed/9262419

Idema T. (2005). *The behaviour and attractiveness of the Lotka-Volterra equations*. Phd, Universiteit Leiden. http://www.ilorentz.org/\~{}idema/publications/maththesis.pdf

Itti L., Koch C. (2001). Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, vol. 10, p. 161–169. http://papers.klab.caltech.edu/84/

Itti L., Koch C., Niebur E., Others. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, n° 11, p. 1254–1259. http://ilab.usc.edu/publications/doc/Itti_etal98pami.pdf

Kadir T., Brady M. (2001). Saliency, scale and image description. *International Journal of Computer Vision*, vol. 45, n° 2, p. 83–105. http://www.springerlink.com/index/T45N2G8543574026.pdf

Koch C., Ullman S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiology*, vol. 4, n° 4, p. 219–27. http://papers.klab.caltech.edu/104/1/200.pdf

Le Meur O. (2005). *Attention sélective en visualisation d'images fixes et animées affichées sur écran : modèles et évaluation de performances - applications*. Thèse de doctorat, Ecole polytechnique de l'Université de Nantes. http://www.irisa.fr/temics/staff/lemeur/publi/LeMeur_These.pdf

Le Meur O., Le Callet P. (2009). What we see is most likely to be what matters: Visual attention and applications. In *International conference on image processing.* Cairo, Egypt. http://www.irisa.fr/temics/staff/lemeur/publi/LeMeur_ICIP09.pdf

Le Meur O., Le Callet P., Barba D., Thoreau D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, n° 5, p. 802–817. http://www.irccyn.ec-nantes.fr/\~{}lecallet/paper/LeMeur-IEEEPAMI06.pdf

Lesser M., Dinah M. (1998). Mind as a dynamical system : Implications for autism. In *In psychobiology of autism : current research & practice.*

Lienhart R., Maydt J. (2002). An extended set of haar-like features for rapid object detection. In *Ieee icip*, vol. 1, p. 900–903. Citeseer. http://mmc36.informatik.uni-augsburg.de/mediawiki/images/c/c3/Icip2002.pdf

Lopez M., Fernandezcaballero A., Fernandez M., Mira J., Delgado A. (2006). Motion features to enhance scene segmentation in active visual attention. *Pattern Recognition Letters*, vol. 27, n° 5, p. 469–478. http://linkinghub.elsevier.com/retrieve/pii/S0167865505002631

Mancas M. (2007). *Computational Attention : Towards attentive computers*. Phd, Faculté Polytechnique de Mons. http://theses.eurasip.org/media/theses/documents/mancas-matei-computational-attention-towards-attentive-computers.pdf

Mozer M. C., Sitton M. (1998). Computational modeling of spatial attention. *Attention*, p. 341–393. http://www.nbu.bg/cogs/events/2002/materials/Mozer/mozer1998.pdf

Murray J. (2003). *Mathematical biology: An introduction*. Berlin, Heidelberg, Springer Verlag.

Navalpakkam V., Arbib M., Itti L. (2005). Attention and scene understanding. In L. Itti, G. Rees, J. Tsotsos (Eds.), *Neurobiology of attention*, p. 197–203. ACADEMIC PRESS. http://ilab.usc.edu/publications/doc/Navalpakkam_etal05noa.pdf

Navalpakkam V., Itti L. (2006). Top-down attention selection is fine grained. *Journal of Vision*, vol. 6, n° 11, p. 4. http://www.journalofvision.org/content/6/11/4.full.pdf

Orabona F., Metta G., Sandini G. (2008). A Proto-object based visual attention model. In L. Paletta (Ed.), *Attention in cognitive systems. theories and systems from an interdisciplinary viewpoint (wapcv)*, p. 198–215. Berlin, Heidelberg, Springer. http://www.springerlink.com/index/71U3T3262424M763.pdf

Park S., An K., Lee M. (2002). Saliency map model with adaptive masking based on independent component analysis. *Neurocomputing*, vol. 49, n° 1, p. 417–422. http://www.ingentaconnect.com/content/els/09252312/2002/00000049/00000001/art00637

Perreira Da Silva M., Courboulay V., Estraillier P. (2011). Objective validation of a dynamical and plausible computational model of visual attention. In *3rd european workshop on visual information processing (euvip)*. http://hal.archives-ouvertes.fr/docs/00/61/77/30/PDF/euvip_perreira.pdf

Peters R., Iyer A., Itti L., Koch C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, vol. 45, p. 2397–2416. http://linkinghub.elsevier.com/retrieve/pii/S0042698905001975

Rensink R. A. (2000). The dynamic representation of scenes. *Visual Cognition*, vol. 7, p. 17–42. http://homepages.rpi.edu/~grayw/courses/cogs6962/papers/REN00_VisCog.pdf

Rissanen J. (1978). Modeling by shortest data description. *Automatica*, vol. 14, p. 465–471.

Spratling M. W., Johnson M. H. (2004). A feedback model of visual attention. *Journal of cognitive neuroscience*, vol. 16, n° 2, p. 219–37. http://www.ncbi.nlm.nih.gov/pubmed/15068593

Sun Y., Fisher R., Wang F., Gomes H. (2008). A computer vision model for visual-object-based attention and eye movements. *Computer Vision and Image Understanding*, vol. 112, n° 2, p. 126–142. http://linkinghub.elsevier.com/retrieve/pii/S1077314208000167

Tatler B. W. (2007). 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, p. 1–17. http://www.journalofvision.org/content/7/14/4.full.pdf

Tatler B. W., Baddeley R. J., Gilchrist I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision research*, vol. 45, n° 5, p. 643–59. http://www.ncbi.nlm.nih.gov/pubmed/15621181

Torralba A., Oliva A., Castelhano M. S., Henderson J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, vol. 113, n° 4, p. 766–86. http://www.ncbi.nlm.nih.gov/pubmed/17014302

Treisman A. (1969). Strategies and models of selective attention. *Psychological Review*, vol. 76, p. 282–299.

Treisman A., Gelade G. (1980). A Feature-Integration Theory of Attention. *Cognitive Psychology*, vol. 136, n° 12, p. 97–136. http://www.yorku.ca/mfallah/bandb/treisman_gelade.pdf

Tsotsos J., Liu Y., Martineztrujillo J., Pomplun M., Simine E., Zhou K. (2005). Attending to visual motion. *Computer Vision and Image Understanding*, vol. 100, n° 1-2, p. 3–40. http://linkinghub.elsevier.com/retrieve/pii/S1077314205000779

Tsotsos J. K. (1990). Analysing vision at the complexity level. *Behavioral. and. Brain. Sciences*, vol. 13, p. 423–469. http://www.cse.yorku.ca/~tsotsos/Homepage%20of%20John%20K_files/bbs-90.pdf

Tsotsos J. K. (2007). *A selective History of Visual Attention.* ECCV 2008 Tutorial. http://www.cse.yorku.ca/~albertlr/attention_tutorial_eccv2008.htm

Van Rullen R., Koch C. (2005). Visual Attention and Visual Awareness. In G. Celesia (Ed.), *Disorders of visual processing, vol 5*, vol. 91125, p. 65–83. Elsevier. http://papers.klab.caltech.edu/277/1/442.pdf

Viola P., Jones M. (2002). Robust real-time object detection. *International Journal of Computer Vision*, vol. 57, n° 2, p. 137–154. http://research.microsoft.com/en-us/um/people/viola/Pubs/Detect/violaJones_IJCV.pdf

Vitay J., Rougier N., Alexandre F. (2005). A distributed model of spatial visual attention. In *Biomimetic neural learning for intelligent robots*, p. 54–72. Springer. http://www.springerlink.com/index/2qwwddx022jy6naq.pdf

Walther D., Koch C. (2006). Modeling attention to salient proto-objects. *Neural networks : the official journal of the International Neural Network Society*, vol. 19, n° 9, p. 1395–407. http://www.ncbi.nlm.nih.gov/pubmed/17098563

Matthieu Perreira Da Silva est maître de conférences à Polytech Nantes depuis septembre 2011. Il effectue ses activités de recherche à l'Institut de Recherche en Communication et Cybernétique de Nantes. Il a obtenu son DEA en image et calcul (2001) ainsi qu'un doctorat en informatique et ses applications (2010) à l'Université de La Rochelle. De 2001 à 2006 il a travaillé en tant qu'ingénieur R&D dans une société spécialisée dans l'identification biométrique. Ses intérêts en recherche incluent la perception humaine, l'attention visuelle, les interactions homme-machine, la curiosité artificielle, l'apprentissage autonome, le traitement d'image et la vision par ordinateur.

Vincent Courboulay, après un diplôme d'ingénieur délivré par Polytech'Orléans (ESPEO) et une thèse de l'Université de La Rochelle, a obtenu la double qualification en 27e et 61e section en 2003. Depuis septembre 2004, il est maître de conférences à l'Université de La Rochelle au département informatique et son laboratoire de rattachement est le L3I. Ses activités scientifiques s'organisent autour de l'extraction d'informations dans les images et les séquences d'images et cela sur trois niveaux distincts, présentant chacun une définition particulière de l'information. Il s'intéresse aussi à l'attention et sa modélisation biologiquement inspirée.