Modèle computationnel d'attention pour la vision adaptative

Matthieu Perreira da Silva

Université de la Rochelle — Ecole doctorale S2i

Thèse

présentée pour obtenir le titre de DOCTEUR en
Informatique et applications

Modèle computationnel d'attention pour la vision adaptative

Matthieu Perreira Da Silva
mperreir@univ-lr.fr

Soutenue publiquement le 10/12/2010 devant un jury composé de :

Rapporteurs	Laurent Itti	University of Southern California
	Hervé Glotin	Université du Sud Toulon-Var
Examinateurs	Anne Guérin	Université de Grenoble
	Olivier Le Meur	Université de Rennes 1
Directeur de thèse	Pascal Estraillier	Université de La Rochelle
Co-encadrant de thèse	Vincent Courboulay	Université de La Rochelle

Laboratoire Informatique, Image, Interaction

Thèse réalisée au Laboratoire Informatique, Image, Interaction
Pôle Sciences & Technologies, Université de La Rochelle
Avenue M. Crépeau
17042 La Rochelle cedex 01

Tél : +33 5 46 45 82 62
Fax : +33 5 46 45 82 42

Web : http ://l3i.univ-larochelle.fr

Sous la direction de Pascal Estraillier pascal.estraillier@univ-lr.fr

Co-encadrement Vincent Courboulay vincent.courboulay@univ-lr.fr
Armelle Prigent armelle.prigent@univ-lr.fr

Financement Allocation de recherche de la Région Poitou-Charentes

Résumé

L'analyse temps réel de la masse de données générée par les mécanismes de gestion de la vision dans les applications interactives est un problème toujours ouvert, promettant des avancées importantes dans des domaines aussi variés que la robotique, l'apprentissage à distance ou les nouvelles formes d'interactions avec l'utilisateur, sans clavier ni souris.

Dans le cadre général de la vision, les algorithmes d'analyse de scène doivent trouver un compromis entre d'une part la qualité des résultats recherchés et d'autre part la quantité de ressources allouable aux différents tâches. Classiquement, ce choix est effectué à la conception du système (sous la forme de paramètres et d'algorithmes prédéfinis), mais cette solution limite le champ d'application de celui-ci. Une solution plus flexible consiste à utiliser un système de vision adaptatif qui pourra modifier sa stratégie d'analyse en fonction des informations disponibles concernant son contexte d'exécution. En conséquence, ce système doit posséder un mécanisme permettant de guider rapidement et efficacement l'exploration de la scène afin d'obtenir ces informations.

Chez l'homme, les mécanismes de l'évolution ont mis en place le système d'attention visuelle. Ce système sélectionne les informations importantes afin de réduire la charge cognitive et les ambiguïtés d'interprétation de la scène.

Nous proposons, dans cette thèse, un système d'attention visuelle, dont nous définissons l'architecture et les principes de fonctionnement. Ce dernier devra permettre l'interaction avec un système de vision afin qu'il adapte ses traitements en fonction de l'intérêt de chacun des éléments de la scène, *i.e.* ce que nous appelons saillance.

A la croisée des chemins entre les modèles centralisés et hiérarchiques (ex : [Koch 85], puis [Itti 98]), et les modèles distribués et compétitifs (ex : [Desimone 95], puis [Deco 04, Rolls 06]), nous proposons un modèle hiérarchique, compétitif et non centralisé. Cette approche originale permet de générer un point de focalisation attentionnel à chaque pas de temps sans utiliser de carte de saillance ni de mécanisme explicite d'inhibition de retour. Ce nouveau modèle computationnel d'attention visuelle temps réel est basé sur un système d'équations proies / prédateurs, qui est bien adapté pour l'arbitrage entre

un comportement attentionnel non déterministe et des propriétés de stabilité, reproductibilité, et réactivité.

L'analyse des expérimentations menées est positive : malgré le comportement non-déterministe des équations proies / prédateurs, ce système possède des propriétés intéressantes de stabilité, reproductibilité, et réactivité, tout en permettant une exploration rapide et efficace de la scène. Ces propriétés ouvrent la possibilité d'aborder différents types d'applications allant de l'évaluation de la complexité d'images et de vidéos à la détection et au suivi d'objets. Enfin, bien qu'il soit destiné à la vision par ordinateur, nous comparons notre modèle au système attentionnel humain et montrons que celui-ci présente un comportement aussi plausible (voire plus en fonction du comportement défini) que les modèles classiques existants.

Mots clés : Attention visuelle, vision par ordinateur, adaptation, systèmes dynamiques.

Computational attention model for adaptive vision

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 : robotics, distance education, or new mouse-less and keyboard-less human computer interaction.

When using 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 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 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.

In this thesis, we propose a visual attention system tailored for interacting with a vision system (whose theoretical architecture is given) so that it adapts its processing according to the interest (or salience) of each element of the scene.

Somewhere in between hierarchical salience based (ex: [Koch 85], then [Itti 98]) and competitive distributed (ex: [Desimone 95], then [Deco 04, Rolls 06]) models, 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.

Our experiments shows that despite the non deterministic behavior of preys / preda-

tors equations, the system exhibits interesting properties of stability, reproducibility and reactiveness while allowing a fast and efficient exploration of the scene. These properties are useful for addressing different kinds of applications, ranging from image complexity evaluation, to object detection and tracking. Finally, while it is designed for computer vision, we compare our model to human visual attention. We show that it is equally as plausible as existing models (or better, depending on its configuration).

Keywords: Visual attention, computer vision, adaptation, dynamical systems.

Remerciements

Je tiens tout d'abord à remercier Laurent Itti et Hervé Glotin d'avoir accepté d'être les rapporteurs de cette thèse. Leurs commentaires m'ont été très utiles pour la préparation de la soutenance et la mise à jour du rapport de thèse pour sa version finale. Ils seront également, à n'en pas douter, une source d'inspiration importante dans la poursuite de mes travaux.

Je remercie également Anne Guérin-Dugué et Olivier Le Meur pour leur participation à mon jury en tant qu'examinateurs, avec une mention spéciale pour Oliver qui a généreusement mis à ma disposition sa base de test, ainsi qu'une partie de ses résultats.

Je tiens à exprimer tout ma gratitude à Pascal Estraillier pour avoir accepté de diriger ma thèse. Il a su laisser à Vincent et moi même une grande liberté dans nos choix scientifiques, tout en nous délivrant des conseils stratégiques éclairés au moment opportun.

Un très grand merci à Vincent Courboulay pour son encadrement scientifique, sa grande ouverture d'esprit et son amitié. Ces quatre années (puisque nous avons commencé l'aventure dans un autre projet un an avant le début de la thèse) ont été bien plus qu'une simple collaboration professionnelle. Nos discussions tant scientifiques que personnelles m'ont permis de prendre de la hauteur, moi qui, avec ma culture d'ingénieur, ai plutôt tendance à vouloir garder les pieds sur terre :-)

J'ai également beaucoup de reconnaissance envers Michel Ménard qui à été mon "coach scientifique" et mon lien avec l'Université pendant les quelques années que j'ai passé en entreprise. Il m'a transmis son goût pour la recherche et m'a aidé à rejoindre le L3i lorsque je l'ai sollicité. Sans lui cette thèse n'aurait certainement pas eu lieu.

Plus généralement, je remercie toute l'équipe du Laboratoire Informatique, Image, Interaction de la Rochelle pour son accueil et pour les qualités humaines de ses membres.

J'ajoute bien entendu une mention spéciale à mes collègues doctorants du bureau 121bis (Mickael, Nathalie et Nicolas), qui ont égayé et enrichi mes journées par de parfois longues (Mika ?) mais finalement fructueuses discussions. Je n'oublie bien entendu

par les autres doctorants (je ne nommerai personne pour ne pas en oublier) avec qui notamment les repas au restaurant Universitaire ont souvent été inoubliables.

Merci également personnel administratif et technique du L3i : Christelle, Kathy et Dominique. On oublie souvent que sans vous, on ne pourrait pas faire grand chose...

Cette thèse a également été l'occasion de découvrir les joies de l'enseignement. Ainsi, je tiens à remercier ceux qui m'ont fait profiter de leur expérience pédagogique, en particulier Anthony, Frédéric et Vincent, car celle-ci a sans aucun doute contribué à améliorer la clarté et la qualité de ce rapport.

Enfin et surtout merci à ma famille : à mes parents pour leur bienveillance depuis maintenant 33 ans (également pour leur relecture du rapport, plus récente), à mon frère qui partage comme moi je goût de la connaissance, à Kélia et Adam qui ont parfois écourté mes nuit mais qui m'ont apporté beaucoup de bonheur et ont su être sages lorsque j'en avais besoin, et à Céline pour son amour et son soutient indispensable dans les dernières semaines de rédaction...

Table des matières

Table des figures

Liste des tableaux

Introduction

Cadre général et objectifs

L'attention, en tant qu'outil de gestion de l'information, nous permet de construire une perception adaptée à nos capacités et nos besoins. L'efficience [1] de ce mécanisme passe par une sélection contextualisée des données les plus pertinentes. Cette sélection est d'autant plus importante dans le cas de la vision biologique, que la quantité de données à traiter est massive : sans l'attention, notre système visuel ne pourrait pas traiter en un temps raisonnable un problème si complexe [Tsotsos 90].

Ces propriétés de gestion de la complexité font de l'attention un mécanisme clé pour la création de systèmes cognitifs artificiels [Paletta 08]. Pourtant, son utilisation est loin d'être systématique, car les modèles computationnels actuels sont généralement soit :

- *réalistes mais lents* : ils sont alors idéaux pour étudier les phénomènes attentionnels chez l'homme et ainsi valider des hypothèses sur le fonctionnement du cerveau, mais non adaptés à une utilisation quasi temps réel dans des applications de vision par ordinateur ;
- *rapides mais peu inspirés biologiquement* : ils sont alors généralement conçus spécifiquement pour une application et ne peuvent être généralisés à d'autres.

Il existe donc peu de modèles proposant un juste équilibre entre vitesse de traitement, plausibilité biologique, et possibilité d'adaptation à une large classe de problèmes.

Dans cette thèse, nous avons tenté de combler ce fossé entre attention et applications en développant un modèle d'attention visuelle spécialement étudié pour s'intégrer dans une large classe d'applications : la vision adaptative. Le terme adaptatif est important car les mécanismes attentionnels de sélection n'ont de sens que dans un monde ouvert et changeant, dans lequel il est nécessaire de lever certaines ambiguités. Dans le cas contraire (monde fermé : contexte figé et bien défini), l'attention n'est pas indispensable et les algorithmes de vision « classiques » sont particulièrement efficaces (cf. chapitre 1).

Dans ce cadre, notre objectif consiste alors à proposer un modèle d'attention compu-

1. L'efficience qualifie la capacité de produire le maximum de résultats avec le minimum d'effort, de dépense.

tationnellement efficient, c'est-à-dire gérant de manière optimale le compromis temps de traitement / qualité de résultats. Cette dernière peut être définie comme la capacité à proposer un mécanisme de sélection de l'information suffisamment plausible et adaptable à un contexte ou à des objectifs changeants.

Contributions

Nos contributions s'établissent à différents niveaux :

- dans le cadre de l'étude des modèles existants et de leurs applications, nous proposons une taxonomie permettant de révéler les propriétés des différentes classes d'algorithmes et leur adéquation à différents types d'applications (en particulier la vision par ordinateur) ;
- sur un plan plus théorique, nous proposons un modèle d'attention hybride hiérarchique / compétitif basé sur des équations proies / prédateurs. Celui-ci traite, *via* un mécanisme unique, différents problèmes inhérents à la construction d'un modèle d'attention : la fusion des différentes sources d'information (intensité, couleur, orientation, etc.), la dynamique des focalisations, et l'inhibition de retour. Nous montrons également que des mécanismes de rebouclage peuvent être introduits dans le système, afin d'adapter son comportement aux contraintes du système de vision hôte, en particulier en terme d'exploration de l'espace ;
- d'un point de vue expérimental, nous caractérisons l'influence des différents paramètres du modèle sur son comportement. Les critères d'étude retenus sont : la fidélité au modèle humain, la stabilité, la reproductibilité, l'exploration de l'espace et la dynamique du système. Les résultats obtenus montrent que notre modèle, bien que simplifié pour être computationellement efficace, affiche une plausibilité du même ordre (voire meilleure) que des modèles reconnus.

Organisation de la thèse

La première partie de la thèse permet de positionner nos travaux dans leur cadre scientifique. Dans le chapitre 1, nous délimitons notre champ d'étude en en précisant le contexte, et en situant nos travaux dans les différentes communautés concernées :

1. les modèles attentionnels ;
2. la vision par ordinateur ;
3. les systèmes adaptatifs.

Nous exprimons également notre problématique selon différents points de vue (applicatif, scientifique et implémentation) et concluons que pour piloter efficacement un système de vision adaptatif, un modèle d'attention spécifique s'avère nécessaire. Afin de mieux appréhender ce dernier, nous présentons brièvement le fonctionnement du système visuel humain et les principales familles de modèles théoriques et computationnels d'attention visuelle (chapitre 2). Chacune de ces familles ne répondant que partiellement

aux contraintes de simplicité algorithmique, extensibilité et gestion dynamique des mécanismes attentionnels, nous concluons cette partie en proposant un nouveau modèle computationnel d'attention visuelle. Celui-ci utilise une approche hybride réunissant la simplicité et l'extensibilité des modèles centralisés, avec la gestion dynamique de la compétition attentionnelle des modèles distribués.

Dans la seconde partie du manuscrit, nous décrivons notre modèle attentionnel. Nous l'étudions en deux phases, une première en boucle ouverte, l'autre en boucle fermée, c'est-à-dire en incluant l'interaction. Nous commençons par décrire la partie purement *bottom-up* (chapitre 3). Nous détaillons comment calculer en temps réel les cartes de caractéristiques grâce à l'utilisation généralisée d'images intégrales. Afin de résoudre le problème de compétition entre différentes cartes de singularité, nous présentons une approche utilisant un système d'équations proies / prédateurs. Nous justifions ce choix par ses propriétés théoriques et computationnelles. Nous exposons également les expérimentations menées, permettant d'étudier l'influence des différents paramètres de notre modèle sur son comportement (fidélité au modèle humain, dynamique, reproductibilité, exploration de l'espace, etc.). Dans un deuxième temps (chapitre 4), nous présentons une version bouclée munie de mécanismes d'auto-adaptation. Nous appliquons le même principe d'étude systématique des paramètres du système. La principale contribution de notre modèle apparaît alors dans ses possibilités d'adaptation dynamique à un objectif ou un contexte changeant.

Enfin, la dernière partie de la thèse est consacrée à la présentation de différentes applications exploitant notre modèle. Celles-ci ont été choisies pour leur variété, permettant ainsi de vérifier la flexibilité de notre approche. Dans le chapitre 5, nous montrons que l'analyse de la dynamique des focalisations attentionnelles générées par notre algorithme, permet de définir une mesure de la complexité des images, basée sur le trajet oculaire. Deux autres applications exploitent les *heatmaps* mises à jour à chaque pas de temps par notre modèle. La première permet d'effectuer un recadrage dynamique d'images ou de flux vidéos (chapitre 6). La seconde permet la segmentation d'images et vidéos en zones d'intérêt par extraction des vignettes les plus saillantes (chapitre 7).

Le modèle attentionnel hiérarchique et compétitif présenté dans cette thèse offre la possibilité, pour un système de vision adaptatif, de posséder un système de prétraitement temps réel permettant de guider l'exploration de la scène. Les capacités de rebouclage du système, ainsi que ses différents paramètres, permettent à tout moment de modifier le comportement attentionnel en fonction de l'objectif visé (exploration, suivi, besoin de reproductibilité, etc.). Une implémentation de l'architecture de vision théorique, présentée dans cette thèse, est une perspective à envisagée afin d'exploiter au mieux les caractéristiques du système attentionnel.

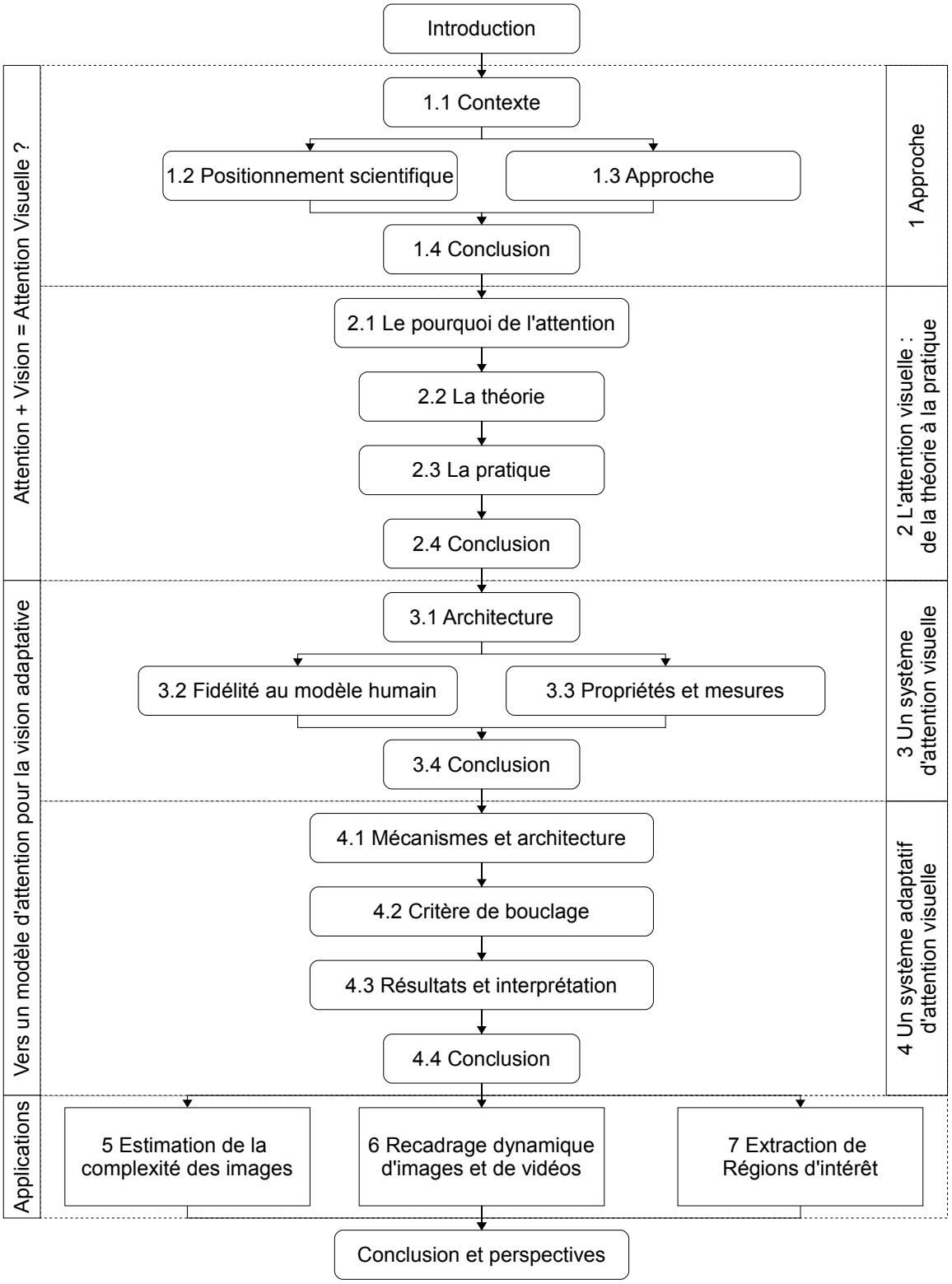


FIGURE 0.0.1: Organisation de la thèse

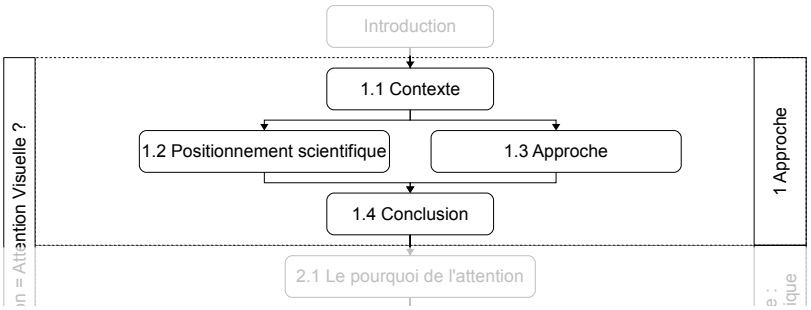
Attention + Vision = Attention Visuelle ?

« If we had really huge brains, say the size of watermelons, attention would play a much smaller role in our behavior. »"

(Mozer et Sitton)

Chapitre 1

Positionnement



L'objectif de ce chapitre est de situer nos travaux dans leur cadre scientifique, théorique et applicatif. Dans la section 1.1, nous précisons le contexte dans lequel ils s'inscrivent. Dans la section 1.2, nous détaillons leur positionnement au sein des communautés scientifique, liées à l'attention, la vision par ordinateur et l'adaptation. Nous exprimons également notre problématique selon différents types de contraintes (applicatives, scientifiques ou d'implémentation). Enfin, dans la section 1.3, nous présentons notre approche de la synergie attention / vision dans un contexte adaptatif.

1.1 Contexte

Lorsque nous observons une scène, sa perception et son interprétation nous paraissent généralement simples et sans ambiguïté. Ce faisant, et malgré 40 années de recherche en vision par ordinateur, la croissance exponentielle de la puissance de calcul disponible, et des caméras de plus en plus précises (voire même 3D), la conception d'un système performant, et surtout générique, est toujours hors de portée.

> « There are today numerous sophisticated methods for extracting visual information, but they seldom work consistently and robustly in the real,

> dynamically changing world »
>
> Eklundh et Christensen [Eklundh 01]

Cela ne remet pas en cause les progrès effectués en matière d'analyse de scène durant ces 40 dernières années. Si l'on considère un domaine d'application précis, et un certain nombre d'hypothèses concernant le contexte d'exécution du système, il est alors possible de concevoir des algorithmes égalant voire dépassant la performance humaine. Les domaines impactés sont variés : détection et suivi d'objets ou de visages, inspection de pièces industrielles, reconstruction de scène, analyse de mouvement, recherche d'information (image ou vidéo) par le contenu, etc. Cependant, dès que l'on souhaite traiter des tâches plus générales, dans un environnement complexe, de nombreux verrous doivent être levés : ambiguité dans l'interprétation des données , explosion de la combinatoire liée à l'exploration des données, etc. Une des voies possibles pour tenter de résoudre ce problème est de chercher des mécanismes efficaces de sélection des informations pertinentes.

On peut trouver des exemples de ce type de mécanisme dans le règne animal, pour lequel l'évolution a favorisé l'émergence du système attentionnel. Chez l'homme, il est étudié intensivement depuis la fin des années 1950 (d'abord par des psychologues, puis des neuropsychologues). Plus récemment, grâce aux neurosciences computationnelles, des modèles informatiques de l'attention (visuelle) sont apparus. Ils permettent de la modéliser et d'en comprendre le fonctionnement.

Les contraintes d'exécution quasi temps réel liées à la vision par ordinateur nécessitent un mécanisme de sélection de l'information efficace. Il semble alors judicieux de s'intéresser à l'attention dans ce cadre. Cependant, sa mise en œuvre soulève un nombre important de questions :

- Comment créer un système attentionnel suffisamment rapide pour qu'il ne reste qu'un prétraitement et permette ainsi l'exécution de vraies tâches de vision ?
- Jusqu'où doit-on pousser le réalisme biologique ?
- Comment prendre en compte les informations liées au contexte dans le mécanisme attentionnel ?

Cette dernière question ouvre le champ à des questions fondamentales sur les capacités d'adaptation des systèmes de vision. En effet, nous avons déjà précisé qu'un système de vision générique doit pouvoir s'adapter à un contexte d'exécution changeant. Chez l'homme, les mécanismes attentionnels abordent ce problème *via* leur voie descendante (*top-down*[1]) guidée par l'intention et le contexte. Celle-ci s'oppose à la voie ascendante (*bottom-up*), guidée par les *stimuli*. Dans le cadre d'un système de vision attentionnel, les mécanismes adaptatifs doivent donc être mis en place, *via* la voie *top-down* du système d'attention.

1. Les voies *bottom-up* et *top-down* sont abordées plus en détail en sous-section 2.2.3.2.

Créer un système de vision attentionnel et adaptatif est un problème vaste et complexe. Nous verrons dans les sections suivantes les domaines que nous avons choisis de couvrir dans cette thèse, et dans quelle mesure ceux-ci seront abordés et résolus.

1.2 Positionnement scientifique

1.2.1 Problématique

Toute recherche est associée à une problématique. Mais qu'entend-on exactement par problématique ?

> « Dans une recherche active, le chercheur ne choisit pas les problèmes à résoudre (l'histoire les lui impose) mais il crée sa problématique, c'est-à-dire que pour résoudre un problème donné il choisit un certain nombre de critères et élabore à partir de ceux-ci son système de recherche »
>
> Dumazedier, Ripert, Loisir et cult., 1966, p.30

La question que nous abordons dans cette thèse et à laquelle nous souhaitons apporter notre contribution est la suivante : comment proposer un modèle adaptatif d'attention permettant de guider, en temps réel, un système de vision en lui fournissant des informations sur l'importance relative des différents éléments de la scène ?

Pour que le tandem système attentionnel / système de vision fonctionne de manière efficace, l'un et l'autre doivent pouvoir s'adapter :
- le système attentionnel doit modifier son comportement en fonction des objectifs du système de vision (détection d'un type d'objet particulier, focalisation sur une zone de la scène ou au contraire exploration de celle-ci, etc.) ;
- le système de vision doit adapter / optimiser ses traitements en fonction de la saillance des différents éléments de la scène.

Pour traiter cette question, nous avons réduit notre champ d'étude en nous appuyant sur trois points de vue.

Point de vue applicatif

Le modèle d'attention visuelle que nous proposons est dédié à un type d'application restreint : la vision par ordinateur. C'est ce cadre, défini en section 1.3, qui a guidé notre étude des modèles existants. Il a également dirigé la construction de notre proposition et le choix des applications l'exploitant.

Point de vue scientifique

Une partie du système attentionnel que nous proposons est basée sur des équations proies / prédateurs, dérivées de celles de Volterra-Lotka [Volterra 28]. Les équations originales forment un système dynamique permettant de simuler l'évolution des quantités de proies et de prédateurs dans un espace sans-dimension. Notre extension de ce modèle s'applique à des cartes d'intérêt 2D. Ainsi les équations régissent le comportement de chacune des entités localisées dans cette espace 2D. Le système obtenu est un système complexe pour lequel des modifications de paramètres à l'échelle microscopique engendrent un comportement à l'échelle macroscopique difficilement prévisible.

Pour mieux cerner l'influence des différents paramètres de notre modèle sur son comportement, nous avons choisi d'étudier les propriétés de notre modèle en fonction de critères issus des contraintes applicatives. Ainsi, nous étudions :

- la fidélité : le modèle se comporte-t-il d'une manière proche de l'attention humaine ?
- la stabilité : le modèle est-il suffisamment fiable pour être utilisé dans une application de vision ?
- la reproductibilité : le modèle a-t-il un comportement déterministe ? Peut-on quantifié sa variabilité ?
- la dynamique : quelle est la réactivité du système face à un changement ?
- l'exploration de l'espace : la stratégie attentionnelle d'acquisition de l'information est-elle efficace ?

Point de vue implémentation

La grande majorité des applications de vision par ordinateur a besoin d'être exécutée en temps réel. Ainsi, certains aspects d'ingénierie sont importants à prendre en compte. Dans la mise en œuvre de notre modèle, nous utilisons différentes techniques afin d'accélérer les calculs. Ces solutions sont détaillées en annexe D, où nous décrivons dans quelle mesure ces choix accélèrent les traitements. Leur impact sur la qualité des résultats est abordée en sous-section 3.2.2.

Nous venons de définir la question que nous souhaitions aborder dans cette thèse et délimité notre champ d'étude. Dans la prochaine sous-section, nous positionnons nos travaux dans les communautés liées à notre problématique.

1.2.2 Communautés visées

Nos travaux sont au carrefour de trois communautés, représentées en figure1.2.1, dont les thèmatiques sont les suivantes :

- système visuel attention et saillance ;
- vision par ordinateur ;
- adaptation.

Sur cette figure, la dimension de chacun des cercles représente l'importance relative de nos contributions. La suite de cette section présente chacune de ces communautés, ainsi que notre positionnement au sein de celles-ci.

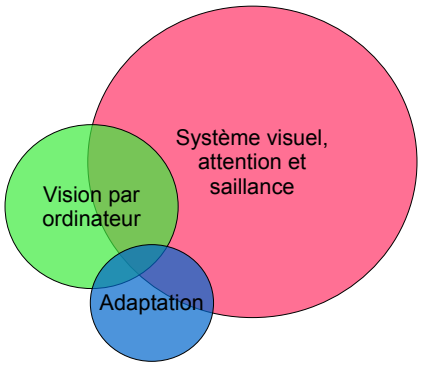


FIGURE 1.2.1: Intersection des différentes communautés abordées dans cette thèse.

Système visuel, attention et saillance

Cette communauté inclut des disciplines variées, allant de la psychologie à l'informatique en passant par la neurophysiologie. Son champ d'investigation concerne l'étude du système visuel avec ou sans prise en compte des mécanismes de l'attention. Nous abordons l'étude théorique du système visuelle en section 2.2.1, mais ne traitons que les modèles computationnels du système visuel lié à l'attention. Le lecteur intéressé par les autres modèles, intégrant ou non des phénomènes de *grouping* (voir l'annexe C sur la théorie de la forme), pourra consulter [Choe 01, Bednar 02, Behnke 03, Hérault 07, Benoit 10] ou encore [Li 01].

Nous ne pouvons préciser la portée de nos contributions dans le domaine de l'attention sans en donner une définition. L'une des plus anciennes et reconnues est donnée par le psychologue William James :

> « Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others. »
>
> [James 90], page 404.

Cependant, cette définition reste encore trop vague par certains aspects et à l'éclairage de la connaissance moderne des mécanismes cognitifs. L'attention est un terme générique

qu'il est nécessaire de caractériser. On peut en effet définir différents types d'attention [Larochelle 00] :

- l'éveil est l'ouverture sensorielle au monde qui nous entoure. Elle se manifeste lors du passage du sommeil à un état éveillé ;
- l'attention sélective ou focalisée est la sélection dans l'environnement d'une source de stimulation jugée, consciemment ou non, plus importante que les autres.
- l'attention maintenue intervient après l'attention sélective, afin de rester focalisé sur la tâche / le *stimulus* en cours pendant un certain temps, évitant ainsi le passage trop rapide d'un *stimulus* ou d'une activité à l'autre ;
- l'attention partagée ou divisée permet de traiter simultanément plusieurs types d'information ;
- la distractivité aux *stimuli* internes correspond à une difficulté à basculer son attention sur des *stimuli* externes. Dit en langage courant, c'est « être dans la lune » ;
- la distractivité aux *stimuli* externes correspond à une difficulté à contrôler son attention pour rester focalisé sur la tâche courante. En langage courant : « ne pas être concentré » ;
- la vigilance est la capacité à percevoir un *stimulus* externe, noyé dans un bruit de fond, au cours d'une tâche d'une certaine durée ;
- la supersistance est l'incapacité à cesser une activité malgré des signaux internes ou externes indiquant de le faire.

Dans le cadre de la vision (naturelle ou artificielle), l'attention est principalement sélective ou focalisée. Notons cependant que l'attention maintenue peut également jouer un rôle, par exemple, dans le suivi de cible. Nous y revenons au chapitre 4 lorsque nous abordons les mécanismes de rebouclage et d'adaptation de notre modèle.

Nous considérons également comme fondamentale la notion de saillance (ou prégnance) visuelle, puisqu'elle est intimement liée aux phénomènes attentionnels. Le dictionnaire Larousse définit d'ailleurs un élément saillant comme un objet « qui ressort sur le reste ; qui attire l'attention ». Bien entendu cette définition n'est pas réduite qu'aux aspects visuels mais peut également s'appliquer dans différents domaines tel que la linguistique [Landragin 04]. Dans le cadre de l'attention visuelle, ce terme est généralement utilisé pour désigner la carte de saillance des modèles computationnels centralisés (voir section 2.3.3.2), permettant de représenter le « potentiel d'attention » des différents éléments de la scène. Saillance n'est cependant pas attention : il est possible d'étudier la saillance de certains attributs visuels (coins, zones de fort contraste, etc.), en dehors de tout modèle d'attention [Hall 02].

Vision par ordinateur

Nos travaux sont également en lien avec la communauté « Vision par ordinateur ». C'est une communauté large, faisant appel à des disciplines variées telles : l'intelligence

artificielle, l'apprentissage automatique, les mathématiques, la neurobiologie, l'imagerie, la physique, le traitement du signal ou la robotique. Son spectre d'applications est large, allant de la détection / suivi d'objet, à la recherche d'images par le contenu.

Les recherches effectuées dans cette communauté ont longtemps été influencées par les travaux de David Marr [Marr 82], dont le paradigme de vision propose une analyse uniquement ascendante, centrée sur les données (figure 1.2.2). Cette approche purement *bottom-up* (cf. sous-section 2.2.3) a également inspiré les modèles d'attention visuelle de la même époque [Koch 85].

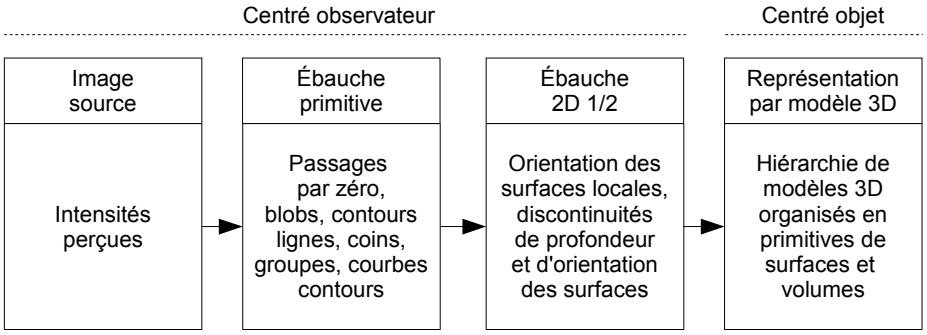


FIGURE 1.2.2: Le paradigme de David Marr. La scène est reconstruite étape par étape, uniquement à partir des données de l'image source.

Cependant, la majorité des applications de vision nécessite l'introduction d'informations concernant le contexte et / ou l'environnement. Cette composante descendante (*top-down*, cf. sous-section 2.2.3) est généralement incorporée de manière figée au moment de la conception du système. Nous avons, par exemple, utilisé ce type d'approche pour réaliser un système de détection et suivi de visages utilisé dans des applications interactives [Perreira Da Silva 07, Perreira Da Silva 08, Perreira Da Silva 09b]. Cela permet de réaliser des algorithmes performants, mais difficilement adaptables à un environnement changeant [Eklundh 01].

Il est alors intéressant de s'inspirer de certains mécanismes de la vision biologique afin d'améliorer les performances et / ou la flexibilité des algorithmes de vision par ordinateur. Cela peut être fait au niveau de l'architecture du système [Draper 07], en adaptant des éléments de théorie de la forme (cf annexe C) [Desolneux 01, Desolneux 04, Desolneux 08]ou par une analyse plus systématique des mécanismes de la vision humaine et leur adaptation à l'ordinateur [Wörgötter 04].

Le système attentionnel apparait alors comme un élément clé de la vision biologique pouvant améliorer les systèmes de vision artificielle. C'est en effet un mécanisme efficace de sélection des informations pertinentes (avec ou sans *a priori*). S'il est de surcroit

adaptable, on peut espérer obtenir une amélioration des performances même en environnement changeant.

Son intégration dans un système de vision artificielle permet de :

- pré-filtrer les informations contenues dans les images d'entrée afin de ne traiter en priorité que celles dont l'importance est jugée suffisante. Le système attentionnel agit alors comme filtre de sélection, permettant de réduire l'ambiguïté liée aux données (en réduisant / triant les données à traiter). C'est également un système d'optimisation de ressources puisque théoriquement la quantité de données à traiter sera moindre. Pour cela, le système attentionnel doit être rapide, afin de laisser suffisamment de ressources processeur libres pour les autres traitements du système de vision. Il doit également posséder une stratégie de sélection d'information adaptée, afin de ne pas diminuer les performances du système ;
- pré-calculer un certain nombre d'attributs visuels qui pourront être réutilisés par l'application de vision [Siagian 07].

Dans nos travaux, les systèmes de vision sont le terreau applicatif servant à définir les contraintes du système attentionnel. C'est à travers eux que sont justifiés les choix effectués lors de la conception. C'est également « à cause » de ce contexte que notre système attentionnel se doit d'être adaptatif.

Adaptation

La communauté étudiant les mécanismes de l'adaptation peut être séparée en deux sous-communautés. La première étudie les systèmes biologiques et sociaux : le système immunitaire, les écosystèmes, le cerveau, le comportement animal, l'évolution génétique, etc. La seconde applique ces mécanismes au domaine de l'informatique. On y retrouve entre autres les systèmes multi-agents ou la reconfiguration automatique de logiciels. Cependant, l'adaptation en informatique n'est pas liée à l'utilisation d'une méthode particulière. Tout système correspondant à la définition de l'adaptation est valide. Selon Thierry Vieville [Vieville 05] :

> « L'adaptation correspond à un processus par lequel un sujet, lorsqu'il enregistre une variation de l'environnement, modifie les paramètres d'un objet, à partir d'un modèle de référence, dans le but d'accomplir une tâche spécifique. »

Pour construire un système adaptatif il est donc nécessaire de définir :

- *un objet*, représentant la chose à adapter (un algorithme, un comportement, une entité, etc.) ;
- *un sujet*, représentant le mécanisme d'adaptation. Lorsqu'il y a auto-adaptation, sujet et objet peuvent être identiques ;
- *un modèle de référence*, qui correspond à la connaissance qu'à le sujet de l'objet et de ses interactions avec l'environnement. C'est à partir de ce savoir qu'une modification adaptée de l'objet pourra être possible ;

– *un environnement*, correspondant au contexte dans lequel la tâche est exécutée ;
– *une tâche*, correspondant à l'objectif que veut / doit atteindre le sujet.

Nous pouvons établir une relation entre cette définition et le système adaptatif de vision basé sur l'attention qui sert de cadre à nos travaux. Le mécanisme d'adaptation *(le sujet)*, lorsqu'il perçoit une variation dans la scène observée *(l'environnement)*, modifie les paramètres du système attentionnel (*l'objet*) afin de réaliser sa tâche de vision *(la tâche)*.

Notre contribution est alors duale :

– mettre en place des mécanismes d'adaptation dans notre système attentionnel. Ces mécanismes seront le *sujet* de l'adaptation ;
– étudier les paramètres du modèle afin de connaitre leur influence sur son comportement et ses performances. Cette étude participe à la constitution du *modèle de référence.*

Nous sommes, dans ce contexte, simples utilisateurs de mécanismes d'adaptation. Ceci afin de concevoir un système de vision performant, utilisant au mieux les propriétés de notre modèle attentionnel.

1.3 Approche

1.3.1 Cadre théorique : la simplexité

Au début de ce chapitre, nos évoquions l'apparente simplicité avec laquelle nous percevons le monde. Pourtant celui-ci est complexe : les informations à analyser sont nombreuses, souvent ambigües 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 [Berthoz 09] 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 [2] doit :

– séparer les différentes fonctions / être modulaire ;
– être rapide
– être fiable ;
– être flexible ;
– posséder de la mémoire ;
– avoir des propriétés de généralisation ;

Compte tenu de ses propriétés, l'attention répond parfaitement à cette caractérisation. Berthoz définit même ce mécanisme sélectif comme une fonction essentielle de notre existence en proposant de remplacer le célèbre « Je pense donc je suis » par un « Je

2. Nous entendons par « système simplexe », tout ensemble d'éléments mettant en œuvre des mécanismes simplexes.

choisis, donc je suis ».

1.3.2 Architecture du système de vision

Compte tenu de l'encrage important de l'attention dans la théorie de la simplexité, nous avons choisi de nous en inspirer pour établir le cahier des charges de notre système de vision adaptative (théorique) et de notre modèle attentionnel.

La figure 1.3.1 présente une vue fonctionnelle d'un système de vision intégrant notre système d'attention adaptatif. Conformément au principe de séparation des fonctions de la théorie de la simplexité [Berthoz 09], on y trouve différents modules :

- un système visuel, permettant de calculer les caractéristiques visuelles qui seront exploitées par le système attentionnel ou les tâches de vision.
- un système attentionnel, s'occupant de résoudre le problème de sélection de l'information visuelle par un mécanisme compétitif.
- des mécanismes d'adaptation, permettant de modifier le comportement du système attentionnel et des différentes tâches de vision.
- des tâches de vision, exploitant les informations fournies par les systèmes visuels et attentionnels.

Les deux premiers modules sont décrits et étudiés en détail dans le chapitre 3. Le chapitre 4 montre comment les mécanismes d'adaptation peuvent influer sur le système attentionnel. La mise en œuvre des différentes tâches de vision exploitant notre modèle dépasse le cadre de cette thèse ; cependant, des exemples d'application de notre modèle dans un cadre plus restreint sont présentées aux chapitres 5, 6 et 7.

1.3.3 Cahier des charges du modèle attentionnel

Le cadre de la théorie de la simplexité [Berthoz 09], le contexte applicatif, ainsi que l'architecture du système de vision adaptatif hôte, imposent différentes contraintes, qui doivent être prises en compte par notre modèle attentionnel.

Fiabilité

Le système doit délivrer le comportement attendu. L'influence éventuelle du paramétrage du système sur la fiabilité devra être étudiée.

Mémorisation

Le système attentionnel est un mécanisme de sélection. La mémorisation ne fait pas partie de ses attributions. Celle-ci est déléguée au système visuel.

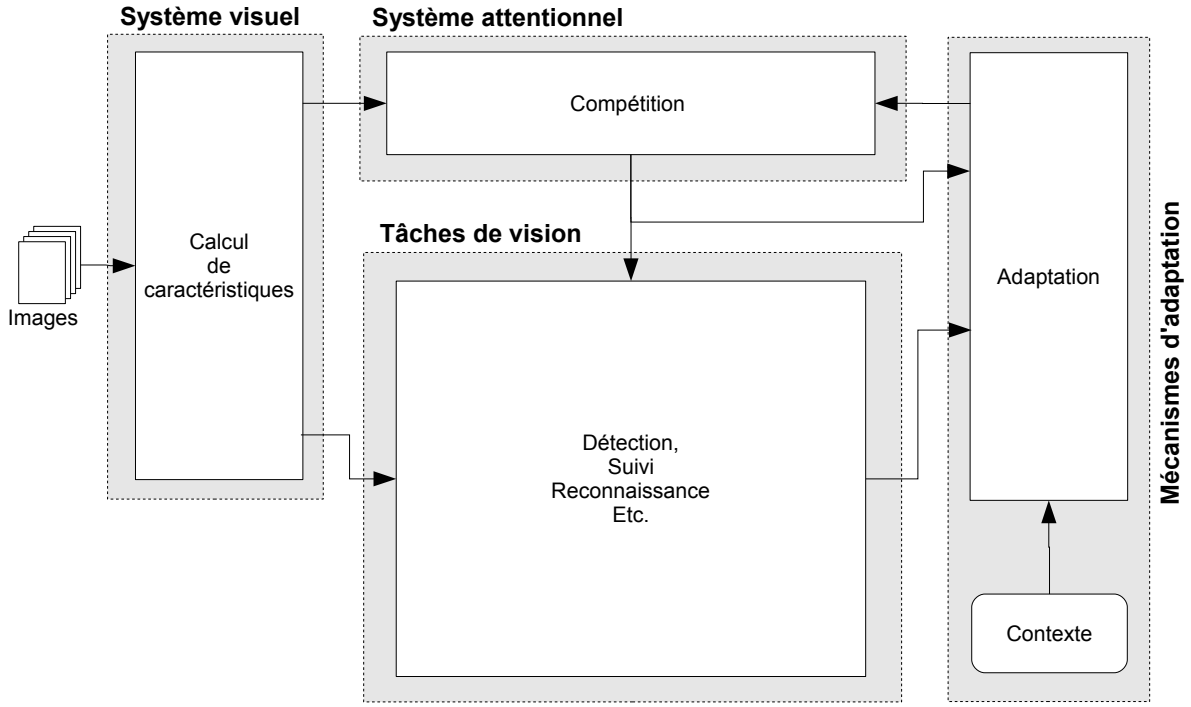


FIGURE 1.3.1: Architecture du système de vision.

Généralisation

Bien que conçu pour la vision par ordinateur, notre système attentionnel doit également être applicable dans d'autres situations. Nous avons veillé en particulier à ce que son comportement reste proche de celui du système attentionnel humain.

Vitesse

Les différents calculs doivent être rapides. Dans un système de vision complet, l'attention est un module au service des autres, il doit donc laisser suffisamment de ressources processeur pour permettre l'exécution concurrente d'autres tâches de vision.

Le système doit pouvoir fournir une estimation de la saillance des éléments de la scène à tout moment, de manière approximative si peu de temps est disponible, ou de manière plus précise si le temps le permet. Cela correspond à la définition d'un algorithme *anytime* [Zilberstein 96]. Ce mécanisme permet d'adapter le compromis temps de calcul / qualité des résultats, en fonction des besoins du système hôte.

Flexibilité

Le système de vision doit pouvoir s'adapter à un contexte (environnement et objectifs) changeant. Pour cela, son système attentionnel doit pouvoir modifier son comportement

via ses paramètres ou ses mécanismes d'adaptation.

Le système doit être facilement extensible. Différents types de caractéristiques doivent pouvoir être prises en compte sans remettre en cause l'architecture du système.

Le système doit être dynamique. Il n'est pas de la responsabilité du système de vision de décider comment doit être parcourue la scène. C'est au système attentionnel d'effectuer ce choix à chaque instant, en fonction de ses contraintes propres, de celles du système de vision, et de celles plus générales du contexte d'exécution.

1.4 Conclusion

Dans ce chapitre nous avons positionné nos travaux dans leur contexte applicatif et scientifique, puis esquissé notre approche de l'attention visuelle à travers la vision adaptative. Cependant, la caractérisation de notre problématique est encore incomplète. Avant de présenter le façon dont nous avons abordé la construction d'un modèle attentionnel adaptatif, il nous faut introduire les différents travaux à partir desquels nous nous sommes basés, ou par rapport auxquels nous devons nous situer. Cela permettra d'affiner notre cahier des charges, en fonction des propriétés des modèles existants.

Points clés

Positionnement

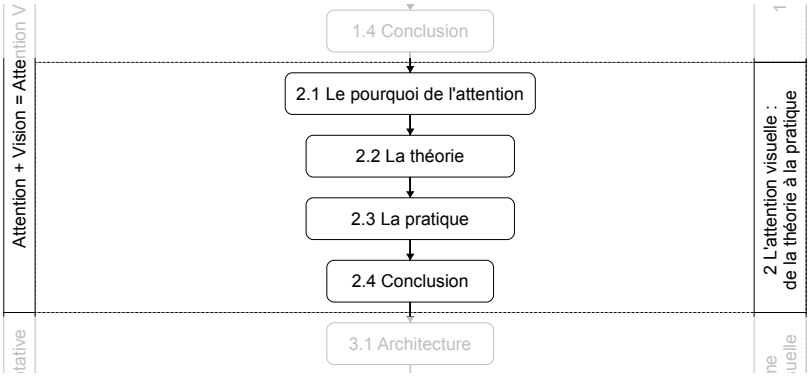
- ❏ Nos contributions sont principalement situées dans la communauté étudiant l'attention et ses différents modèles computationnels.
- ❏ Les systèmes de vision sont le socle applicatif de cette thèse. Nous y puisons les différentes contraintes guidant la conception du modèle attentionnel.
- ❏ Nous utilisons les principes de l'adaptation afin de concevoir un système attentionnel pouvant coopérer avec son système de vision hôte.
- ❏ L'attention est un mécanisme simplexe

Contributions

- ❏ Nous proposons un cahier des charges pour la mise en œuvre computationnelle d'un modèle d'attention dédié à la vision par ordinateur.
- ❏ Nous transposons les propriétés des mécanismes simplexes de Berthoz au cadre de la modélisation computationnelle de l'attention.

Chapitre 2

L'attention visuelle : de la théorie à la pratique



Nous utilisons la vision bien plus que tous nos autres sens : livres, télévision, internet, une majorité de nos sources d'information est visuelle. Notre vie quotidienne (déplacements, orientation, coordination des mouvements) est également guidée par cette modalité. Ces différentes tâches nécessitent un système visuel particulièrement développé (on estime que près d'un tiers de notre cerveau participe au traitement de l'information visuelle [Chalupa 03]).

Pour atteindre le niveau de performance que nous lui connaissons aujourd'hui, les mécanismes de l'évolution ont lentement créé, adapté et optimisé notre système visuel. Lorsque nous devons créer un système de vision artificiel, il apparaît alors intéressant de s'inspirer de ce que la nature a longuement élaboré et affiné. Cependant, l'architecture des ordinateurs est plutôt éloignée de celle du cerveau. Malgré des progrès récents (multi-cœurs, traitement sur GPU, cloud computing, etc.) les traitements y sont encore principalement sériels alors que notre cortex est massivement parallèle. Même si leur mise en œuvre est possible, il n'est donc pas toujours computationnellement efficace de

copier fidèlement tous les mécanismes de notre système visuel. Une étude plus générale des concepts utilisés s'avère cependant d'une aide précieuse pour améliorer le traitement de l'information dans les algorithmes de vision par ordinateur. L'un de ces concepts-clés permet de sélectionner les informations les plus pertinentes dans le flot de données auquel nous sommes confrontés : c'est l'attention visuelle.

Dans ce chapitre, nous proposons d'étudier comment passer des théories à la pratique. En effet, pour comprendre les apports et les limites des modèles computationnels existants, il est important de pouvoir les situer dans une théorie plus générale de l'attention visuelle. Pour cela, nous commençons par répondre à la question « à quoi sert l'attention visuelle ? » (section 2.1). Puis, en section 2.2, nous présentons les principaux modèles théoriques ainsi que leurs propriétés. Enfin, en section 2.3, nous abordons les modèles computationnels d'attention visuelle et leurs applications.

2.1 Le pourquoi de l'attention

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.

L'œil et la rétine ne capturent pas une image fidèle du monde. La répartition des différentes cellules photosensibles sur la rétine n'est pas homogène, le centre de celle-ci (la fovéa) contient beaucoup plus de récepteurs que la périphérie. Il en découle une représentation bien plus précise au centre de notre champ de vision (figure 2.1.1). Ainsi, il est nécessaire de déplacer nos yeux afin de capturer tous les détails de la scène et en construire une représentation cohérente [Rensink 00].

Notre cortex visuel décompose le signal qu'il reçoit en différentes caractéristiques de complexité croissante (intensité, couleur, orientation, puis coins, lignes intersections, etc.), générées dans des zones séparées (mais interconnectées) du cerveau [VanRullen 03]. Cette spécialisation se retrouve également dans le traitement de la reconnaissance des objets et leur localisation (hypothèse des voies centrales et dorsales). Notre cerveau doit alors reconstruire une représentation cohérente à partir d'un ensemble d'informations relativement hétérogènes.

Comme nous le constatons, 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.

FIGURE 2.1.1: Flou rétinien. Les images capturées par la rétine sont bien plus précises au centre (ici on a focalisé sur le visage de la jeune fille habillée en blanc).

La sélection comme conséquence de nos capacités limitées

Cette théorie, qui est la plus largement répandue, 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.

Cette thèse a été initiée par [Broadbent 58] lors de sa définition de l'attention sélective précoce (voir annexe A). Elle fut ensuite reprise par de nombreux modèles théoriques [Deutsch 63, Treisman 60, Treisman 69, Treisman 80]. Plus récemment, Tsotsos [Tsotsos 90] a utilisé cette hypothèse afin de démontrer que si l'on se place du point de vue de la complexité, il est nécessaire de mettre en place certains mécanismes (attentionnels) afin de rendre possible le processus de vision. C'est également le postulat de base d'un grand nombre de modèles computationnels d'attention visuelle [Itti 98, Ouerhani 03a, Frintrop 07].

La théorie de la simplexité [Berthoz 09], présentée au chapitre 1, place également l'attention parmi les mécanismes clés de simplification de la complexité. Cependant elle ne présente pas l'attention comme une conséquence de nos capacités limitées, mais comme un outil permettant d'utiliser au mieux nos capacités en fonction de nos besoins.

La sélection comme objectif fonctionnel

Une théorie alternative [Allport 87, Neumann 87, van der Heijden 97] consiste à considérer que nos capacités de traitement ne sont en rien limitées. Pour les partisans de cette théorie, avoir un plus gros cerveau ne nous dispenserait pas des mécanismes attention-

nels. L'attention trouve alors son origine dans une autre explication. 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 (cf. B), et nos mains ne peuvent manipuler qu'un (voire deux) objets simultanément, etc. Ce sont donc nos capacités d'action qui sont limitées et imposent une sélection de l'information perçue afin de pouvoir la traiter correctement.

Un autre argument évolutif vient étayer cette théorie. Certains *stimuli* ont une valeur de survie plus importante (repérer le mouvement d'un serpent le plus rapidement possible, sans avoir à analyser toute la scène visuelle est un avantage évolutif). On peut également démontrer qu'il est sous optimal de traiter toute l'information visuelle que nous recevons et qu'un mécanisme de sélection est nécessaire pour effectuer correctement cette tâche [Harel 09]. Dans le même esprit, Van Rullen & Koch [Van Rullen 05] expliquent que dans des cas très simples (objets isolés, familiers et bien contrastés), l'organisation de notre système visuel pourrait permettre le traitement des objets sans faire appel aux processus attentionnels. Cependant, les scènes visuelles auxquelles nous sommes confrontés sont généralement beaucoup plus complexes et bruitées. Dans ce contexte, les représentations que nous générons sont confuses et ambiguës. L'attention sert alors de biais permettant de sélectionner la plus probable. C'est la thèse défendue par la théorie de la compétition biaisée, que nous abordons en section 2.3.3.1

Conclusion

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 05]. Il est également indispensable à la construction d'une représentation cohérente de notre environnement [Rensink 00] et à la détection de changements dans celui ci [Rensink 97].

La prochaine section présente les principaux modèles théoriques d'attention visuelle. Ceux-ci proposent des mécanismes permettant d'expliquer les processus mis en œuvre lors du déploiement de l'attention.

2.2 La théorie

Les processus attentionnels entrent en jeu dans de nombreuses zones de notre cerveau. L'attention module notre perception visuelle mais aussi auditive ; elle entre également en jeu dans de nombreuses tâches cognitives : perception, mémorisation, remémoration, contrôle des mouvements, etc. Son étude est un domaine complexe, aux frontières de nombreuses disciplines (neuropsychologie, neurobiologie, neurosciences cognitives, etc.)

Dans cette section nous ne nous attarderons pas sur l'anatomie du système attentionnel. Les zones impliquées sont nombreuses et varient selon les différentes approches / théories (voir section 2.2.2.3). Nous décrivons différentes théories de l'attention (et

en particulier de l'attention visuelle), ainsi que les différents concepts et propriétés associés. Bien que présentant les principales théories, la sélection proposée n'est pas exhaustive. Les lecteurs désirant la compléter pourront consulter [Driver 01],[Tsotsos 05a, Tsotsos 07]ou [Styles 06].

2.2.1 Le système visuel

Il est difficile d'aborder les théories de l'attention visuelle sans avoir évoqué le fonctionnement du système visuel. Nous ne présentons ici que les caractéristiques qui nous paraissent importantes pour la compréhension des théories décrites en section 2.2.2. Des informations plus précises pourront être trouvées en annexe B et dans des ouvrages de référence tels que [Hubel 95] ou [Rodieck 03].

L'étude de la rétine et des différentes aires du cortex visuel (figure 2.2.1) permet de déterminer les principales caractéristiques du système de vision humain [Behnke 03]. Nous les présentons succintement ci-dessous :

- la rétine n'a pas une résolution spatiale uniforme. Elle est plus précise en son centre (la fovéa) qu'à sa périphérie ;
- l'information visuelle est séparée en canaux chromatiques (couple d'opposition rouge / vert et jaune / bleu) et achromatique (luminance) ;
- l'information visuelle est séparée en différents canaux sensibles au mouvement ou à une orientation et une fréquence spécifiques ;
- le principal mécanisme de transformation de l'information visuelle est le filtrage centre-périphérie (*center-surround*) ;
- deux voies de traitement des données issues de la rétine coexistent :
  - la voie dorsale (voie « où ») gère les aspects spatio-temporels des signaux reçus. Elle travaille rapidement sur des signaux de faible résolution ;
  - La voie ventrale (voie « quoi ») est impliquée dans la reconnaissance des objets.

  Dans la sous-section suivante nous présentons les principales contributions théoriques qui sont maintenant reconnues comme les fondements de l'étude de l'attention visuelle.

2.2.2 Théories fondatrices

2.2.2.1 La théorie d'intégration des attributs

Ce modèle particulièrement influent, a été développé par Treisman [Treisman 80] au tout début des années 80 et a servi de base à de nombreux modèles computationnels à partir de la fin des années 1990. Le point de départ de la *« Feature Integration Theory »* est la découverte du « problème » de correspondance (*binding problem*).

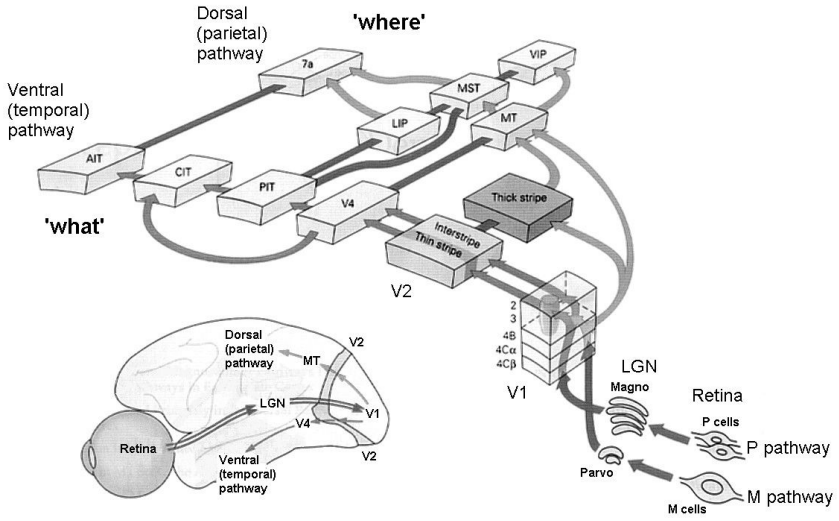


FIGURE 2.2.1: Le système visuel humain : de la rétine aux différentes aires du cortex visuel. Illustration issue de [Behnke 03].

Ce « problème » est lié à la façon dont l'information visuelle est traitée par notre cerveau. En effet, les différentes caractéristiques (forme, couleur, mouvement, etc.) des *stimuli* auxquels nous sommes soumis sont encodées dans des zones partiellement indépendantes. De plus, nos structures corticales hiérarchiques (champs récepteurs de plus en plus grands) sont organisées de façon à rendre la détection de ces caractéristiques relativement indépendante de leur position sur la scène visuelle. Enfin, le plus souvent, les scènes que nous observons comportent plusieurs objets. Un mécanisme permettant de mettre en correspondance les différentes caractéristiques détectées, afin de construire une représentation cohérente des objets, est alors nécessaire.

Treisman propose que ce mécanisme de correspondance (*binding*) soit pris en charge par l'attention visuelle sélective. Pour résoudre le problème, elle propose que la scène soit balayée par un faisceau attentionnel de taille variable (figure 2.2.2). Ce faisceau bloque les informations qui ne sont pas situées dans son rayon. Ainsi, il est possible de faire correspondre toutes les caractéristiques détectées dans cette zone afin d'en construire une représentation cohérente. En déplaçant le faisceau au cours du temps, on construit progressivement une perception globale de la scène.

2.2.2.2 Processus attentionnels automatiques et contrôlés

Une autre contribution, qui a marqué de manière importante les communautés liées à l'étude de l'attention, est celle de Schneider [Schneider 77]. Celle-ci introduit une distinction entre deux types de processus de traitement de l'information : automatiques et contrôlés (tableau 2.1). Ces travaux serviront notamment de base aux études de

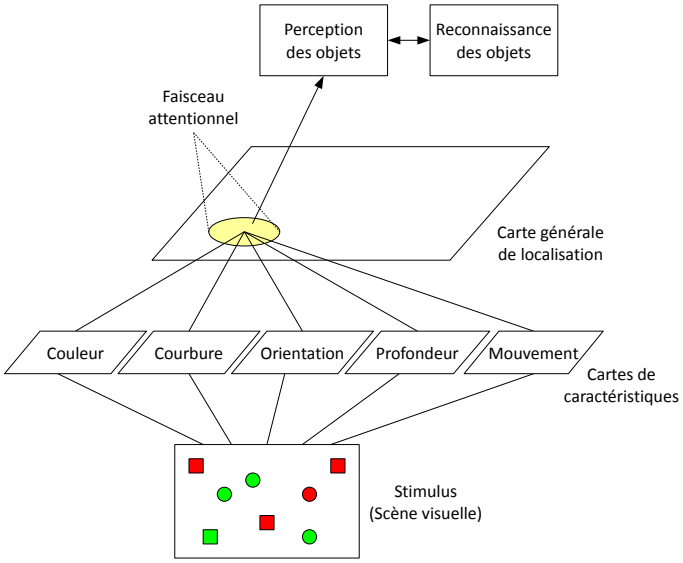


FIGURE 2.2.2: La théorie d'intégration des attributs [Treisman 80].

[Posner 80] qui mettra à jour deux types d'orientation de l'attention : endogène et exogène (voir 2.2.3).

	Processus automatique	Processus contrôlé
Type d'exécution	parallèle	sériel
Vitesse d'exécution	rapide	lente
Contrôle en cours d'exécution	nul	complet
Niveau de conscience	faible	élevé
Implication de l'attention	non nécessaire	indispensable
Effort cognitif	faible	important

TABLE 2.1: Caractérisation des processus automatiques et contrôlés de traitement de l'information (d'après [Schneider 77]).

2.2.2.3 Les modèles d'attention en tripode

Contrairement au modèle de Treisman qui propose une explication des mécanismes internes de l'attention, les modèles regroupés dans cette sous-section se focalisent sur une explication plus anatomique du phénomène. C'est une approche complémentaire aux théories vues précédemment, puisqu'elle autorise une vision plus « systémique » de l'attention, permettant de dégager un ensemble de blocs fonctionnels qui, dans une approche computationnelle, peuvent guider l'architecture des modèles.

Le modèle de Mesulam

[Mesulam 81] définit un réseau comprenant trois régions principales (figure 2.2.3) : le cortex pariétal supérieur, le cortex cingulaire et une région appelée *frontal eye field* située dans le cortex frontal. Ces régions sont associées à des tâches spécifiques et contiendraient toutes une carte spatiale nécessaire à l'expression de l'attention :

- la région pariétale serait responsable de la représentation sensorielle. Elle contiendrait une carte perceptive du monde extérieur ;
- la région cingulaire serait liée à la motivation et à nos émotions. Elle permettrait de diriger notre attention en fonction de notre état interne. Elle contiendrait une carte de motivation, permettant de réguler l'allocation spatiale de l'attention ;
- la région frontale serait liée à l'exécution de nos activités motrices. Elle contiendrait une carte d'exploration, permettant de coordonner l'exploration, l'atteinte ou la fixation d'une cible.

Ces zones seraient connectées entre elles, ainsi qu'à différentes structures corticales et sous-corticales.

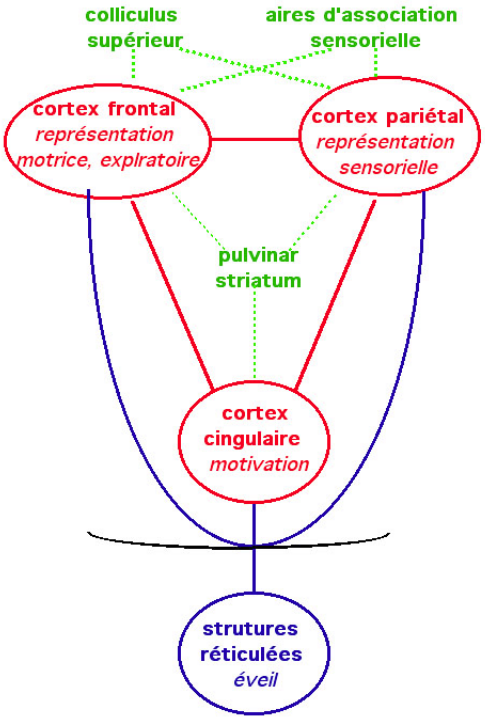


FIGURE 2.2.3: Le réseau attentionnel de Mesulam [Mesulam 81].

Le triple réseau attentionnel de Posner

[Posner 90] propose également un réseau triple (figure 2.2.4). Celui-ci a des points communs avec celui de Mesulam, mais il propose un découpage différent et regroupe les aires corticales dans des unités fonctionnelles communes.

Le cortex pariétal supérieur, le pulvinar et le colliculus supérieur sont regroupés au sein d'un réseau attentionnel dit postérieur. Celui ci serait chargé de l'orientation de l'attention. Les différentes aires seraient impliquées respectivement dans le désengagement, l'engagement et le mouvement de l'attention vers une nouvelle cible.

Le cortex cingulaire et l'aire motrice supplémentaire sont groupés dans le réseau attentionnel antérieur. Celui-ci serait impliqué dans la détection des objets et la sélection des réponses appropriées.

Enfin, le locus cœruleus et ses connexions vers les cortex pariétal et cingulaire forment le troisième réseau, dont le rôle serait d'assurer le maintient de la vigilance.

Ce modèle postule également que ces modules sont indépendants du reste du système cognitif. En effet, des lésions situées sur une ou plusieurs des régions inclues dans les trois différents réseaux, engendrent des déficiences uniquement attentionnelles (la mémoire ou le langage sont par exemple intacts).

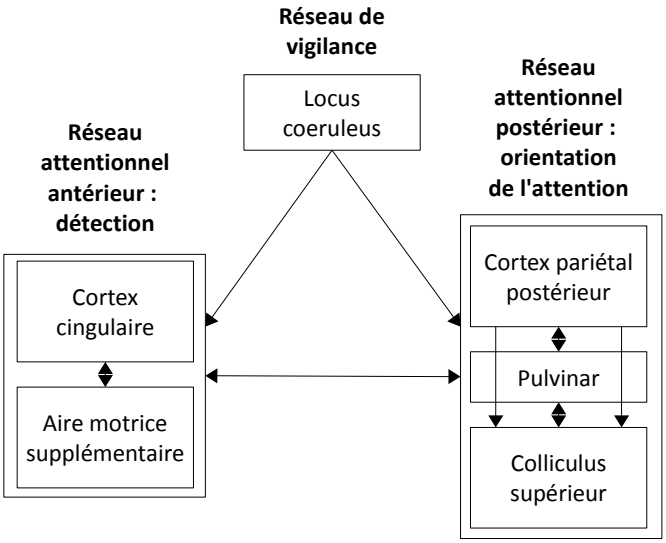


FIGURE 2.2.4: Le triple réseau attentionnel de Posner [Posner 90].

Posner a également contribué à différentes avancées importantes dans la compréhension des phénomènes attentionnels. Il est notamment à l'origine de l'intégration des notions d'attention ouverte (*overt*) et couverte (*covert*) [Posner 80], que nous abordons en section 2.2.3.1. D'un point de vue expérimental, il est également le père du paradigme

de l'orientation spatiale signalée (*attention cueing*) [Posner 78]. Cette méthode expérimentale (et ses dérivés) a permis de nombreuses avancées dans l'étude de l'attention top-down (section 2.2.3.2).

Les fenêtres attentionnelles de Laberge

[Laberge 95] se démarque des deux modèles précédents par sa description des mécanismes internes d'allocation de l'attention.

Les trois aires principales concernées par le modèle de Laberge sont :

- le cortex antérieur dont le rôle serait le contrôle de l'attention (sélection, préparation et maintien) ;
- le cortex postérieur impliqué dans l'expression de l'attention, et notamment sa localisation ;
- le pulvinar (thalamus), qui serait un filtre permettant le rehaussement de la cible de l'attention ainsi que l'inhibition latérale autour de la cible.

Ce modèle considère que l'attention est allouée assez tardivement. Celle-ci consisterait à rehausser les informations issues d'un codage sensoriel réalisé automatiquement. Ce rehaussement correspondrait à une fenêtre temporelle liée, selon sa durée, à une focalisation de l'attention ou à un processus d'attention préparatoire.

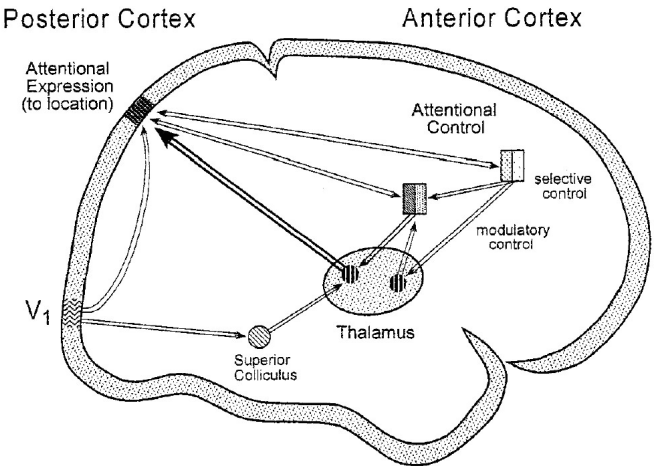


FIGURE 2.2.5: Le réseau attentionnel de Laberge [Laberge 95].

2.2.3 Les (petites) attentions...

Dans les sous-sections précédentes nous avons décrit différents modèles d'attention. L'étude de ces modèles fait ressortir différents concepts que nous présentons ici séparément afin de les rendre plus « saillants ». Ce sont en effet des notions clés qui serviront

de base à la présentation et l'analyse des modèles computationnels présentés dans le chapitre suivant.

2.2.3.1 Attention ouverte ou couverte

L'attention ouverte (*overt attention*)

Ce type d'attention est le plus simple à observer : lorsque nous portons notre attention sur un objet, nos yeux se déplacent afin de fixer cet objet. C'est à partir de cette constatation, somme toute triviale, que les premières ébauches de la notion d'attention ont été définies [1].

> While the sense organs are occupied with one object, they cannot be simultaneously be moved by another so that an image of both arises. There cannot therefore be two images of two objects but one put together from the action of both.
>
> Hobbes (1655)

L'un des pionniers de l'étude de l'attention visuelle ouverte est Yarbus [Yarbus 67]. Il a beaucoup étudié le lien entre attention visuelle et mouvement des yeux (figure 2.2.6). Il a notamment montré que la trajectoire de notre regard lors de l'exploration d'une scène dépendait de la tâche demandée (figure 2.2.7), montrant ainsi que l'allocation de notre attention (ouverte) n'est pas un processus uniquement *bottom-up*, mais également *top-down* (voir 2.2.3.2).

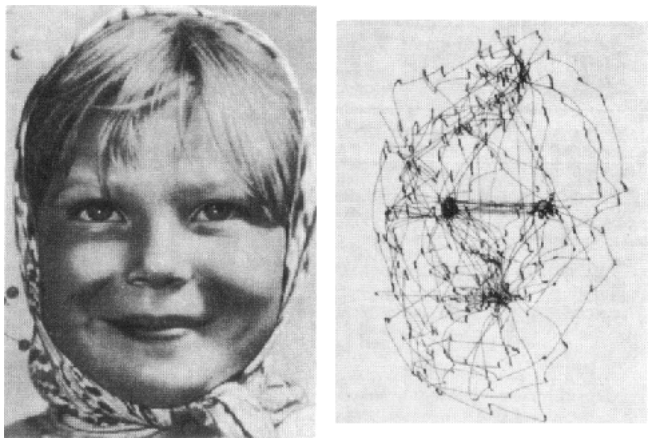


FIGURE 2.2.6: Trajectoire oculaire lors de l'observation d'un visage [Yarbus 67].

1. On pourra consulter le chapitre introductif de [Itti 05b] pour un historique des recherches concernant l'attention visuelle.

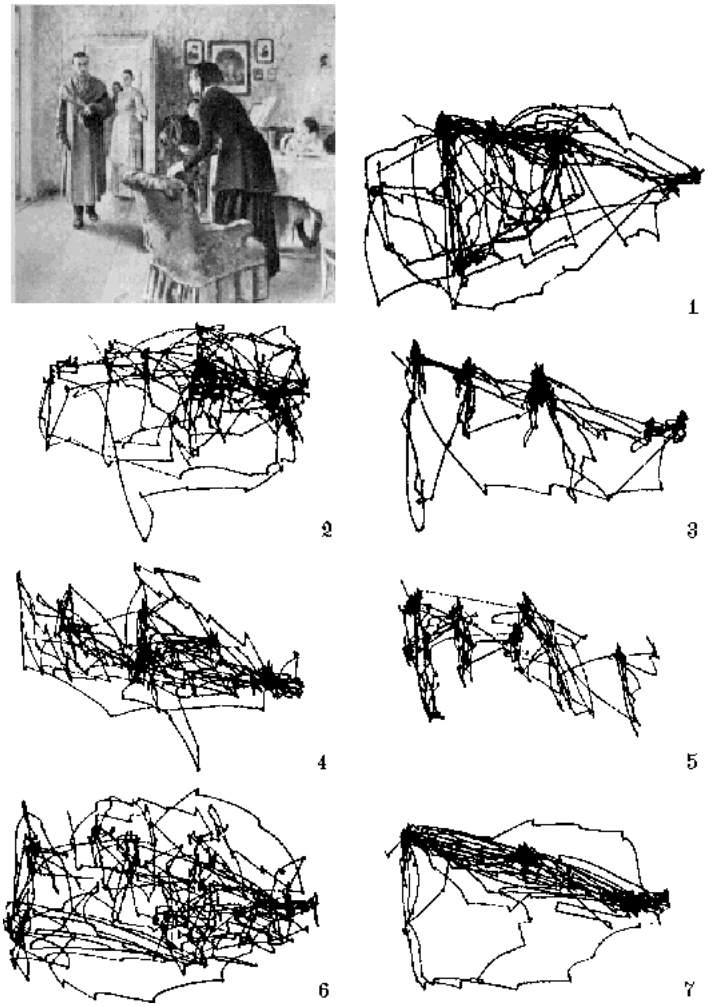


FIGURE 2.2.7: Les mouvements des yeux sont dépendants de la tâche confiée à l'observateur [Yarbus 67]. (1) observation libre. Avant l'expérience on a demandé aux sujets (2) déterminez la situation matérielle de la famille, (3) donnez l'âge des personnes, (4) essayez de deviner ce que la famille faisait avant l'arrivée du « visiteur surprise », (5) mémorisez les vêtements portés par les personnes, (6) mémorisez la position des personnes et objets dans la pièce (7) estimez depuis combien de temps le « visiteur surprise » n'a pas vu la famille. Chaque enregistrement a duré 3 minutes.

L'attention couverte (*covert attention*)

L'attention couverte correspond à notre faculté à focaliser notre attention sur une cible (objet ou position) sans déplacer nos yeux. Puisqu'elle n'implique aucun mouvement oculaire, son observation est particulièrement difficile. Cependant, dès 1896, on trouve des traces d'expériences d'Hermann Von HelmHoltz rapportant ce phénomène [Nakayama 89]. Sa prise en compte dans des modèles d'attention visuelle est bien plus tardive. On la doit à Posner[Posner 80] qui effectue le lien entre attention ouverte et couverte en mettant en place des expériences ingénieuses d'orientation spatiale signalée (*attention cueing*). Le principe de ces expériences est de mesurer le temps de réaction d'un sujet lors d'une tâche de recherche de cible. Avant l'affichage de la cible, on présente au sujet un indice (valide ou non) permettant au système d'attention couverte d'anticiper la position d'apparition de la cible. Ces expériences montrent que lorsque l'indice est valide, le temps de réaction du sujet est plus rapide que sans indice. A l'opposé, lorsque l'indice n'est pas valide, le temps de réaction du sujet est plus lent que sans indice. Posner conclut à un modèle d'attention séquentiel : l'attention couverte est allouée avant l'attention ouverte afin de faciliter le déplacement des yeux vers des *stimuli* « intéressants ».

Deux autres approches tentent d'établir un lien entre attention ouverte et couverte :

- Le modèle d'indépendance [Klein 80]. Attention ouverte et couverte sont indépendantes, elles se déploient simultanément car elle sont déclenchées par la même source de données : les informations visuelles.
- La théorie de l'attention pré-motrice [Rizzolatti 87]. L'attention couverte ne serait qu'un mécanisme préparatoire aux saccades oculaires. L'attention devient alors un produit dérivé du système moteur.

2.2.3.2 Les approches exogène (*bottom-up*) et endogène (*top-down*)

Comme nous l'avons évoqué en section 2.2.2.2, Schneider [Schneider 77] propose une séparation entre processus attentionnels automatiques et contrôlés. Cette notion a été étendue, notamment par Posner [Posner 80] afin de définir deux concepts utilisés couramment dans les modèles d'attention, aussi bien théoriques que computationnels.

L'attention exogène *(*également appelée ascendante ou *bottom-u*p) représente l'ensemble des processus automatiques, déclenchés par les *stimuli* externes captés par notre système visuel. C'est par ces mécanismes que, par exemple, nous tournons la tête si nous percevons un mouvement brusque à la périphérie de notre champ de vision.

L'attention endogène *(*également appelée descendante ou *top-down)* est volontaire et dépend, par exemple, de nos objectifs. C'est typiquement le type d'attention que nous mettons en œuvre lorsque nous jouons au fameux jeu « où est Charlie ? » (figure 2.2.8).

Ne dépendant pas de facteurs extérieurs, l'attention *bottom-up* est plus simple à modéliser que son homologue *top-down*. Ainsi, les premiers modèles computationnels

d'attention visuelle étaient basés exclusivement sur celle-ci [Koch 85, Itti 98]. Nous voyons en section 2.3 que c'est également le cas de nombreux modèles actuels. Cependant, comme l'évoque [Itti 01b], il est difficile chez l'homme de séparer les deux types de traitement (*bottom-up* et *top-down*). Dans cette approche, Mozer [Mozer 98] et Deco [Deco 04] ont été précurseurs en proposant des modèles intégrant nativement ces deux aspects. D'autres modèles computationnels récents intègrent également une influence *top-down* mais cette fois dans des modèles à l'origine purement *bottom-up* [Navalpakkam 05b, Navalpakkam 05a, Torralba 06].

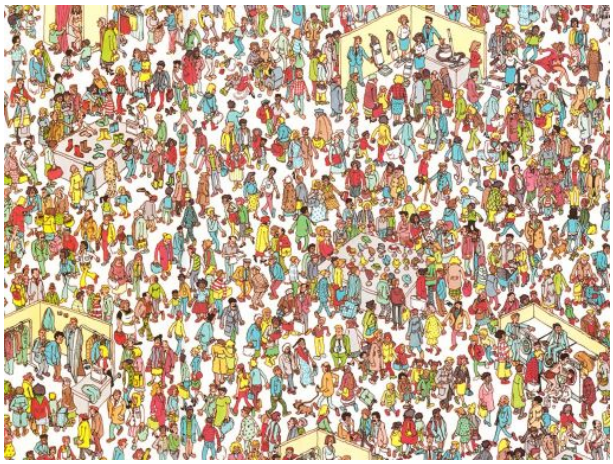


FIGURE 2.2.8: « Où est Charlie ? », application type de l'usage de notre attention exogène. L'objectif est de trouver un personnage grand et maigre, au pull rayé rouge, avec un bonnet, une canne et des lunettes.

2.2.3.3 Attention orientée espace ou objet

Les études de psychologie comportementale ont permis de révéler deux modes d'allocation de l'attention visuelle. Posner [Posner 78, Posner 80, Posner 90], montre par exemple que l'attention peut être dirigée dans des zones plus ou moins grandes de la scène visuelle. Ce type de conclusion a permis l'établissement de modèles où l'attention est vue comme une sorte de faisceau [Treisman 80], gradient [LaBerge 89] ou zoom [Eriksen 85].

D'autres études [Duncan 84] suggèrent que l'allocation de l'attention pourrait être dirigée vers des objets entiers, qui seraient le résultat d'un regroupement (*grouping*) effectué pré-attentivement. La théorie de l'attention orienté objet permettrait de faire le lien entre attention et théorie de la forme (Annexe C). Elaborée par les psychologues gestaltistes [Guillaume 37], cette théorie présuppose l'existence de différentes règles (continuité, proximité, similitude, etc.) utilisées par notre cerveau pour structurer la scène visuelle.

Puisque les deux phénomènes sont observés, il semblerait que les deux mécanismes soient impliqués dans l'allocation de l'attention. Il est cependant difficile de savoir pour l'instant s'il s'agit de deux processus indépendants, ou deux manifestations différentes d'un même processus. La solution viendra certainement des modèles plus complexes développés récemment, intégrant à la fois attention *bottom-up* et *top-down* [Hamker 05b, Torralba 06]. En effet, l'attention *top-down* étant principalement basée objet (notre représentation du monde est sémantique), on peut établir un lien entre une attention spatiale *bottom-up* et une attention objet t*op-down* [Ji 08].

2.2.3.4 Attention centralisée ou distribuée

L'idée d'une carte centrale de représentation spatiale de l'attention est partagée par de nombreux modèles théoriques. Pour Treisman [Treisman 80], à l'origine de ce concept, elle est appelée *carte générale de localisation* (master map of locations), pour Kock & Ullman [Koch 85] c'est une carte de saillance (*saliency map*), pour Wolfe [Wolfe 89] c'est une carte d'activation (*activation map*), pour Fecteau [Fecteau 06] c'est une carte de priorité (*priority map*). Selon les modèles développés, les caractéristiques et le rôle de cette carte centralisée sont variables, mais l'esprit d'une carte centralisée reste commun. Ce principe est séduisant car il permet de représenter dans une même carte le potentiel attentionnel de toute la scène visuelle. Cependant, il n'existe aucune donnée scientifique attestant de manière indiscutable l'existence d'une carte de représentation centrale unique. En effet, selon les études elle serait située dans le colliculus supérieur [Kustov 96], le LGN [Koch 85], V1 [Li 02], V4 [Mazer 02], le pulvinar [Robinson 92], le frontal eye fields [Thompson 05] ou le cortex pariétal[Gottlieb 07].

Puisqu'il n'est pas avéré qu'il existe une représentation centralisée de la saillance dans notre système attentionnel, une alternative est de considérer que cette représentation n'existe pas. Dans ce cas, l'attention est distribuée et devient une propriété émergente de la compétition entre les différents *stimuli* pour obtenir le focus attentionnel. On peut classer dans cette catégorie le modèle théorique de Desimone & Duncan [Desimone 95] ou le modèle computationnel de Deco & Rolls [Deco 04, Rolls 06] que nous décrivons plus loin dans ce chapitre (section 2.3).

Qu'il soit centralisé ou distribué, orienté espace ou objet, *bottom-up* ou *top-down*, un modèle d'attention doit se baser sur un certain nombre d'attributs visuels. C'est en effet à partir de ces attributs qu'il pourra déterminer la singularité de certains *stimuli* et focaliser le système visuel sur ceux-ci. Dans la prochaine sous-section, nous effectuerons un panorama des attributs potentiellement impliqués dans les processus attentionnels, et les classerons par plausibilité.

2.2.4 Les attributs utilisés

Pour sélectionner les zones de notre champ visuel qui devront être traitées en priorité, notre système attentionnel a besoin de se baser sur l'analyse d'un certain nombre d'attributs. Leur connaissance est nécessaire à la construction d'un modèle plausible d'attention visuelle. C'est pourquoi de nombreuses études de neuropsychologie ont tenté de les déterminer.

Dans [Wolfe 00, Wolfe 04], Wolfe effectue une synthèse des différentes recherches du domaine et établit une liste des attributs impliqués dans le déploiement de l'attention visuelle. Le tableau 2.2 présente ce classement. La plausibilité des différents attributs est déterminée en fonction des données expérimentales disponibles.

Indiscutables	Probables	Possibles	Cas douteux	Très peu probables
– Couleur – Mouvement – Orientation – Taille	– Clignotement – Polarité de luminance – Décalage de Vernier [a] – Profondeur stéréo – Indices picturaux de profondeur – Forme – Terminaisons de lignes – Fermeture – Structure topologique – Courbure	– Ombres – Brillance – Expansion – Nombre – Proportions	– Nouveauté – Identité des lettres – Catégorie alphanumérique	– Intersection – Flot optique – Changements de couleur – Volumes 3D – Visages – Nom de la personne – Catégorie sémantique

a. Décalage entre deux séries de traits espacés avec un pas différent. On utilise par exemple ce décalage pour améliorer la précision des pieds à coulisse.

TABLE 2.2: Attributs entrant potentiellement en compte dans le déploiement de l'attention, d'après [Wolfe 04].

Parmi ces attributs, tous n'ont pas la même complexité. On remarque que les attributs basiques, calculés par le système visuel primaire (couleur, orientation, mouvement, polarité de luminance — différence de contraste on-off ou off-on — , clignotement), sont plus probablement liés aux mécanismes attentionnels que des attributs de bien plus haut niveau (visages, nom de la personne, nouveauté, catégorie sémantique, flot optique) dont

la génération, plus complexe, semble peu compatible avec un mécanisme pré-attentif rapide.

2.2.5 Conclusion

Au début de cette section, nous avons présenté les principaux modèles théoriques qui ont contribué à établir les fondements de l'attention visuelle. Bien que loin d'être exhaustif (on pourra consulter [Mole 09] pour une liste plus complète), le sous-ensemble que nous avons choisi permet d'illustrer les principales caractéristiques de l'attention. Nous pouvons les résumer en quatre points :

- notre cerveau ne possède pas de zone unique dédiée à l'attention. Celle-ci est le fruit de l'intéraction entre les différentes aires corticales impliquées dans la perception, le contrôle moteur, ou la planification des actions ;
- l'attention est un mécanisme sélectif, permettant de régler le problème de la correspondance (*binding*) en effectuant un lien spatial entre les différents attributs (non localisés) calculés dans des zones séparées de notre cortex visuel. ;
- l'attention est un phénomène aux multiples visages (tableau 2.3). Sur de nombreux points son interprétation est duale, elle peut être : exogène et automatique ou endogène et contrôlée, dirigée spatialement ou sur des objets, représentée de manière centralisée ou distribuée, déployée de manière ouverte (par le déplacement des yeux) ou couverte (par une focalisation mentale).
- les attributs qui la guident sont multiples : intensité / contraste, couleur, orientation ; mais peuvent être également : forme, fins de lignes, courbure ou visage et catégorie sémantique.

Propriétés	Alternative
Processus mis en œuvre	automatiques ou contrôlés
Déploiement	ouvert ou couvert
Mode d'allocation	spatial ou objet
Source	endogène ou exogène
Représentation interne de la saillance	centralisée ou distribuée

Table 2.3: Les différents visages de l'attention visuelle.

La complexité du phénomène attentionnel ouvre la porte à une multiplicité d'interprétations théoriques que nous avons évoquées tout au long de cette section. A la fin des années 90, la puissance de calcul des ordinateurs devient suffisante pour envisager une mise en œuvre computationnelle de ces multiples modèles. De nouvelles théories continuent à voir le jour, mais elles 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'approche proposés et les différents domaines d'application (étude théorique, vision par ordinateur, traitement d'image, etc.).

2.3 La pratique

Les modèles théoriques présentés dans la section précédente proposent des pistes pour expliquer les mécanismes de l'attention visuelle. Cependant, ils ne fournissent pas suffisamment de détails pour permettre leur mise en œuvre computationnelle directe. Les modèles computationnels apparus à partir de la fin des années 1990 proposent de combler ce fossé. Ils permettent la simulation de différents phénomènes attentionnels et ainsi une comparaison directe avec les propriétés du modèle humain.

Ces modèles ne sont pas cantonnés à l'étude théorique de l'attention, de nombreuses applications en vision par ordinateur, traitement d'images, recherche d'images par le contenu, etc. sont également possibles. Dans chaque type d'application, les contraintes de conception du système attentionnel sont différentes : un modèle dédié à l'étude théorique n'a pas besoin d'être aussi rapide que son homologue dédié à la vision par ordinateur temps réel. Ainsi, pour comprendre les différentes approches de l'attention computationnelle, il est nécessaire de connaître ses applications. C'est l'objet de la sous-section 2.3.1.

L'analyse des différents modèles passe également par la compréhension de leurs propriétés. Une partie de celles-ci, partagée avec les modèles théoriques, a été décrite en sous-section 2.2.3. La sous-section 2.3.2 revient sur leur mise en œuvre et décrit les propriétés spécifiques aux modèles computationnels. Nous pouvons alors aborder la description de ces différents modèles en sous-section 2.3.3. Avant de conclure sur l'adéquation de chaque famille de modèles avec les contraintes d'un système de vision temps réel, nous évoquons en sous-section 2.3.4 la question de l'adaptation dans les modèles computationnels.

2.3.1 Pour quoi faire ?

A quoi peut servir un modèle computationnel d'attention visuelle ? La question est importante, bien qu'un peu provocatrice, car les contraintes issues de l'application imposent des choix dans la conception des modèles attentionnels. Beaucoup d'applications utilisent un modèle d'attention « sur-étagère » (souvent le modèle d'Itti [Itti 98], car il est disponible librement), mais comme nous le verrons en section 2.3.2 avec l'article de Draper [Draper 05] concernant l'invariance des modèles d'attention, il est important de connaître les limites des modèles afin de choisir le plus adapté. Dans cette sous-section, nous décrivons les différentes utilisations des modèles attentionnels et tentons d'en extraire un jeu de contraintes qui servira de grille de lecture lors de notre description des

modèles existants (section 2.3.3).

2.3.1.1 Étude des phénomènes attentionnels

Les premiers modèles computationnels n'avaient pas de vocation autre que de proposer une mise en œuvre informatique permettant de vérifier les propriétés des modèles théoriques développés dans les années 70 et 80. Leur objectif était d'être le plus fidèle au modèle biologique étudié. Leur validation était effectuée en vérifiant que ces systèmes pouvaient reproduire les phénomènes de recherche parallèle (*pop-out*) ou sérielle (*conjonctive search*) observés chez l'humain (figure 2.3.1).

Depuis, les modèles attentionnels sont utilisés pour bien d'autres tâches (que nous décrivons dans la suite de cette sous-section). Mais de nombreux modèles récents et en particulier les modèles d'attention distribués (décrits en sous-section 2.3.3), à vocation neuromimétique, sont encore conçus spécifiquement pour la simulation des différents phénomènes attentionnels observés chez l'homme..

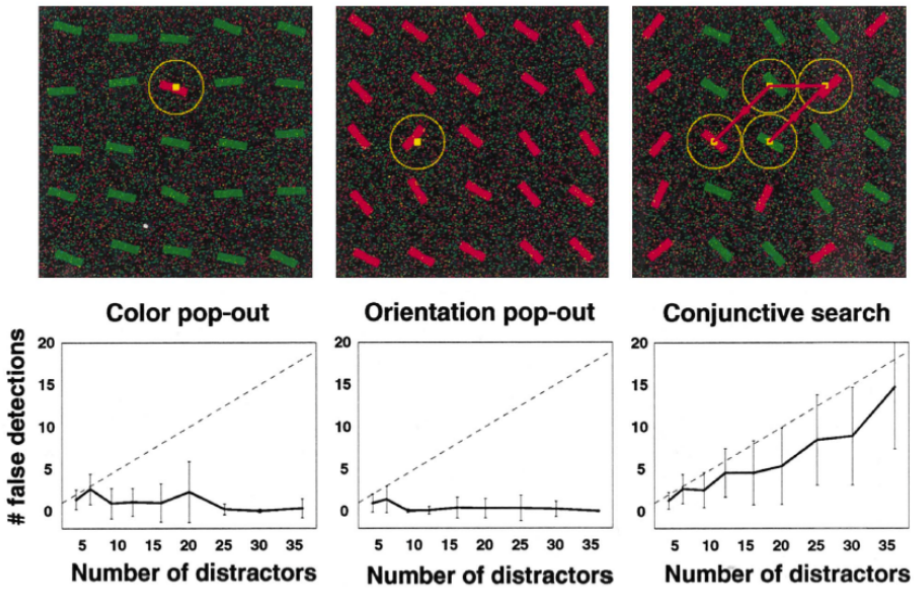


FIGURE 2.3.1: Exemples de phénomènes attentionnels étudiés chez l'homme et reproduits par les modèles computationnels.
Sur les deux images de gauche, les éléments cibles sont trouvés dans un temps constant quel que soit le nombre de distracteur. Sur l'image de droite par contre, le temps de recherche de la barre rouge dont l'orientation est différente des autres augmente en fonction du nombre de distracteurs.

2.3.1.2 Ergonomie / Publicité

L'ergonomie des interfaces homme-machine (IHM) et la publicité sont également deux grands consommateurs d'études sur l'attention. Cependant, celles-ci sont généralement réalisées *via* les techniques du *mouse-tracking* et de l'*eye-tracking* (voir 3.2.1.2) sur un échantillon représentatif d'utilisateurs. Ces études sont relativement lourdes à mettre en place et coûteuses car elle requièrent beaucoup d'interventions humaines et un équipement honéreux.

Depuis peu, des services complètement automatisés d'étude de l'attention visuelle sont proposés, par exemple, par le site Israelien Feng-GUI (`http://www.feng-gui.com/`). Le site annonce que son algorithme est le fruit d'une composition de différents algorithmes d'attention visuelle. En l'absence de publication ou de brevet concernant la méthode utilisée, il est difficile d'évaluer les performances du système et sa validité scientifique. Cependant, le panel des applications proposées est intéressant :

- analyse de l'impact attentionnel de photos, logos, sites-webs et toute autre création visuelle ;
- aide à la définition de l'organisation optimale des éléments sur une page web ;
- sélection de la publicité la plus efficace visuellement sur une page web donnée (figure 2.3.2) ;
- recadrage automatique d'images, basé sur l'attention (voir plus loin les contributions d'Olivier le Meur [Le Meur 06, Meur 07] à ce sujet).

FIGURE 2.3.2: Exemple d'utilisation d'un modèle d'attention visuelle pour sélectionner la publicité la plus adaptée à une page web. Outil proposé par le site Feng-GUI.

Ces solutions semblent être un bon complément à l'étude de l'attention *via* les tech-

niques plus coûteuses d'*eye-tracking*. En l'absence de prise en compte du sens et d'éléments *top-down*, elles ne peuvent cependant pas les remplacer pour l'instant. Ce lien pourrait, par exemple, être fait en combinant un modèle d'attention visuelle avec les travaux de [Chanceaux 09] permettant de simuler un trajet oculomoteur correspondant à une recherche d'information d'une page web, à partir de sa structure et de sa sémantique.

2.3.1.3 Applications intégrant un module attentionnel

Nous abordons ici des applications plus classiques de l'attention en vision par ordinateur et traitement d'images. Dans ce cas, les modèles attentionnels sont utilisés comme un outil, intégré au sein d'un système plus complexe. Des contraintes différentes de celles de la modélisation biologique entrent alors en compte (précision, temps réel, utilisation mémoire, etc.).

Les différents types d'applications ont été hiérarchisés en trois familles :
- la vision, contenant plusieurs sous-familles ;
  - vision active ;
  - reconnaissance d'objets ;
  - détection d'objets ;
  - suivi d'objets ;
  - robotique ;
  - détection de nouveauté ;
- la recherche d'images par le contenu ;
- le traitement d'images, également scindé en plusieurs sous-familles :
  - la segmentation ;
  - le recadrage ;
  - la compression ;
  - le résumé de vidéo.

Dans la suite de cette sous-section, nous décrirons comment les modèles d'attention sont utilisés dans chacune de ces sous-familles.

Vision

Cette première famille d'applications, la vision par ordinateur, est la candidate la plus naturelle à l'utilisation de modèles computationnels d'attention car les tâches à effectuer sont proches de celles réalisées par la vision humaine. Les mécanismes mis en œuvre chez l'homme peuvent alors s'avérer pertinents.

Vision active

Chez l'homme, l'attention visuelle sert à orienter notre regard. Il apparait alors naturel d'utiliser un système computationnel d'attention visuel pour effectuer une tâche

similaire avec le système visuel (caméra(s) motorisée(s)) d'un robot. Les méthodes mises en œuvre sont de complexité variable. Vijayakumar [Vijayakumar 01] utilise un système d'attention relativement simple, basé sur la sélection par un réseau *winnner takes all* (*WTA*) de la zone d'activité la plus importante dans le flot optique des images capturées. Choi [Choi 06] utilise un modèle plus complet, décrit plus en détails en section 2.3.3.2. Dankers [Dankers 07] propose un modèle hiérarchique biologiquement inspiré (les différents modules du système sont calqués avec les différentes aires du cortex visuel primaire) utilisant de nombreuses caractéristiques : intensité, couleur, orientation, flot optique, profondeur et même, possibilité de collision.

Reconnaissance d'objets

La reconnaissance d'objets est également un domaine où l'attention joue un rôle important. Selon les modèles, celle-ci est mise en œuvre de manière très variée. Certains modèles intègrent attention et reconnaissance dans un même processus [Rybak 98, Deco 04, Ji 08]. Une autre approche consiste à guider la reconnaissance grâce à un processus attentionnel externe [Kadir 01, Fay 05, Walther 05, Walther 06a]. Enfin, une dernière approche consiste à doter de capacité d'apprentissage la partie *top-down* d'un système d'attention [Frintrop 05b, Rebhan 08]. Le système attentionnel ainsi obtenu n'effectue pas de reconnaissance d'objet à proprement parler (il n'a pas les même performances), mais il facilite grandement le travail de l'algorithme de reconnaissance qui sera connecté en aval.

Détection d'objets

Variation « subtile » de la reconnaissance d'objets, la détection consiste à trouver des objets appartenant à une classe sémantique de plus haut niveau : on effectuera une reconnaissance du visage de monsieur Dupont, mais on détectera les différents visages d'une scène. Cette détection de visages pourra d'ailleurs être épaulée par un système attentionnel [Rybak 98, Ban 04]. Une autre tâche de détection souvent associée aux modèles attentionnels est celle des panneaux routiers [Bremond 06] (c'est d'ailleurs un des exemples de l'article d'Itti [Itti 98] dès 1998). L'objet de la détection peut également être une cible en mouvement, comme c'est le cas dans [Hu 09]. Ces travaux redéfinissent la notion d'objet saillant dans le cadre de l'analyse de vidéo, comme l'objet que le caméraman souhaite suivre.

Suivi d'objets

Comme nous venons de le voir, les modèles d'attention peuvent également servir à faciliter le travail de modules de suivi d'objets. C'est par exemple le cas dans les travaux de Zhang [Zhang 05] où la saillance de l'élément cible est utilisée comme mesure de similarité d'un modèle de suivi par filtrage particulaire. Dans [Ouerhani 03b], le suivi est effectué simultanément sur plusieurs éléments de la scène. Les zones les plus saillantes

sont détectées grâce à un modèle d'attention, puis suivies d'une trame à l'autre par un module de *tracking* exploitant les informations de saillance.

Robotique et véhicule intelligent

Que ce soit pour la navigation autonome de robots [Aziz 06] ou de véhicule intelligent [Michalke 10], les systèmes attentionnels se doivent d'être rapides. Pour atteindre cet objectif, les modèles proposés sont souvent construits dans l'objectif d'optimiser la tâche de vision cible, et ne s'inspire que très vaguement du modèle d'attention humain [Liu 06b]. Ceci n'est cependant pas systématique puisque Frintrop [Frintrop 05a] propose un modèle à la fois computationellement efficace et biologiquement plausible.

Détection de nouveauté

Une approche naïve permettant de détecter les éléments nouveaux d'une scène consiste à comparer l'image courante avec une version antérieurement mémorisée. Cette approche n'est cependant pas très robuste, car le bruit, les déplacements éventuels de la caméra ou des objets peuvent considérablement perturber ce processus de détection. Pour contourner ce problème, différentes solutions basées sur l'attention visuelle ont été développées. Le modèle attentionnel sert alors de filtre, indiquant les zones importantes de l'image à comparer, réduisant ainsi les problèmes dus au bruit. Pour vérifier que les éléments saillants sont nouveaux, il est alors nécessaire de connecter le système attentionnel à un système de reconnaissance et de mémorisation. Pour ce dernier, différentes options sont proposées dans la littérature : détection de nouveaux clusters pour Gaborski [Gaborski 03], comparaison des chemins formés par les différentes focalisations attentionnelles pour Ban [Ban 06], réseau de neurones *GWR* (*Grow When Required*) ou analyse en composante principale pour Viera Neto [Vieira Neto 07].

Notons que l'approche de Baldi liée à la quantification de la surprise dans des vidéos [Baldi 05] pourrait également être utilisée dans ce cadre.

Recherche d'images par le contenu (*CBIR*)

Les algorithmes de cette seconde famille d'applications utilisent des attributs visuels locaux ou globaux afin de calculer des descripteurs, permettant de définir la similarité entre images.

Lorsque des descripteurs globaux sont utilisés (histogramme couleur de l'image par exemple), la prise en compte de l'ensemble des pixels de l'image pose un problème de segmentation fond / objets d'intérêt. Deux mêmes objets présentés sur des fonds de couleurs différentes risquent fort de ne pas être pris en compte comme similaires. A l'inverse, deux objets différents présentés sur des fonds de couleurs proches seront évalués comme similaires. Bamidele [Bamidele 04] propose d'utiliser un modèle attentionnel pour ne calculer les descripteurs globaux que sur les points des régions les plus saillantes de l'image, limitant ainsi grandement le problème d'influence du fond.

Lorsque des descripteurs locaux sont utilisés, ils sont généralement calculés autour de points d'intérêt déterminés par des méthodes telles que le détecteur de Harris, SIFT, SURF, etc. Ces méthodes ont été créées afin d'extraire des points possédant des propriétés d'invariances à diverses transformations (bruit, échelle, translation, rotation, etc.), mais n'extraient par forcement des points saillants (donc *a priori* distinctifs) de l'image. Kadir [Kadir 01] propose une méthode de calcul de la saillance d'une image et de l'échelle associée aux points les plus saillants, permettant de combiner la stabilité et l'invariance des points d'intérêt avec le pouvoir descriptif, *a priori* plus fort, des zones attirant l'attention.

Traitement d'images

Les applications listées dans cette troisième famille relèvent du traitement d'images au sens large. On y trouve donc des outils, pouvant être utilisés par exemple en vision (segmentation), ou des applications plus liées au multimédia (recadrage, compression et résumé de vidéo).

Segmentation

Les algorithmes de segmentation d'images ou de vidéos utilisent différents critères (couleur, gradient, flot optique, etc.) afin de séparer l'image en différentes zones homogènes. Dans ce cadre, un modèle attentionnel peut être utilisé pour fournir les points de départ (*seeds*) de l'algorithme de segmentation. Ainsi Ouerhani [Ouerhani 03a] combine modèle attentionnel et la méthode du *Seed Region Growing*, alors que Zhang [Zhang 08] utilise les focalisations attentionnelles comme points de départ d'un algorithme de ligne de partage des eaux (*watershed*).

D'autres algorithmes de segmentation basés sur l'attention [Yu 07] n'extraient qu'une seule région, correspondant à la zone la plus saillante de l'image. Une autre approche consiste à effectuer une segmentation fond / éléments de premier plan, en seuillant une carte de saillance orientée régions [Achanta 08]. Cette méthode est également applicable dans le cadre de la vidéo afin de segmenter les objets en mouvement [Maki 00, Lopez 06].

Recadrage

Nous avons vu précédemment que les modèles attentionnels pouvaient être utilisés pour déterminer une région correspondant à la zone la plus intéressante dans l'image. Le Meur [Le Meur 06, Meur 07] exploite ce principe pour recadrer des images en fonction de leur saillance. Celles-ci peuvent alors être affichées de manière plus lisible sur des périphériques d'affichage de petite taille.

Compression

Puisque les modèles attentionnels permettent de hiérarchiser l'importance de chacun des pixels d'une image, on peut utiliser ce classement afin d'adapter localement le taux de compression d'une image [Ouerhani 03a, Lee 05] ou une vidéo [Itti 04] en fonction de sa saillance. On améliore ainsi le taux de compression, tout en garantissant une qualité correcte dans les zones importantes de l'image.

Résumé de vidéo

Le résumé de vidéo est un outil intéressant pour faciliter le parcours de grandes bases multimédia (les campagnes TRECVID 2005 à 2008 y étaient d'ailleurs en partie consacrées). Il est cependant difficile de trouver le bon équilibre entre la réduction de la quantité d'information et le respect du sens. La segmentation sémantique des vidéos étant une tâche particulièrement complexe, des solutions alternatives sont explorées. C'est le cas de la segmentation attentionnelle proposée par Hua [Hua 05]. Dans cette approche, le score attentionnel de chaque trame est évalué en fonction de diverses modalités : visuelle, auditive et en partie sémantique. Les trames à inclure dans le résumé de la vidéo sont alors extraites en sélectionnant les *maxima* de la courbe d'attention ainsi calculée.

2.3.1.4 Bilan

Les applications décrites dans cette sous-section, quoique très variées, utilisent généralement des modèles d'attention adaptés à leurs particularités. Ainsi, la majorité des modèles n'est appliquée qu'à un type de problème, pour lequel il a été conçu. Dans le tableau 2.4, nous proposons une caractérisation des différentes applications en fonction d'un jeu de contraintes que nous nommerons *FAIRED* :

- fidélité au modèle humain ;
- possibilités d'adaptation du système : a-t-on besoin de modifier le comportement du système en fonction du contexte ?
- invariance du modèle à différentes transformations (rotation, translation, changement d'échelle) et / ou répétabilité du système (si celui-ci n'est pas totalement déterministe) ;
- rapidité des calculs nécessaires à la détermination des différentes focalisations ;
- capacités d'extension : est-il nécessaire que le modèle puisse prendre en compte facilement de nouvelles caractéristiques ?
- importance de la gestion de la dynamique de l'attention : faut-il générer des focalisations, ou une simple carte de saillance suffit-elle ?

Ces contraintes et leur impact sur la construction d'un système de vision, serviront de base pour l'analyse des différents modèles présentés en sous-section 2.3.3. Cependant, avant de décrire ces modèles, nous nous intéresserons à leurs propriétés, que nous décrirons dans la sous-section suivante.

	Fidèle	Adaptable	Invariant	Rapide	Extensible	Dynamique
Etude de l'attention	●●●	●	●	●	●	●●●
Ergonomie / publicité	●●●	●	●	●	●●	●●
Vision	●●	●●●	●●	●●●	●●●	●●●
CBIR	●	●	●●●	●●	●●●	●
Traitement d'images	●●	●●	●●	●●	●●●	●●

TABLE 2.4: Contraintes *FAIRED* liées aux différents types d'applications.

2.3.2 Généralités et propriétés

Effectuons maintenant un panorama des propriétés d'une sélection de modèles attentionnels dans l'objectif de mettre à jour leurs caractéristiques communes, ainsi que certaines particularités plus spécifiques à quelques approches originales.

2.3.2.1 Attributs spatiaux ou temporels

Comme nous l'avons vu en section 2.2.4, notre système attentionnel semble utiliser certains attributs visuels au détriment d'autres. Partant de ce constat, la grande majorité des modèles est construite autour de trois caractéristiques principales : l'**intensité**, la **couleur**, et l'**orientation**.

Cependant, ces caractéristiques n'étant pas les seules potentiellement impliquées dans le déploiement de l'attention, certains modèles utilisent des attributs supplémentaires / alternatifs.

Attributs spatiaux

Différentes formes de **symétrie** (radiale, axiale et couleur) sont utilisées par Koostra [Kootstra 08] afin de démontrer qu'un modèle basé uniquement sur cet attribut génère des prédictions d'égale (voire meilleure) qualité que les modèles basés sur le contraste. C'est également le cas du modèle de Sela [Sela 97] où la symétrie est utilisée comme unique caractéristique attentionnelle. Enfin, dans [Aziz 08b] et [Choi 06, Dong 06], elle est utilisée comme attribut complémentaire.

Certains modèles destinés à la robotique prennent également en compte la **profondeur**. C'est le cas de [Ouerhani 03a], où celle-ci est obtenue à partir d'images stéréoscopique. Dans [Frintrop 05a] par contre, cette information est obtenue grâce à un scanner laser 3D.

La **taille** ou l'**excentricité** sont rarement utilisées car leur détection requiert une segmentation en objets. Seuls certains modèles d'attention objet (cf. 2.3.2.4) peuvent intégrer ce type d'attributs [Aziz 08b, Avraham 10]. La même remarque est applicable

à la **forme**, utilisée par le modèle objet de Lopez [Lopez 06].

La détection de **teinte chair** [Rapantzikos 03, Dong 06] ou de **visage** [Ma 03] est également parfois utilisée comme composante *top-down* de certains modèles afin d'améliorer les performances de ceux-ci en présence d'images comportant des personnes / visages. C'est cependant une caractéristique de plus haut niveau, aucun des modèles que nous avons étudiés ne l'intègre dans sa partie *bottom-up*.

Enfin, de manière beaucoup plus anecdotique, les **fins de lignes** sont utilisées par Mozer [Mozer 98], la **courbure** par Milanese [Milanese 94], et les **contours** par [Choi 06, Dong 06]

Attributs spatio-temporels

De nombreux modèles d'attention ne travaillent qu'à partir d'images statiques. D'autres acceptent des séquences d'images ou vidéos. Dans ce dernier cas, le calcul du mouvement peut être effectué de différentes façons.

L'utilisation de simples différences d'images est parfois suffisante. Cette différence peut être faite sur des images au nombre de couleurs réduit afin de limiter le bruit [Lopez 06], sur une pyramide multi-résolutions [Milanese 94] ou encore, dans les modèles d'attention objet, sur les images représentant les différents blobs couleur détectés [Orabona 08].

Plus fréquemment, on calcule le « classique » flot optique *via* une des 3 principales classes d'algorithmes :

- les méthodes de correspondance de blocs : le mouvement est estimé par alignements de blocs (le plus souvent par moindre carré ou corrélation) entre deux trames successives (avec ou sans multi-résolutions). Dans [Le Meur 05b] on calcule également le mouvement dominant de l'image, que l'on soustrait alors au mouvement local afin d'obtenir un mouvement relatif.
- les méthodes basées sur l'énergie [Itti 05a, Belardinelli 09] : on utilise des filtres spatio-temporels orientés dans le domaine de Fourier.
- les méthodes différentielles [Marat 10] ou basées sur le gradient [Rapantzikos 03, Ouerhani 03a] : le flot optique est alors estimé en fonction des dérivées de l'intensité selon le temps et l'espace.

Une alternative est d'intégrer le calcul du mouvement dans un modèle connexionniste neuromimétique dédié à l'attention spatio-temporelle [Tsotsos 05b]. Dans ce cas le mouvement est intégré au modèle et il est difficile de le séparer des autres composantes du système attentionnel.

Une dernière approche dans le traitement du mouvement pour l'attention consiste à utiliser la théorie de l'information afin de faire ressortir les éléments les plus improbables temporellement. Cette méthode ne nécessite pas de calcul explicite du mouvement, c'est

l'écart de l'état actuel à la prédiction basée sur les états temporels passés qui remplace ce calcul. Dans [Mancas 07] cette méthode est appliquée directement sur l'historique des valeurs de chaque pixel de l'image, alors que dans [Bruce 08] une analyse en composantes indépendantes spatio-temporelles est d'abord réalisée sur les données afin de réduire leur dimension.

Quelques modèles sont basés uniquement sur le mouvement [Tsotsos 05b, Belardinelli 09]. Dans ce cas il n'y a pas de problème de conflit entre attributs statiques et spatio-temporels. Cependant, dans la majorité des cas, les modèles intègrent une branche « attention dynamique » en complément du modèle d'attention statique [Milanese 94, Tsotsos 95, Rapantzikos 03, Ouerhani 03a, Mancas 07, Bruce 08, Orabona 08, Bruce 09]. On obtient ainsi deux cartes de saillance $S_{statique}$ et $S_{dynamique}$.

Se pose alors la question du mécanisme à utiliser pour combiner ces deux cartes. Deux solutions sont typiquement utilisées [Bur 07a, Bur 07b, Bur 07c] :

- mélanger les deux types d'attention par une combinaison linéaire du type $S_{globale} = \alpha S_{statique} + (1-\alpha) S_{dynamlque}$, avec $\alpha \in [0;1]$. Attention statique et dynamique sont alors en compétition et α permet de régler la contribution relative de chacun à l'attention globale.
- définir l'attention dynamique comme prioritaire. On a alors :

$$S_{globale} = \begin{cases} S_{dynamique} & si\ \max(S_{dynamique}) > T \\ S_{statique} & sinon \end{cases}$$

  T étant un seuil de basculement vers l'attention dynamique. Cette stratégie peut être appliquée soit globalement (on utilise toute la carte statique ou toute la carte dynamique) ou localement (on choisit la valeur de $S_{statique}$ou de $S_{dynamique}$individuellement pour chaque pixel.

Les résultats obtenus dans [Bur 07a, Bur 07b] semblent indiquer que la stratégie « attention dynamique prioritaire localement » est la plus proche du comportement humain.

2.3.2.2 Propriétés locales ou globales

Comme nous l'avons évoqué en section 2.2.1 la majorité des cellules de la rétine, puis du cortex visuel primaire fonctionne selon le principe « centre-périphérie ». Une comparaison est effectuée entre la réponse au centre d'une cellule (ou d'une groupe de cellules) et son proche voisinage. Cela permet de faire ressortir les différences locales de contraste. De nombreux modèles d'attention se sont inspiré de ce traitement local (un point et son voisinage proche) afin de rendre saillantes les zones de l'image aux fortes différences locales de contraste (ou orientation ou couleur).

Une autre approche, moins justifiable au sens biologique, tout au moins pré-attentivement, consiste à considérer comme saillants les éléments globalement rares. On peut alors classer les modèles d'attention selon leur approche locale ou globale de la saillance

[Mancas 09].

Les modèles de la seconde catégorie sont peu nombreux, mais proposent des approches intéressantes et avec des résultats comparables à l'approche classique. Ainsi [Itti 05a] propose un modèle bayesien permettant de quantifier la surprise dans des vidéos. Mancas [Mancas 07] propose un modèle d'attention basé sur la théorie de l'information, calculant la rareté globale de différentes caractéristiques. Plus récemment, Avraham [Avraham 10] utilise un modèle stochastique afin de calculer la saillance d'une image. Enfin, Torralba [Torralba 06] introduit le calcul de propriétés globales dans la partie top-down de son modèle attentionnel afin de guider l'attention *via* des éléments de contexte (nature de la scène).

2.3.2.3 *Bottom-up*, *Top-down* ou les deux

Le système attentionnel humain est influencé par deux sources d'information : endogène (*bottom-up*) et exogène (*top-down*). Un modèle artificiel d'attention visuelle devrait donc prendre en compte ces deux composantes. Cependant, la partie *top-down* est difficile à appréhender car dépendant, par sa définition, du contexte. C'est certainement pour cette raison que la grande majorité des modèles de la fin des années 1990 et du début des années 2000 ne traitait que de la composante *bottom-up* [Koch 85, Tsotsos 95, Sela 97, Itti 98, Kootstra 08, Itti 00, Kadir 01, Park 02, Bruce 03]. On notera cependant quelques exceptions avec des modèles *bottom-up* et *top-down* très proches de la théorie [Bundesen 87, Desimone 95, Bundesen 98], mais aussi avec des approches plus computationnelles [Ahmad 92, Milanese 94, Mozer 98]. Dans ce dernier cas, l'implémentation du mécanisme *top-down* consiste, soit en la possibilité de modifier globalement [Ahmad 92] ou localement [Mozer 98] le poids de certaines cartes, soit au calcul d'une carte *top-down* à partir d'un système de reconnaissance d'objets [Milanese 94].

De manière assez logique, aucun modèle n'est purement *top-down*, puisque cette influence est considérée comme une modulation contextuelle de l'attention *bottom-up* (guidée par les données). L'ajout d'une composante *top-down* permet donc de créer un système d'attention supervisé, mieux à même de s'adapter à un contexte d'utilisation spécifique. Ainsi, une grande partie des modèles récents propose d'intégrer les deux sources d'attention, la partie *top-down* pouvant être prise en compte en utilisant :

- une modification globale ou locale des poids de chaque caractéristique (intensité, couleur, orientation, etc.) en fonction des propriétés de l'objet à reconnaître. Cela peut être réalisé soit en fournissant directement une ou des cartes de caractéristiques correspondant à la cible [Sun 03, Deco 04, Hamker 05a, Aziz 08a, Orabona 08], ou par un processus d'apprentissage permettant de calculer les poids séparant aux mieux la cible du fond [Frintrop 05b, Choi 06].
- une carte *top-down* issue de reconnaissance de visage [Ma 03, Dong 06], ou de teinte chair [Rapantzikos 03].
- une carte *top-down* issue d'une procédure de catégorisation de scène [Navalpakkam 05b,

Torralba 06].
– une carte *top-down* créée à partir de la moyenne des observations de différents sujets *via* un procédé d'eye-tracking ou de mouse-tracking [Mancas 09].

2.3.2.4 Attention spatiale ou objet

Tout comme de nombreux modèles théoriques d'attention, la majorité des modèles computationnels est basée sur une représentation spatiale. En effet, celle-ci est plus simple à mettre en œuvre et est plus cohérente avec le rôle « bas-niveaux » de l'attention visuelle. Cependant, comme nous l'avons vu en sous-section 2.2.3, certaines théories (en particulier la théorie de la forme, Annexe C) suggèrent que l'attention pourrait jouer un rôle dans les phénomènes de groupement perceptuel (*perceptual grouping*) permettant à notre système visuel de résoudre le problème de la segmentation figure / fond (*figure / ground segmentation*). Dans le cadre de la vision par ordinateur, l'attention objet permet une sélection et détection des objets bien plus aisées mais nécessite une phase de segmentation, difficilement plausible biologiquement, car intervenant très tôt dans le traitement visuel. La plupart des modèles objets sont donc des modèles computationnels, revendiquant peu une « réalité biologique » [Lopez 06, Geerinck 09, Aziz 09a, Avraham 10]. On trouve également des modèles objets biologiquement inspirés [Sun 03, Sun 08] mais ceux-ci sont plus rares et l'on préfère plutôt utiliser la notion de proto-objet [2] plus compatible avec les modèles théoriques [Walther 06b, Orabona 08].

2.3.2.5 Gestion explicite du focus d'attention

Lorsqu'il déploie son attention ouverte, le système visuel humain parcourt la scène en effectuant différentes fixations et saccades oculaires. Beaucoup de systèmes computationnels fonctionnent sur le même principe : ils génèrent un ensemble de fixations, permettant d'explorer les images qui lui sont présentées.

Cependant, il n'est pas toujours nécessaire de déterminer ces différentes fixations. Un modèle d'attention peut uniquement être utilisé pour générer une carte de saillance, donnant l'importance relative de chacun des points de l'image. C'est le cas par exemple des modèles de Le Meur [Le Meur 05a] ou Mancas [Mancas 07].

Les deux méthodes sont justifiables et cohérentes :
– dans un cas, la fidélité biologique est privilégiée. En produisant (directement ou non) les fixations, on génère un type de données similaire à celui mesuré lors des expériences oculométriques. C'est également la manière la plus simple de procéder lorsque l'on utilise un modèle non centralisé (connexionniste ou à compétition biaisée).

2. Ensemble de pixels ou caractéristiques regroupés avant le processus attentionnel, et pouvant être « transformé » en objet après le processus attentionnel [Rensink 00].

– dans l'autre cas, la carte de saillance est souvent suffisante pour effectuer des comparaisons avec la vision biologique. En effet, ces comparaisons sont généralement effectuées en mesurant la similarité entre une carte de saillance ou une *heatmap* générée à partir des fixations du système artificiel, et une *heatmap* générée à partir d'expérimentations d'*eye-tracking* (cf. section 3.2.1.2). On évite ainsi une phase de conversion des fixations en *heatmap*. De plus, un ensemble de fixations peut toujours être généré à partir d'une carte de saillance en utilisant un réseau *Winner Takes All* et un mécanisme d'inhibition de retour (c'est le procédé utilisé par exemple par [Itti 98] ou [Frintrop 05a].

2.3.2.6 Invariance des modèles d'attention

La plupart des modèles d'attention ne sont pas conçus pour être invariants aux transformations appliquées aux images qui leur sont fournies (translation, rotation, changement d'échelle). Cela peut être dû aux modèles (reproduction de la plus forte sensibilité humaine à certaines orientations par exemple) ou à son implémentation (choix des filtres et échelles de traitement, etc.).

Si l'objectif du modèle est la simulation de l'attention visuelle humaine, cela ne pose aucun problème. On peut difficilement démontrer que les fixations humaines sont invariantes aux translations ou rotations (différentes expériences d'*eye-tracking* sur une même image et un même sujet, avec les mêmes instructions donnent lieu à des trajectoires oculaires différentes). Cela n'empêche aucunement que notre perception soit invariante à ces transformations.

Cependant, dans le cadre de l'utilisation des modèles d'attention dans certaines applications de vision par ordinateur, cette non-invariance peut être problématique. Il convient alors de pouvoir l'évaluer afin d'en mesurer les effets et adapter l'algorithme ou en choisir un différent. Dans [Draper 05], les auteurs montrent que le *Neuromorphic Vision Toolkit* de Laurent Itti [Itti 98] n'est pas invariant et que des modifications dans son implémentation permettent de résoudre en grande partie ce « défaut ». L'évaluation est malheureusement limitée à ce seul algorithme, mais on peut imaginer que de nombreux autres modèles souffrent du même « problème ».

Il ne faut cependant pas se focaliser sur cet aspect, car rien ne spécifie que d'une manière générale un système attentionnel doit être invariant aux différentes transformations. Il ne sera nécessaire d'étudier cet aspect que dans le cas où l'application cible nécessite cette invariance (dans [Draper 05], celle-ci était un système de reconnaissance basé sur l'apparence).

2.3.2.7 Bilan

Dans cette section nous avons vu que les différents modèles d'attention proposaient une grande variété de solutions pour résoudre un même problème. Certaines propriétés sont communes, notamment les attributs pris en compte (généralement au moins l'intensité, la couleur et l'orientation). Mais on retrouve également la dualité quasi systématique constatée lors de la description des propriétés des modèles théoriques d'attention (sous-section 2.2.3) : cheminement *bottom-up* / *top-down*, orientation spatiale / objet, propriétés locales / globales, etc.

En conséquence, les angles d'attaque pour effectuer une taxonomie de ces différents algorithmes sont nombreux. On peut cependant séparer les modèles computationnels en deux grandes familles [Hannagan 06] : les modèles distribués et les modèles centralisés. Cette catégorisation, que nous abordons dans la prochaine sous-section, est intéressante car elle s'applique également aux modèles théoriques sous-jacents, ce qui ne serait pas le cas d'une classification basée sur des aspects plus computationnels.

2.3.3 Deux grandes familles de modèles

Nous avons choisi de séparer la présentation des différents modèles computationnels d'attention en deux familles, basées sur des concepts duaux. Les partisans des modèles d'attention distribuée considèrent que l'attention est une propriété émergente de la compétition biaisée (par l'évolution, l'apprentissage ou le contexte) entre les différents *stimuli* visuels. L'attention ne serait pas spécifiquement codée dans une carte topographique. Les partisans des modèles centralisés pensent eux qu'au contraire, l'attention est codée dans une carte topographique 2D qui sert de référence pour l'allocation de l'attention *via* différents mécanismes (*winner takes all*, inhibition de retours, etc.). Dans cette sous-section, nous effectuerons un panorama (non exhaustif) des différents modèles existants dans ces deux familles.

2.3.3.1 Attention distribuée : modèles basés sur la compétition biaisée

Fondements

Les modèles d'attention distribuée trouvent leurs racines dans les écoles neuroscientifique et connexionniste. Ils sont généralement conçus dans une approche neuromimétique. Leur niveau de granularité peut aller jusqu'à celui des neurones, ceux-ci étant parfois même simulés très précisément[Rolls 06]. L'étude expérimentale de ces modèles peut alors se rapprocher de celle du cerveau humain, en permettant la mesure des signaux individuels de chaque cellule.

La majorité de ces travaux sont inspirés du modèle de compétition biaisée proposé par Desimone et Duncan [Desimone 95] au milieu des années 1990. Celui-ci peut être résumé de la façon suivante :

- L'attention n'est pas un faisceau mental (tel que décrit par Treisman) parcourant les éléments de la scène visuelle à haute vitesse. C'est au contraire une propriété émergente d'une compétition lente, issue des interactions entre les traitements visuels parallèles effectués sur tout le champ visuel.
- Les objets dans le champ visuel sont en compétition pour l'allocation des ressources cognitives limitées et le contrôle moteur de l'attention.
- Cette compétition est biaisée par des mécanismes *bottom-up* permettant de séparer les objets de leur fond, et par des mécanismes *top-dow*n permettant de sélectionner les objets les plus adaptés à la tâche / au contexte courant(e). Ce biais peut être contrôlé par la position et / ou les différentes caractéristiques des objets de la scène.

Les modèles

Nous allons maintenant aborder deux sous-classes de modèles : ceux étudiant la réponse isolée de certains neurones et ne faisant que peu état du comportement macroscopique du système ; et les modèles plus computationnels, permettant une utilisation en condition réelle, sur des images naturelles.

A la limite de la théorie

Le modèle proposé par Bundesen [Bundesen 87, Bundesen 98] est basé sur une compétition temporelle entre les différents éléments de la scène visuelle. Ceux-ci sont traités en parallèle par le système visuel afin d'être catégorisés et mis en mémoire à court terme. L'élément dont la catégorisation et la mémorisation se termine en premier sera celui qui recevra le focus d'attention.

Sur ce même principe de la compétition temporelle, [VanRullen 03, VanRullen 04] propose que l'attention n'est pas un processus à part, mais qu'elle est liée à l'ordre des impulsions neuronales lors du processus de reconnaissance des différents éléments de la scène. Ainsi lors de l'analyse d'une scène, la saillance serait représentée par le premier train d'impulsions généré par les différents neurones du système visuel. Celui-ci permettrait par exemple de catégoriser une scène très rapidement (moins de 150ms), à partir de ses éléments les plus saillants. Les impulsions suivantes serviraient alors à moduler cette réponse initiale, en fonction d'une influence *top-down*. Le modèle proposé par VanRullen n'est pas directement un modèle d'attention, mais il fournit une explication des phénomènes attentionnels au sein d'un mécanisme plus large de reconnaissance d'objets, architecturé autour de neurones encodant l'information grâce à l'ordre de leurs impulsions (*rank order coding*).

Pour Spratling [Spratling 04], la compétition biaisée entre les différents *stimuli* visuels est effectuée grâce au rebouclage (*feedback*) entre les neurones des différentes régions corticales. L'information ne suit pas un parcours linéaire (*feedforward*) à travers les

différentes couches neuronales : les interactions entre les différentes couches permettent une compétition amenant à une sélection des *stimuli* les plus saillants.

Modèles plus computationnels

Le modèle de *selective tuning* proposé par Tsotsos [Tsotsos 95] en 1995 puis étendu plus récemment au mouvement [Tsotsos 05b] permet d'effectuer une sélection attentionnelle sur un réseau de neurones pyramidal de traitement visuel. Ce modèle procède en deux temps :

- l'information visuelle suit son chemin dans les différentes couches du réseau de neurones jusqu'à atteindre la couche de plus haut niveau ;

- un mécanisme de *winner takes all* (*WTA*) sélectionne sur cette couche l'unité ayant la plus forte réponse et inhibe les autres. Le processus se poursuit ensuite sur les unités de la couche inférieure reliées à l'unité ayant gagné le *WTA*, et est répété ainsi pour chacune des couches de la pyramide.

A la fin de ce processus de sélection *top-down* (on procède du haut vers le bas de la pyramide), on a localisé précisément la source visuelle qui avait générée une forte réponse au sommet de la pyramide. Les neurones ne participant pas au traitement du *stimulus* visuel le plus saillant sont désactivés.

Le modèle présente essentiellement les mécanismes de sélection attentionnelle, laissant un grand choix dans la construction de la pyramide d'analyse visuelle. Différents exemples sont donnés afin d'obtenir des modèles d'attention basés sur la luminance, l'orientation de lignes ou le flot optique.

Deco, Stringer et Rolls [Deco 04, Stringer 00] poussent l'intégration des systèmes de reconnaissance d'objet et d'attention beaucoup plus loin (figure 2.3.3). Tout comme Ji et Weng [Ji 08], ils proposent un modèle de traitement visuel découpé selon deux voies : la voie « *what* » (dorsale) est chargé du processus de reconnaissance ; la voie « *where* » (ventrale) interagit avec celle-ci afin de prendre en compte l'attention spatiale. La modélisation du système visuel ainsi créé est très fine : les différentes aires corticales sont représentées et leurs interactions modélisées.

Le processus d'attention visuelle, basé sur la compétition biaisée, est ainsi mis en œuvre à deux niveaux : au niveau microscopique par un mécanisme d'inhibition locale des neurones voisins (*lateral inhibition*) ; au niveau macroscopique, par l'interaction des modules correspondant aux différentes aires du cortex visuel (V1, V2, V4, IT).

Résumé

Les modèles distribués basés sur la compétition biaisée permettent de modéliser très finement l'attention visuelle. Leur complexité est cependant à la hauteur de leur fidélité biologique. Simuler les interactions complexes des différentes aires corticales demande des ressources importantes, rendant l'utilisation temps réel de tels modèles (sur un or-

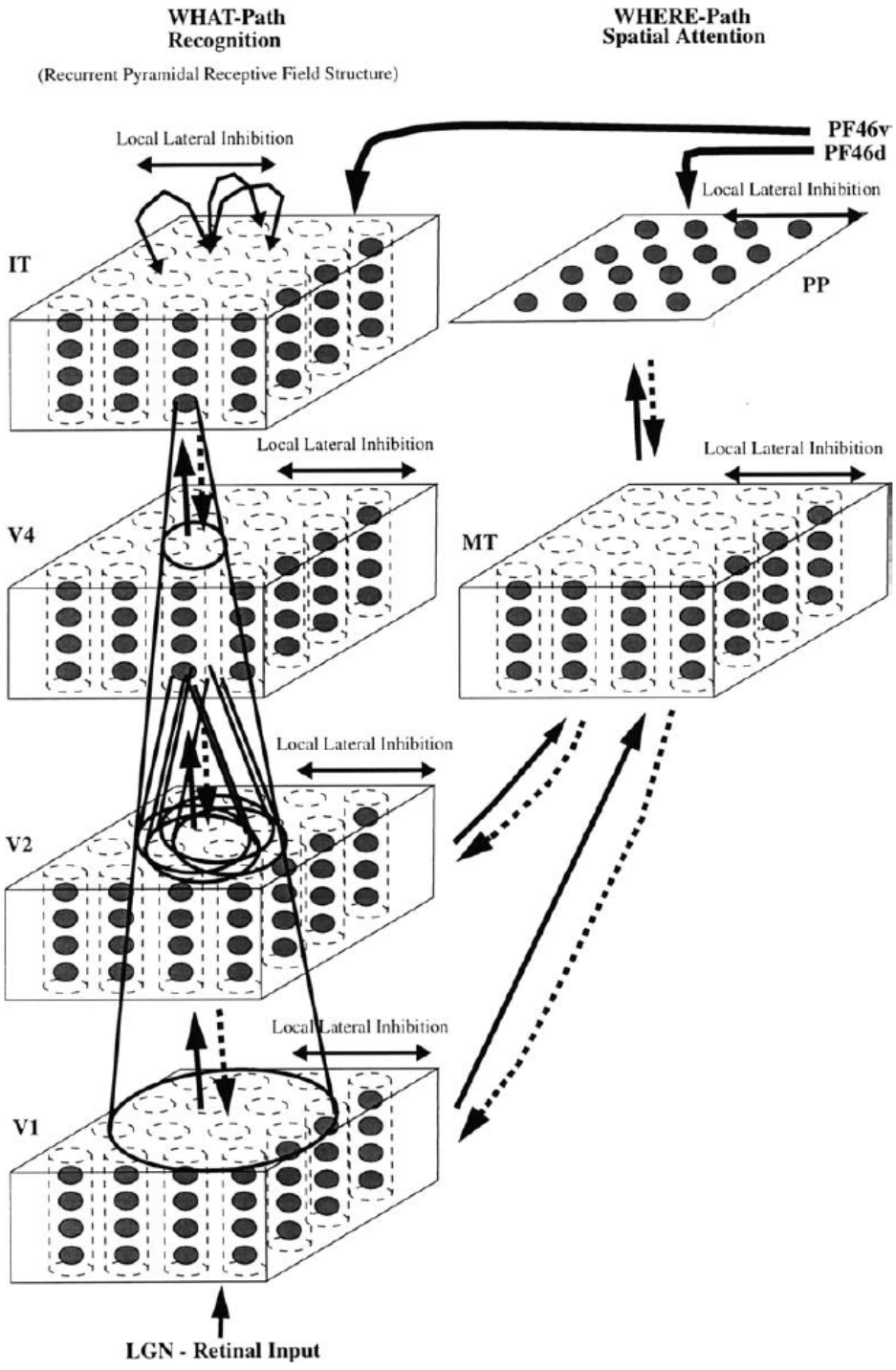


FIGURE 2.3.3: Le modèle distribué de Deco, Rolls et Stringer. D'après [Deco 04]

dinateur standard) peu crédible pour l'instant. Une approche compétitive de l'attention à un niveau de granularité plus grossier (supérieur à celui du neurone) pourrait permettre de bénéficier des avantages de la compétition (résolution efficace du problème de la sélection d'information) de manière plus adaptées à la vision par ordinateur.

L'attention peut également être modélisée de manière moins compétitive et plus centralisée. Nous abordons les modèles utilisant ce paradigme dans la suite de cette section.

2.3.3.2 Attention centralisée : modèles à carte de saillance

Les modèles d'attention centralisée se situent dans la continuité des travaux précurseurs d'Anne Treisman [Treisman 80]. Selon la *Feature-Integration Theory*, l'attention est codée dans une carte centrale interne (dont le nom varie selon les théories : carte maîtresse, carte de saillance, etc.) permettant une représentation du champ visuel. Bien que des études plus récentes n'aient encore pas réussi à prouver l'unicité de représentation de la saillance dans notre cerveau, ce modèle est très populaire car il propose une explication simple, computationellement efficace, et au pouvoir explicatif avéré.

L'hypothèse de l'attention centralisée étant plus populaire que sa version distribuée, le nombre de modèles basés sur ce paradigme est important. Nous proposons une taxonomie en 5 familles, composée des modèles (figure 2.3.4) :

- hiérarchiques ;
- statistiques et probabilistes ;
- basés sur la théorie de l'information ;
- connexionnistes ;
- algorithmiques.

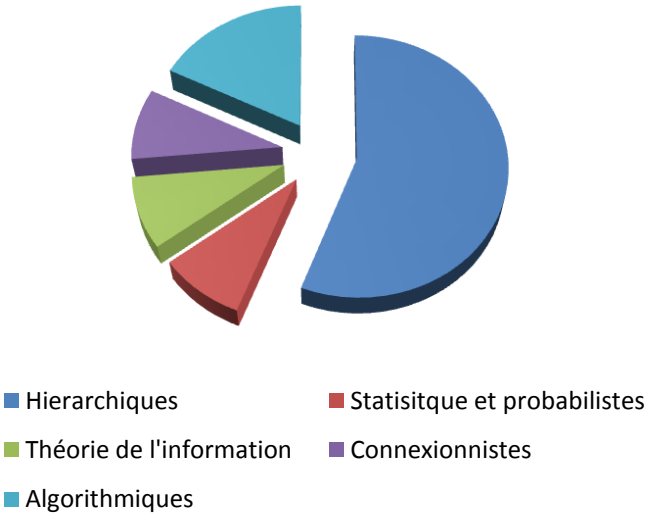


FIGURE 2.3.4: Répartition des algorithmes étudiés dans cette sous-section.

D'autre classements existent [Tsotsos 07, Le Meur 09], d'autant plus que certains algorithmes font appel à différentes méthodes et qu'il est difficile de les placer dans une seule famille. La taxonomie proposée a pour principal avantage de bien distinguer les différents types d'approches de l'attention centralisée.

Hiérarchiques

Ce type d'algorithme construit, à partir d'une image initale, une hiérarchie de différentes cartes de caractéristiques, qui seront progressivement combinées jusqu'à obtenir une représentation centrale unique : la carte de saillance. L'un des modèles les plus influents de cette classe d'algorithmes est le modèle de Laurent Itti [Itti 98, Itti 00]. Celui-ci doit sa grande popularité à différents facteurs :

- c'est un des premiers modèles computationnels d'attention ;
- il est basé sur des théories attentionnelles influentes [Treisman 80, Koch 85] ;
- son architecture, biologiquement inspirée, est simple et efficace ;
- son implémentation (code source et exécutables) est disponible librement à travers le *Neuromorphic Visionl Toolkit*[3] (*iNVT*), ce qui permet de s'en servir comme base pour d'autres modèles, ou d'effectuer facilement des comparaisons avec celui-ci ;
- il est en constante amélioration depuis sa création.

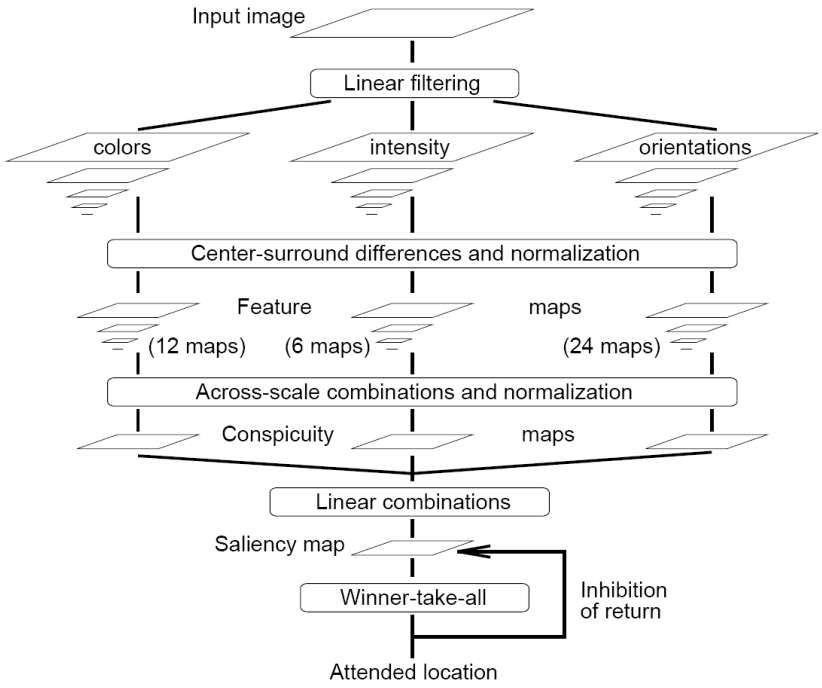


FIGURE 2.3.5: Le modèle hiérarchique centralisé de Laurent Itti[Itti 98]

3. Disponible à l'adresse suivante : `http://ilab.usc.edu/toolkit/`.

Le modèle étant décrit en détail dans de nombreuses thèses et publications, nous ne décrirons ici que ses principes généraux. L'architecture de ce modèle *bottom-up* (figure 2.3.5) est basée sur les étapes suivantes. L'image source est décomposée en différents canaux perceptuels (dans la version de base de l'algorithme : intensité, couleur et orientation). Puis, une représentation multi-échelles est construite à partir de ces différents canaux. Une opération de filtrage centre-périphérie est alors effectuée par soustraction de certains niveaux des pyramides multi-échelles afin d'obtenir différentes cartes de caractéristiques. Celles-ci sont ensuite normalisées par un opérateur $\mathcal{N}$ permettant de renforcer les cartes ne contenant qu'un petit nombre de pics d'activité (zones saillantes) puis sommées afin d'obtenir 3 cartes de singularité (intensité, couleur, orientation). Ces cartes sont également normalisées avec l'opérateur$\mathcal{N}$, puis sommées afin d'obtenir une carte de saillance. La génération des différentes focalisations est effectuée grâce à un réseau *Winner Takes All* (*WTA*), sélectionnant la zone d'activité maximale de la carte de saillance, couplé à un mécanisme d'inhibition de retour, désactivant temporairement les zones déjà visitées afin que le focus d'attention n'y revienne pas immédiatement.

Ce modèle en a inspiré d'autres, comme par exemple l'implémentation temps réel de Ouerhani [Ouerhani 03a] qui a exploré l'utilisation de puces SIMD [4] dédiées au traitement d'images dans le cadre de la modélisation de l'attention visuelle. Rapantzikos [Rapantzikos 03] propose quant à lui une extension du modèle original par l'ajout du mouvement (également pris en compte par les évolutions plus récentes de [Itti 98]) et de la détection de teinte chair. Koostra [Kootstra 08] propose un des premiers modèles d'attention entièrement basé sur la symétrie. Choi [Choi 06] et Dong [Dong 06] ajoutent également la prise en compte de la symétrie, et proposent une extension *top-down,* basée sur les réseaux Fuzzy-ART [5], afin que le modèle puisse apprendre à inhiber ou renforcer certaines zones du champ visuel en fonction du contexte. Enfin, Walther [Walther 06b] étend le modèle d'Itti en proposant un mécanisme permettant de déterminer le proto-objet correspondant à chaque nouveau focus attentionnel.

Nous souhaitons insister particulièrement sur deux modèles dérivés de [Itti 98] car ils apportent des améliorations intéressantes et complémentaires :

- Le modèle d'Olivier Le Meur [Le Meur 05a] (figure 2.3.6) pousse l'inspiration biologique bien plus loin que le modèle original. Sans utiliser d'approche neuromimétique telle que celle utilisée dans les modèles distribués, il modélise finement de nombreux phénomènes psycho-visuels : fonctions de sensibilité aux contrastes, décomposition en canaux perceptuels complexes, phénomène de masquage inter et intra cartes, etc. Cette approche montre qu'il est possible d'obtenir une modélisation très fine avec une approche plutôt macroscopique (le comportement individuel des neurones du cortex visuel n'étant pas du tout abordé).
- Le modèle de Simone Frintrop [Frintrop 05a] utilise une approche plus orientée « vi-

4. *Single Instruction Multiple Data*
5. *Adaptive Resonance Theory*

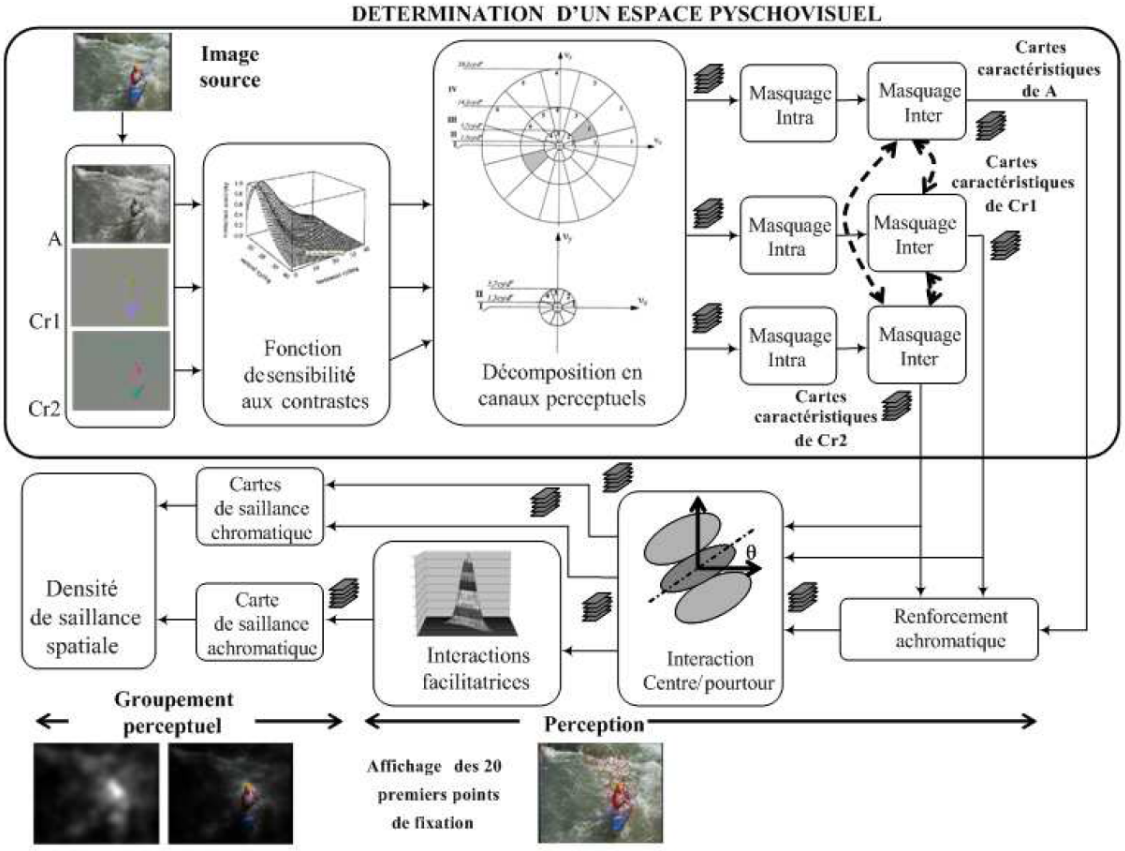


FIGURE 2.3.6: Le modèle hierarchique d'attention de Le Meur [Le Meur 05a].

sion par ordinateur ». Il étend le modèle d'Itti sur différents points : séparation des cartes de caractéristiques On/Off et Off/On pour mieux prendre en compte certains effets de *pop-out*, utilisation d'un espace couleur psychovisuel, modélisation plus précise des filtres centre-périphérie, etc. De plus, ces modifications sont effectuées en prenant en compte la rapidité d'exécution [Frintrop 07], prouvant qu'il est possible d'obtenir un modèle d'attention computationnellement efficace au comportement plausible.

La famille des modèles hiérarchiques contient également de nombreux autres modèles. C'est par exemple le cas du modèle de Milanese [Milanese 94]. Bien que celui-ci soit un des premiers modèles computationnels d'attention visuelle, il intègre un ensemble relativement complet de fonctionnalités, non présentes dans des modèles plus récents. Il est ainsi composé :

- d'un système *bottom-up* hiérarchique calculant des cartes de caractéristiques et de singularité dont la fusion en une carte de saillance est assurée par un algorithme de relaxation ;
- d'une gestion du mouvement sous la forme d'un système d'alerte (*alerting subsystem*) produisant une carte d'alerte ;
- d'un système *top-down* mis en œuvre par une mémoire associative distribuée (*distributed associative memory*) permettant d'apprendre certains objets.

Certains modèles hiérarchiques récents sont basés sur une attention objet plutôt que spatiale. Parmi ceux-ci on peut citer les modèles de Sun [Sun 03, Sun 08], Liu [Liu 06a], Achanta [Achanta 08] et Geerinck [Geerinck 09]. Ces travaux diffèrent sur la façon de segmenter les objets ou régions (segmentation couleur, *watershed* sur l'intensité du gradient, etc.), mais ont tous le même objectif : proposer une focalisation sur des régions, plutôt que sur des points.

Enfin, d'autres modèles sont plus orientés sur le traitement du mouvement. Belardinelli [Belardinelli 09] propose une approche basée uniquement sur celui-ci, utilisant le filtrage spatio-temporel. Celui-ci est réalisé par une combinaison de filtres de Gabor 2D ($x+t$ et $y+t$). Une autre approche, proposée par Marat [Marat 10], utilise une rétine artificielle générant des signaux différents aux voies traitant l'information statique ou dynamique. Une fois traitées, ces informations sont fusionnées de manière adaptative afin d'obtenir une carte de saillance spatio-temporelle unique.

Statistiques et Probabilistes

La saillance d'un objet pouvant être définie comme sa singularité par rapport aux autres, il est assez naturel d'utiliser une théorie probabiliste ou statistique afin de relier la saillance aux éléments les moins probables d'une scène. Dans ce cadre, différentes approches sont possibles.

Park [Park 02] n'utilise pas de filtrage centre-périphérie classique, mais génère les dif-

férents filtres permettant de calculer la carte de saillance *via* une analyse en composantes indépendantes (*ICA*).

Hamker [Hamker 05b, Hamker 05a] utilise la théorie de l'inférence basée sur les populations (dérivée de l'inférence bayésienne) afin de moduler la saillance *bottom-up* par une connaissance *top-down*, modélisant la saillance « attendue » sur un type d'image donné (figure 2.3.7).

Egalement pour un contrôle *top-down*, [Torralba 06] utilise l'inférence bayesienne sur des caractéristiques globales de l'image afin de biaiser l'estimation *bottom-up* de la saillance.

La théorie bayesienne est également utilisée par Itti et Baldi [Itti 05a, Baldi 05] pour proposer non pas un modèle d'attention, mais une théorie de la surprise, prédisant de manière tout à fait convaincante les fixations oculaires de sujets regardant des vidéos.

Enfin, Avraham [Avraham 10] utilise la théorie bayesienne pour calculer la probabilité des différentes régions d'une image d'être un élément saillant de la scène (segmentée auparavant).

Théorie de l'information

En lien direct avec les théories probabilistes, les modèles basés sur la théorie de l'information postulent que notre cerveau utilise les mécanismes attentionnels afin de maximiser la quantité d'information acquise. Estimée localement, celle-ci peut alors servir à définir la saillance d'une image. Différentes approches du calcul de la quantité d'information sont possibles.

Gilles [Gilles 96] propose une explication de la saillance en terme de complexité locale, pouvant être mesurée par l'entropie de Shannon des attributs locaux de l'image. Kadir [Kadir 01] reprend cette définition et étend le modèle en utilisant le maximum d'entropie pour déterminer l'échelle des éléments saillants dans une analyse multi-échelles.

Bruce [Bruce 03] propose d'utiliser une mesure de l'information propre (*self-information*) afin de construire des opérateurs de filtrage non-linéaire, utilisés pour normaliser des cartes de singularité avant leur fusion, dans une architecture proche de celle proposée par Itti [Itti 98]. Il réutilise ce principe dans ses travaux avec Tsotsos [Bruce 09, Bruce 08], combinant analyse en composantes indépendantes (*ICA*) [Park 02] et mesure de l'information propre afin d'obtenir une estimation de la saillance d'une image (figure 2.3.8).

Enfin, Mancas [Mancas 07] propose une approche très complète de la saillance basée sur l'information propre. Il propose des modèles adaptés à différentes modalités : 1D (son), 2D (images) et 2D+t (vidéo). Son approche couvre également l'attention sans *a priori* (*bottom-up*) ou avec *a priori* (*top-down*).

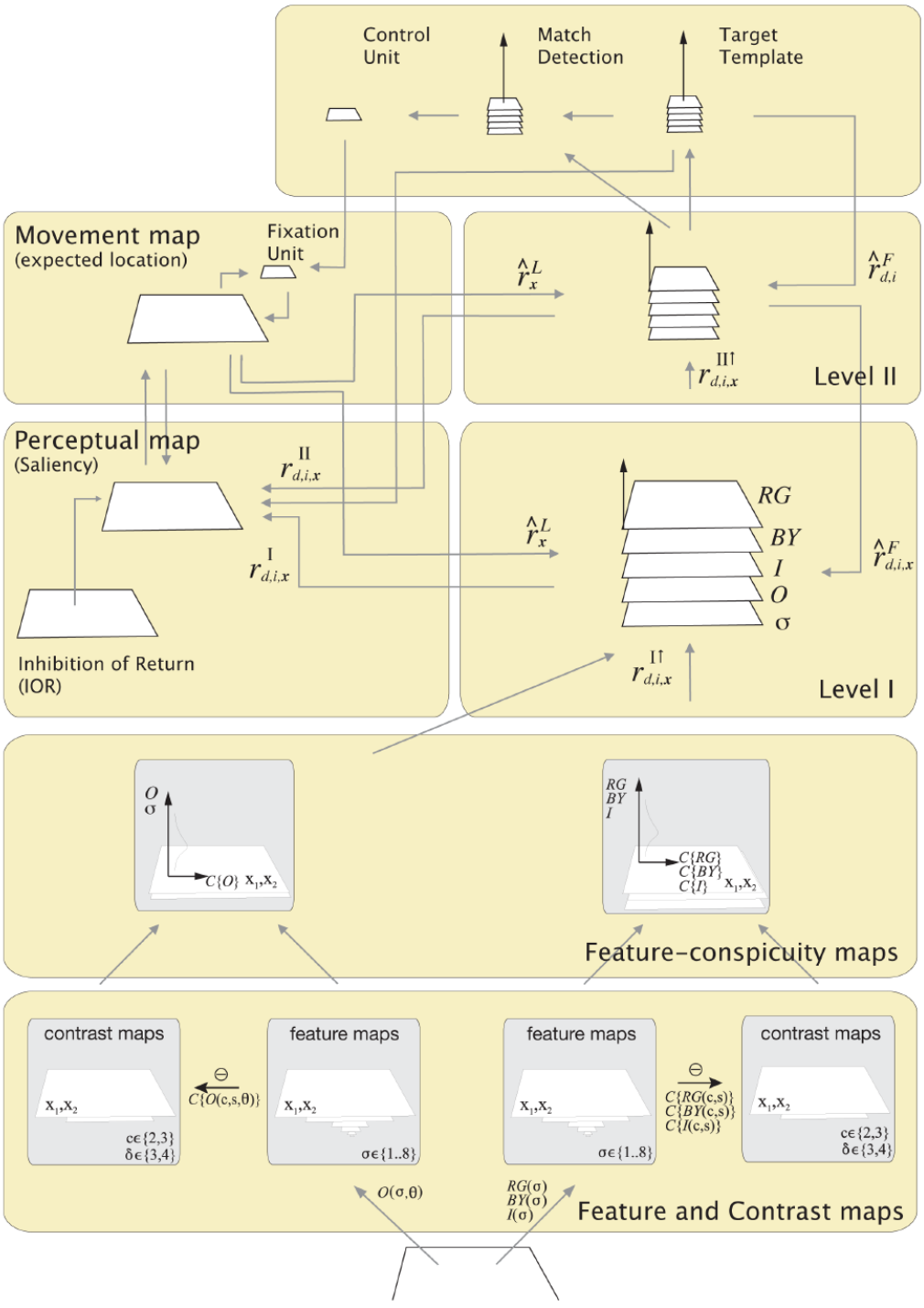


FIGURE 2.3.7: Le modèle de Hamker [Hamker 05b].

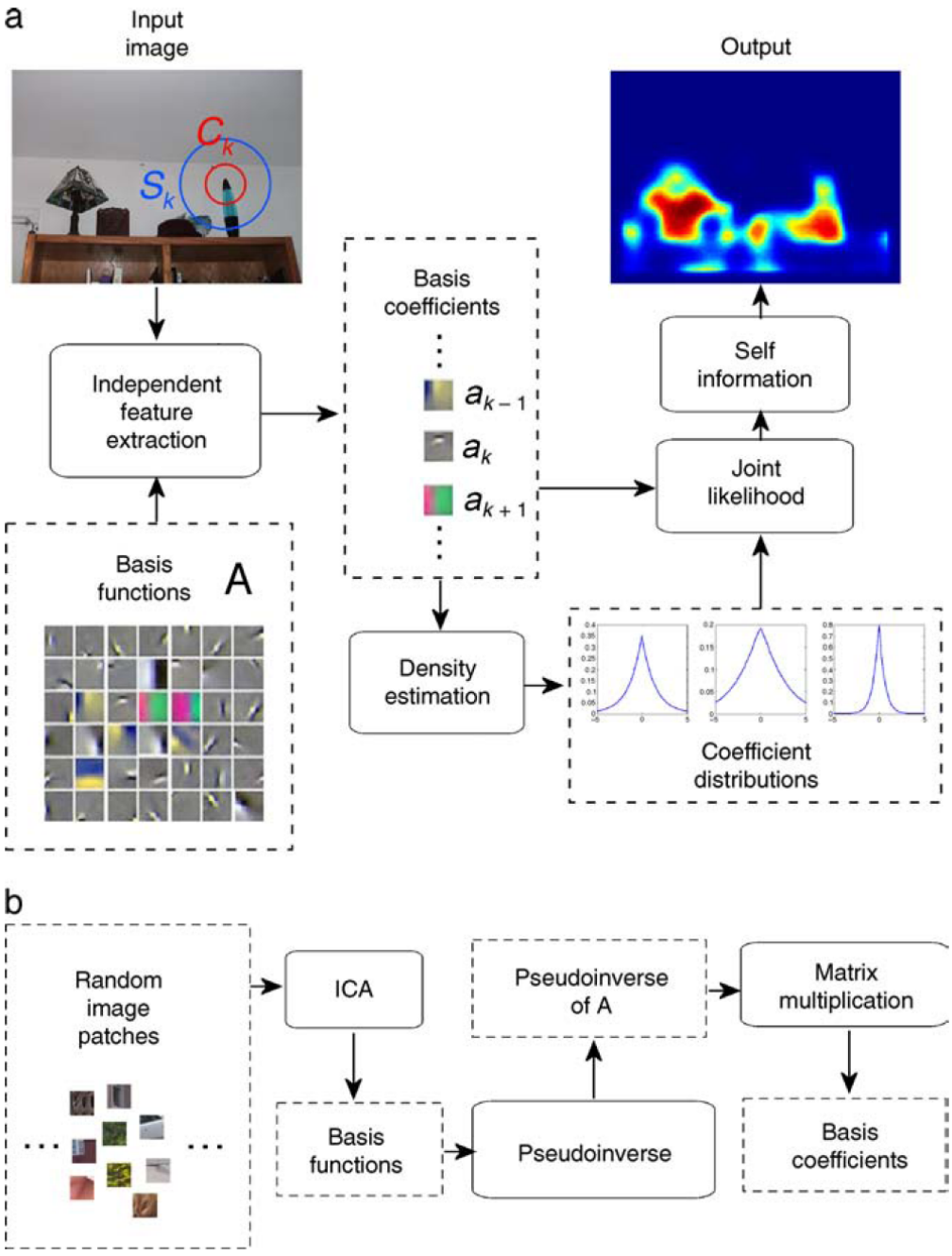


FIGURE 2.3.8: L'approche, basée théorie de l'information, de Bruce et Tstotsos [Bruce 09].

Connexionnistes

Les modèles de cette catégorie utilisent principalement des réseaux de neurones comme mécanisme attentionnel. Bien que basés sur la compétition, ces modèles font partie des algorithmes centralisés car ils utilisent une carte de saillance. Cependant, ils ne permettent pas de construire la carte saillance mais utilisent celle-ci afin de générer un ensemble de focalisations attentionnelles ou de moduler directement les traitements d'un processus de reconnaissance.

L'un des premiers modèles connexionnistes utilisant une représentation centralisée de la saillance est le modèle VISIT [6] de Ahmad [Ahmad 92]. Celui-ci découpe le traitement attentionnel en trois réseaux indépendants, connectés à une mémoire de travail (figure 2.3.9). Le réseau de blocage (*gating network*) supprime toute activité non située dans une région déterminée (le focus d'attention). Le réseau de priorité (*priority network*) sélectionne les zones d'intérêt en utilisant des informations *bottom-up* et/ou *top-down*. Ces informations doivent être fournies *via* une carte centrale appelée carte de priorités (*priority map*). Le réseau de contrôle (*control network*) effectue le lien entre les deux réseaux précités ; il détermine et séquence les changements de focus d'attention.

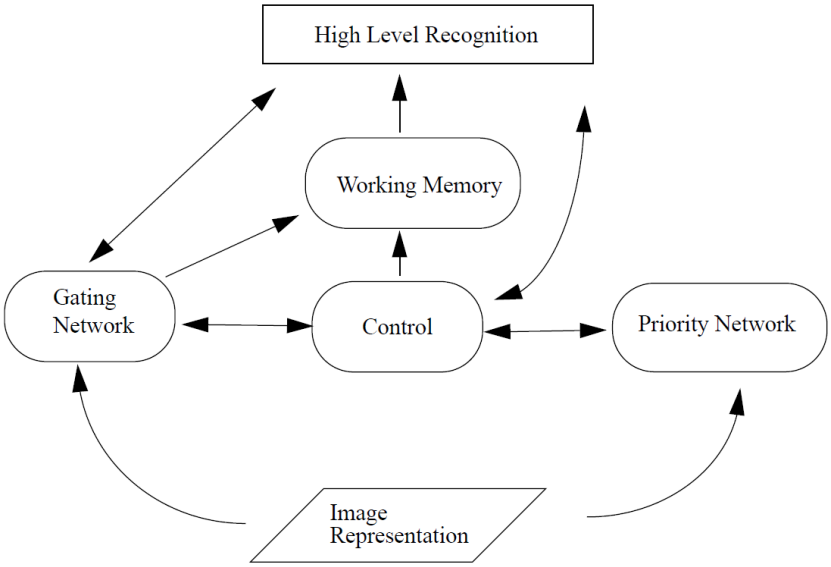


FIGURE 2.3.9: Le modèle connexionniste d'Ahmad [Ahmad 92].

Le principe du blocage (*gating*) est également utilisé par Mozer et Sitton [Mozer 98] afin de proposer un modèle associant reconnaissance d'objet et processus attentionnels. La partie reconnaissance est réalisée *via* un réseau de neurones pyramidal *feedforward*

6. *VIsual Search ITeratively*

classique. La modulation attentionnelle est alors effectuée en insérant une carte attentionnelle entre les deux premiers niveaux de la pyramide. Cette carte permet d'inhiber certaines zones, qui ne participeront plus à la tâche de reconnaissance. La manière de calculer la carte d'attention n'est pas précisée, mais les mécanismes présentés permettent, à partir d'une carte d'attention prédéfinie, d'expliquer de nombreux phénomènes attentionnels.

Le modèle de Vitay et Fix [Vitay 05, Fix 08] utilise un mécanisme attentionnel bien différent des deux modèles cités précédemment. La modélisation est effectuée à un niveau de granularité plus large. Les réseaux de neurones sont ainsi remplacés par des champs neuronaux. Ce formalisme permet de décrire le fonctionnement macroscopique d'un ensemble des neurones, à partir d'un système d'équations différentielles. Celles-ci modélisent les mécanismes d'excitation et inhibition couramment observés lors de l'interaction des différents neurones d'un réseau traditionnel. Ces champs neuronaux sont exploités pour proposer un mécanisme de sélection simple et efficace, permettant le déploiement de l'attention visuelle spatiale.

Algorithmiques

Les modèles appartenant à cette catégorie proposent diverses méthodes difficilement classables dans les précédentes. Souvent dédiées à une application précise, les approches n'en sont pas moins intéressantes.

Lopez [Lopez 06] propose un modèle utilisant uniquement des caractéristiques de forme et de mouvement afin d'améliorer la segmentation de vidéo. Le modèle attentionnel proposé est construit en fonction de l'application cible et est donc difficilement généralisable. Il montre cependant qu'un modèle attentionnel peut améliorer substantiellement certaines tâches de vision par ordinateur.

Le modèle d'Orabona [Orabona 08] est basé sur la notion de proto-objet, que nous avons introduite en section 2.3.2.4 (figure 2.3.10). L'image est tout d'abord segmentée en blobs de couleur uniforme. Une carte de saillance est ensuite calculée en effectuant un filtrage centre-périphérie entre chaque blob et son voisinage. La carte ainsi obtenue représente la saillance de chacun des proto-objets de la scène.

Une approche similaire est exploitée par le modèle d'Aziz et Mertsching [Aziz 08b, Aziz 09a]. L'image segmentée en régions de couleurs uniformes est utilisée pour calculer différentes caractéristiques (contraste de couleur, taille, symétrie, orientation et excentricité). Une carte de saillance globale est alors calculée en fonction de la rareté locale de chaque caractéristique dans chacune des régions de l'image. Notons que ce modèle est l'un des rares modèles d'attention (avec [Sela 97] et [Kootstra 08]) à prendre en compte la symétrie.

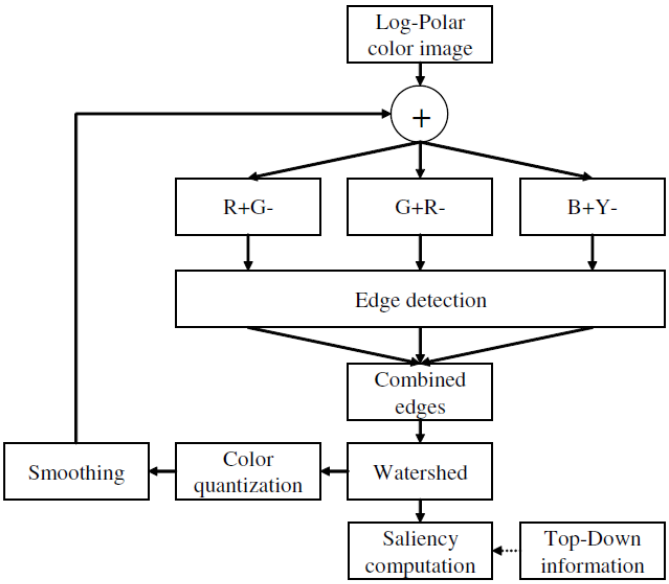


FIGURE 2.3.10: Le modèle d'attention proto-objets d'Orabona [Orabona 08].

Résumé

Nous proposons dans le tableau 2.5 un résumé des avantages et inconvénients de chaque type de modèle hiérarchique d'attention. Malgré l'hétérogénéité des approches utilisées dans chacune des familles présentées dans cette sous-section, il est possible de dégager quelques caractéristiques communes.

2.3.3.3 *Bilan*

Le tableau 2.6 résume les avantages et inconvénients des deux grands types d'approche pour la modélisation de l'attention visuelle. L'approche distribuée est fidèle à la réalité biologique, règle efficacement le problème de la compétition, mais est plus lourde à mettre en œuvre et à étendre. L'approche centralisée propose généralement des solutions plus efficaces computationnellement, mais sa gestion de la dynamique (évolution du focus d'attention dans le temps) n'est pas native et nécessite l'adjonction de méthodes connexionnistes couteuses (WTA + inhibition de retour).

Les modèles attentionnels que nous avons décrits dans cette sous-section ne fonctionnent généralement pas seuls. Ils sont connectés à d'autres modules, dans un système de vision ou de traitement d'images plus vaste. Dans ce cadre, il est nécessaire qu'ils communiquent avec les autres modules et qu'ils puissent s'adapter en fonction du contexte. En section 1.2.2, nous avons donné une définition de l'adaptation et effectué un lien entre celle-ci et notre application cible. Dans la section suivante, nous entrons plus en détail

Type de modèle	Avantage(s)	Inconvénient(s)
Hiérarchiques	Simplicité Efficacité computationnelle Facilement extensible	Méthode de fusion des cartes souvent criticable
Statistiques / probabilistes	Bonne modélisation de la différence par rapport au voisinage Prise en compte de l'information *top-down*	Capacité explicative Plausibilité biologique
Théorie de l'information	Cadre théorique fort Formalisation de la rareté Prise en compte de l'information *top-down*	Capacité explicative Plausibilité biologique
Connexionnistes	Bonne gestion de la compétition entre différentes sources d'information Possibilité de coupler attention et reconnaissance Gestion de la dynamique	Travaille généralement à partir de cartes de saillances fournies
Algorithmiques	Bien adapté aux applications de vision	Eloigné du modèle biologique

TABLE 2.5: Avantages et inconvénients des différents modèles centralisés.

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

TABLE 2.6: Avantages et inconvénients des modèles centralisés et ditribués.

dans les mécanismes utilisés et mettons en avant quelques applications adaptatives liées à notre domaine d'étude.

2.3.4 Et l'adaptation ?

Les systèmes adaptatifs ou auto-adaptatifs sont généralement associés à des agents artificiels, capables de modifier leur comportement individuel ou collectif en fonction d'un changement dans leur environnement. Les systèmes multi-agents correspondent assez bien à cette définition mais la notion d'adaptation est beaucoup plus large. Comme le note Thierry Vieville [Vieville 05], l'adaptation peut être vue de différentes façons :

- Comme *choix d'architecture*, en concevant une architecture modulaire. Un module d'observation supervise le fonctionnement du système global et adapte les paramètres des autres modules. Il peut également supprimer ou ajouter des modules à la demande en fonction de l'état du système. Ce type d'adaptation au niveau système est par exemple utilisé dans les mécanismes de contrôle du déroulement du scénario d'un jeu proposé au sein de notre laboratoire et dans lequel nous avons fourni un modèle d'observation de l'attention du joueur [Rempulski 09].
- Comme mécanisme d'*apprentissage paramétrique*. L'apprentissage est un concept beaucoup plus large que celui d'adaptation. Dans le cadre de l'apprentissage, les propriétés à estimer sont complexes et les information fournies *a priori* faibles. L'adaptation peut être vue comme une version limitée de l'apprentissage où les seules valeurs à estimer sont des paramètres du modèle. Ainsi, alors que certains systèmes d'attention visuelle utilisent l'apprentissage pour piloter leur branche *top-down* [Frintrop 05b], nous avons choisi de mettre en place un mécanisme d'adaptation (chapitre 4) permettant de modifier certains paramètres de notre modèle attentionnel en fonction d'un critère de performance (exploration de l'espace par exemple).

Dans ces deux « visions » de l'adaptation, la notion de bouclage apparaît en filigrane : pour pouvoir modifier et adapter un système ou ses paramètres, il faut pouvoir l'observer puis le modifier. Les sorties du système servent à en modifier les entrées : on crée une boucle de rétroaction.

Alors que dans les cadres plus généraux de la robotique [Camus 07] ou de l'indextion d'images [Tollari 05a, Tollari 05b], le terme d'adaptation est souvent mentionné , il est beaucoup plus rarement évoqué dans le cadre de l'attention visuelle. Pourtant, comme le souligne Garbay [Garbay 00], en s'appuyant sur les travaux de Tsotsos [Tsotsos 90] concernant l'attention, la vision et la complexité, l'attention est un mécanisme d'adaptation permettant de changer les paramètres du système de vision (*via* les focalisations) afin d'obtenir le meilleur compromis entre objectif poursuivi et ressources engagées. Selon Garbay, ce mécanisme d'adaptation concerne les connaissances *a priori* (système attentionnel) et l'évaluation *a posteriori* de la réussite de la tâche (mécanismes de révision / réparation). Il propose alors un schéma général des mécanismes de focalisation et

d'adaptation (figure 2.3.11).

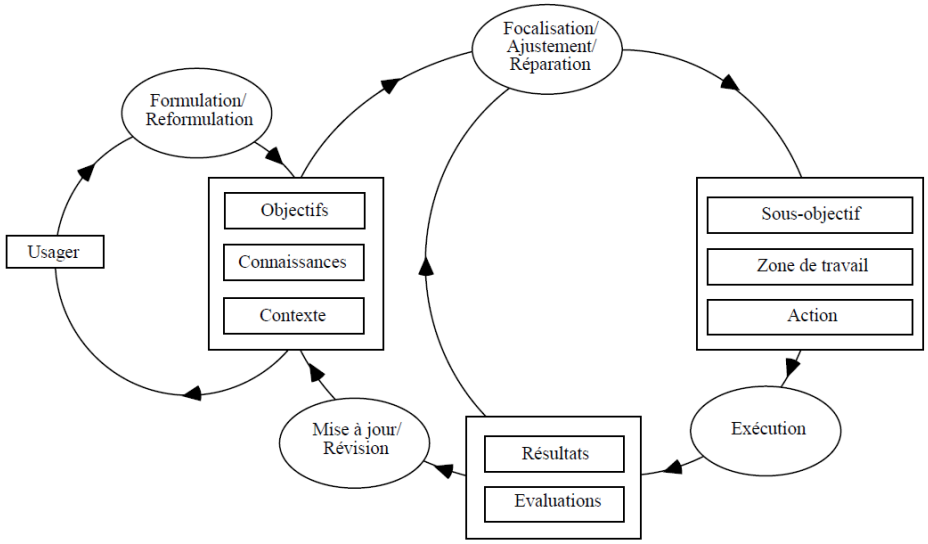


FIGURE 2.3.11: Enchaînement des mécanismes de focalisation et d'adaptation. D'après [Garbay 00].

2.4 Conclusion

Dans la première section de ce chapitre, nous avons abordé la question de la justification des mécanismes attentionnels. Nous avons en particulier relayé l'hypothèse la plus répandue, selon laquelle l'attention serait un mécanisme de sélection destiné à « compenser » nos capacités de traitement limitées.

Dans une seconde section nous avons introduit les principales théories fondatrices de l'étude de l'attention, principalement basées sur cette hypothèse de capacités limitées. Leur étude a permis de mettre en exergue différentes propriétés des modèles attentionnels théoriques, que l'on retrouve également dans leurs pendants computationnels.

Enfin, dans la dernière section, nous avons tout d'abord effectué un panorama des différentes utilisations des modèles computationnels d'attention. Celui-ci nous a amené à définir un jeu de contraintes, permettant de caractériser chaque type d'application. Nous avons ensuite effectué une première taxonomie des modèles en fonction de leurs propriétés (bottom-up / top-down, spatiale / object, etc.), puis de leur appartenance à une famille de modèle (distribué, centralisé-hierarchique, centralisé-statistique, etc.). A partir de ces deux taxonomies, nous avons caractérisé l'adéquation entre les différentes familles de modèles, et le jeu de contraintes défini en section 2.3.1.

On peut ainsi constater que chaque famille de modèles répond partiellement à l'ensemble des contraintes que nous avons associées à l'application « système de vision temps réel et adaptable » (tableau 2.7). Compte tenu des propriétés de fidélité, invariance, dynamique et adaptation des modèles distribués, et des propriétés de rapidité et extensibilité des modèles hiérachiques, on peut conclure qu'une approche hybride entre ces deux solutions permettrait d'obtenir les propriétés désirées.

Ce type d'approche a déjà été partiellement exploré par les modèles centralisés connexionnistes, puisqu'ils combinent généralement un modèle hérarchique pour générer une carte de saillance puis une approche distribuée pour calculer les différentes focalisations attentionnelles. Cependant, dans ces modèles, la compétition entre les différents attributs (intensité, couleur, orientation ,etc.) est effectuée par le système hiérarchique. On ne bénéficie alors pas du principe de compétition biaisée entre ces différentes sources de saillance. Ainsi, nous pensons qu'il est intéressant de ne pas passer par une représentation centralisée de la saillance. Nous proposons plutôt de connecter les cartes de caractéristiques ou de singularité à un modèle compétitif, qui pourra alors pleinenement jouer son rôle de gestion de la dynamique des focalisations et de la compétition entre les différents attributs. Dans les deux chapitres suivants, nous présenterons cette solution et étudierons ses propriétés selon les critères *FAIRED* définis en tableau 2.7.

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

TABLE 2.7: Adaptation des différents modèles aux contraintes d'un système de vision. La première ligne du tableau correspond aux objectifs que nous avons définis. En vert : les critères atteints ou dépassés par les différentes familles de modèles.

Points clés

Positionnement

- ❏ De part nos capacités de traitement et / ou d'action limtées, l'attention est un mécanisme clé de gestion rapide de l'information.
- ❏ Les principaux attributs pris en compte pour le déploiement de l'attention sont établis de manière consensuelle : intensité, couleur, orientation et mouvement.
- ❏ A l'inverse, l'attention est un phénomène aux multiples visages. Sur de nombreux points son interprétation est duale, elle peut être : exogène et automatique ou endogène et contrôlée, dirigée spatialement ou sur des objets, représentée de manière centralisée ou distribuée, déployée de manière ouverte ou couverte.
- ❏ Bien que rarement associée à l'attention, l'adaptation est un processus nécessaire à une simulation efficace de l'attention.

Contributions

- ❏ Nous définissons un jeu de critères pour la caractérisation d'un modèle d'attention (critères *FAIRED*).
- ❏ Nous établissons une taxonomie des différentes approches computationnelles de l'attention.
- ❏ Nous en déduisons le degré d'adéquation des différentes familles de modèles à notre cahier des charges.

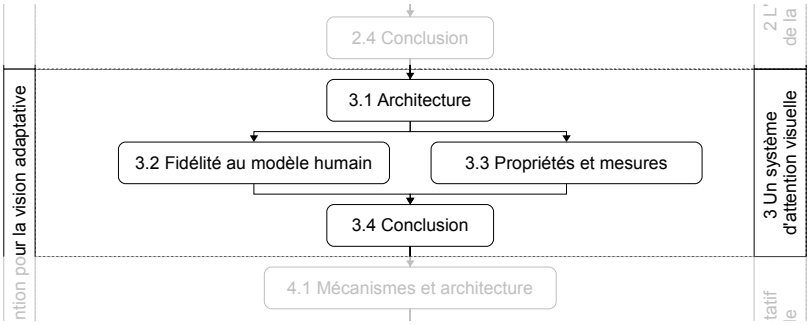
Vers un modèle d'attention pour la vision adaptative

« L'intelligence, c'est la faculté d'adaptation. »

(Andre Gide)

Chapitre 3

Un système d'attention visuelle



Dans cette seconde partie, nous présentons notre modèle d'attention visuelle, dédié à la vision adaptative. Nous effectuons l'étude de notre système en deux étapes : la première (ce chapitre) décrit la partie *bottom-up* du modèle ; la seconde (le chapitre 4) aborde les mécanismes d'adaptation formant la partie *top-down.*

Dans ce chapitre, nous commençons par décrire l'architecture de notre système (section 3.1), puis nous évaluons la plausibilité de ses prévisions en le comparant à une vérité terrain issue d'expérimentations oculométriques (section 3.2). Enfin, en section 3.3, nous étudions en détail l'influence des différents paramètres du modèle en fonction de critères dérivés de notre cahier des charges (défini en sous-section 6.2.1) et du bilan effectué après l'analyse des modèles computationnels existant et de leurs applications (section 2.4). Ces critères concernent : la stabilité, la reproductibilité, l'exploration de l'espace, et la dynamique de notre modèle.

3.1 Architecture

Notre modèle attentionnel est composé de deux éléments (figure 3.1.1) aux fonctions distinctes : un système visuel et un système attentionnel.

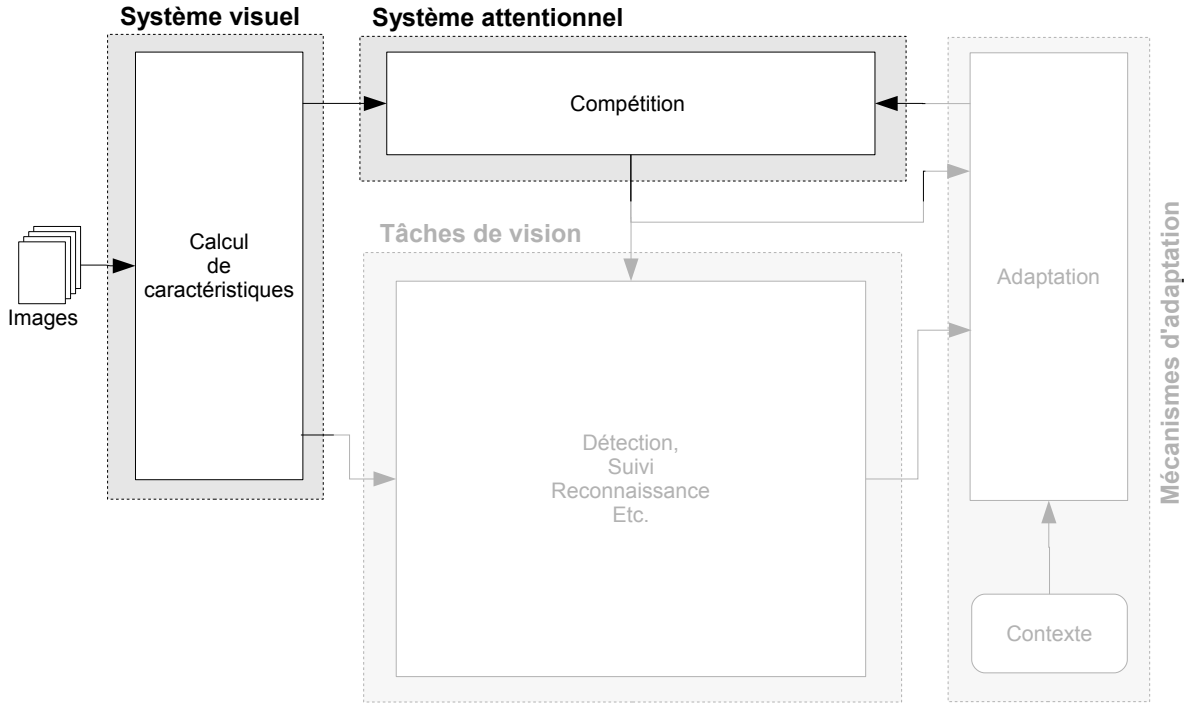


FIGURE 3.1.1: Parties du système présentées dans ce chapitre : nous ne traitons que de l'attention visuelle.

Le système visuel est inspiré des modèles d'attention centralisés hiérarchiques, et en particulier des modèles d'Itti [Itti 98] et Frintrop [Frintrop 05a]. La scène visuelle est décomposée en différentes caractéristiques selon une approche multi-résolutions[Courboulay 02]. 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 singularité 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.

Les prochaines sections décrivent en détail chacune des étapes de ces deux systèmes, formant notre modèle hybride d'attention *bottom-up*. Celui-ci permet de bénéficier des avantages de modèles hiérarchiques (rapidité, extensibilité) et distribués (fidélité, dynamique, adaptation) tout en limitant leurs inconvénients (fusion des cartes critiquable pour les modèles hiérarchiques, et complexité et difficulté d'ajout de nouvelles caractéristiques pour les modèles centralisés).

3.1.1 Schéma général et contributions

La figure 3.1.2 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[Itti 98] dans ses travaux.

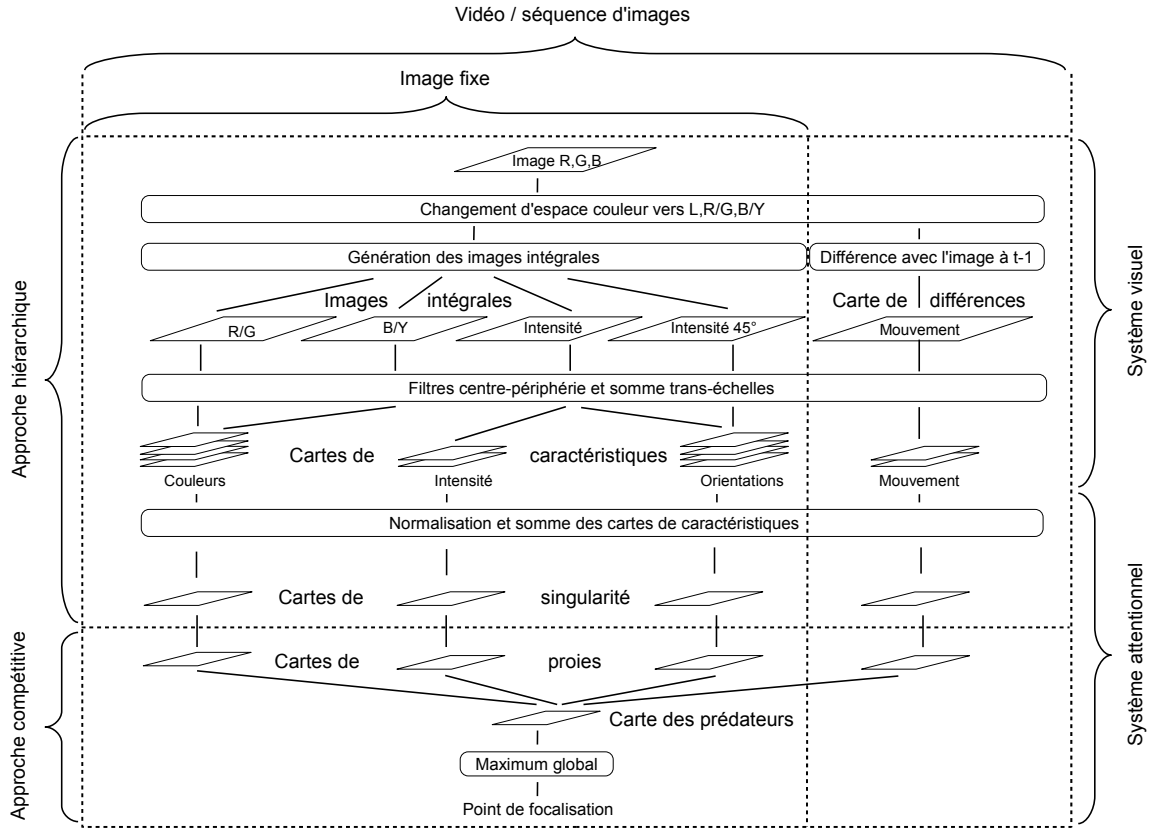


FIGURE 3.1.2: Schéma général du système d'attention visuelle.

Le système visuel est traité avec une approche purement hiérarchique, très inspirée de l'approche de[Itti 98] et des améliorations apportées à son modèle par [Frintrop 05a]. A partir de l'image d'entrée, on construit plusieurs cartes par caractéristique traitée *(feature*

maps) : 2 pour l'intensité (réponses aux filtres centre-périphérie on/off et off/on), 4 pour la couleur (rouge, vert, jaune et bleu), 4 pour l'orientation (0°, 45°, 90° et 135°) et 2 pour le mouvement (réponses aux filtres centre-périphérie on/off et off/on).Ces cartes sont ensuite fusionnées en 4 cartes de singularité *(conspicuity maps)* représentant la saillance de chacun des éléments de la scène analysée d'un point de vue intensité, couleur, orientation et mouvement. La majorité des traitements est commune au traitement des images fixes et des séquences d'images (le seul traitement spécifique à ces dernières est la génération de la carte de mouvement).

Notre contribution dans cette partie du modèle est la suivante :

- optimisation du modèle afin d'obtenir de bonnes performances computationelles (génération des cartes en temps réel) ;
- introduction d'un mécanisme de pseudo-flou rétinien computationnellement efficace, permettant au modèle de « simuler » la résolution variable de la rétine ;
- proposition d'un opérateur de normalisation des cartes de caractéristiques et de singularité basé sur la théorie de l'information. Celui-ci permet un renforcement des caractéristiques rares, sans avoir à déterminer de seuil ou recourir à une procédure itérative coûteuse.

Le système attentionnel utilise une approche hybride entre les traitements hiérarchiques et compétitifs. Il aurait été possible de réaliser un système purement compétitif (en fournissant les cartes de caractéristiques au système attentionnel) mais cette solution aurait alourdi inutilement le système. Nous avons donc choisi de garder une phase de prétraitement hiérarchique permettant de générer une carte de singularité par caractéristique. Ce modèle attentionnel se démarque des modèles de l'état de l'art en proposant une alternative originale basée sur une analogie proies / prédateurs : la compétition entre les différentes sources d'information est modélisée comme une lutte entre différentes espèces dans un écosystème fermé. Notre contribution dans cette seconde partie du modèle est la suivante :

- proposition d'un mécanisme unique pour la compétition des cartes de singularité et la génération des focalisations attentionnelles ;
- introduction d'une part d'aléatoire, permettant de modéliser et de faire varier la « curiosité » du modèle d'attention ;
- prise en compte de la préférence centrale lors de la génération des focalisations attentionnelles ;
- caractéristation de l'influence des paramètres du système sur son comportement.

Une partie de ces contributions a été valorisée dans[Perreira Da Silva 09a] et [Perreira Da Silva 10b].

3.1.2 Images d'illustration

Dans ce chapitre, nous illustrons les différents algorithmes par des résultats, générés à partir de l'une ou l'autre des deux images présentées en figure 3.1.3. A l'entrée du système, les images sont redimensionnées à une largeur de 256 pixels, la hauteur étant

calculée pour respecter le ratio largeur / hauteur original. Concernant les vidéos, nous utilisons une séquence d'images synthétiques (256 × 186, 100 trames) dont quelques trames sont représentées en figure 3.1.4 (la version complète est disponible en ligne : `http://www.youtube.com/watch?v=6sHcxPPs4UA`).

FIGURE 3.1.3: Images exemples, servant à illustrer les différents algorithmes de ce chapitre. A gauche l'image « Parrots », représentative des caractéristiques d'intensité et couleur. A droite, l'image « Road » représentative des caractéristiques d'orientation

Afin d'assurer une meilleure lisibilité des différentes images présentées, celles-ci ont été normalisées afin que la valeur des pixels appartienne à l'intervalle [0 ;255]. Cette normalisation n'est utilisée que pour la présentation des figures. Si une normalisation est effectuée par un des algorithmes de notre modèle cela sera spécifié explicitement lors de sa description.

3.1.3 Calcul des cartes de singularité

La partie « système visuel » de notre modèle d'attention permet de générer les cartes de singularité 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 [Sela 97, Sun 08].

Les traitements appliqués différant légèrement en fonction de la méthode choisie, nous avons découpé notre présentation du calcul des cartes de singularité en trois parties. En sous-section 3.1.3.1 nous décrivons les prétraitements communs aux deux approches, puis détaillons les traitements appliqués lorsque tout le champ visuel est traité de la même manière (sous-section 3.1.3.2). Enfin, en sous-section 3.1.3.3 nous expliquons comment il est possible d'appliquer un flou-rétinien à l'image, tout en accélérant le calcul des différentes cartes de caractéristiques.

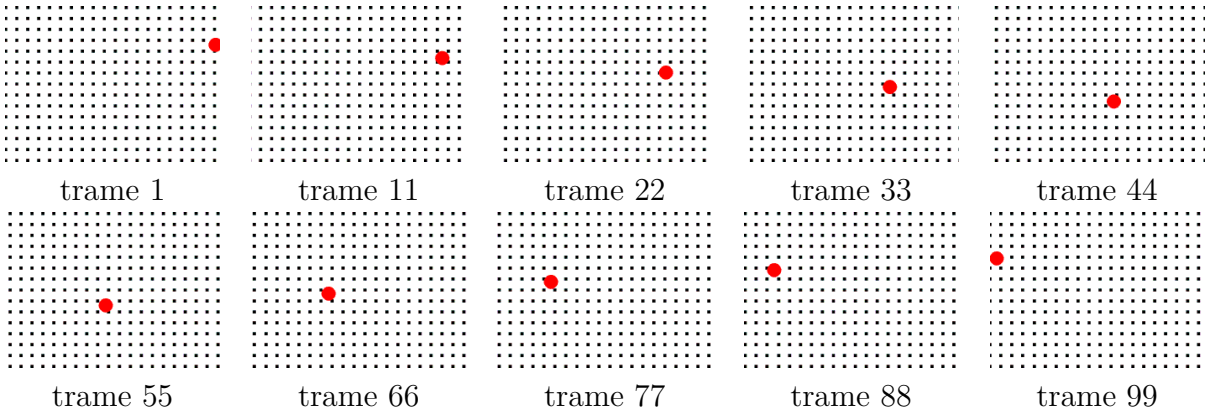


FIGURE 3.1.4: Exemple de séquence « Boule + Grille », illustrative des traitements du mouvement de notre système attention. Une boule rouge se déplace de droite à gauche de haut en bas puis de bas en haut pendant qu'une grille de points noirs se déplace de gauche à droite.

3.1.3.1 Prétraitements communs

Changement d'espace couleur

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 05a] 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 privilégier l'espace L,R/G,B/Y défini comme suit :

$$I_L = \frac{I_R + I_G + I_B}{3} \tag{3.1.1}$$

$$I_{R/G} = 127 + \frac{N_R - N_G}{2} \tag{3.1.2}$$

$$I_{B/Y} = 127 + \frac{N_B - N_Y}{2} \tag{3.1.3}$$

avec

$$\begin{aligned} N_R &= I_R - L \\ N_G &= I_G - L \\ N_B &= I_B - L \\ N_Y &= \frac{N_R + N_G}{2} \end{aligned}$$

Cet espace couleur est semblable à celui utilisé dans [Itti 98] à un détail près : les canaux couleurs sont normalisés en soustrayant la valeur de l'intensité et non en la divisant par celle-ci. Cette normalisation additive et non multiplicative est très rapide et ne pose pas de problème en pratique car notre système encodant les images dans des nombres réels (voir annexe D), il est tout à fait possible d'obtenir des intensités négatives. Précisons malgré tout que la phase de conversion de couleur représente un temps négligeable dans l'ensemble du traitement attentionnel. Si l'application cible de notre modèle d'attention nécessite plus de fidélité biologique, on pourra choisir un espace plus approprié sans trop pénaliser la réactivité du système.

Calcul des images intégrales

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 multi-résolutions. Ces traitements sont généralement réalisés *via* le calcul de pyramides gaussiennes [Itti 98, Choi 06]. Pour accélérer les calculs, nous avons choisi une approche alternative, inspirée de [Frintrop 07], qui permet de remplacer le calcul des pyramides multi-résolutions par des images intégrales [1] [Viola 02] (cf. Annexe D.3.1). A la différence de [Frintrop 07] 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 (cf. sous section 3.1.3.2).

Compte tenu des différentes caractéristiques à prendre en compte, nous devons pré-calculer 4 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

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

proposée par [Lienhart 02].
Les traitements effectués ensuite dépendent de la prise en compte de la résolution spatiale variable de la rétine humaine. Dans la sous-section suivante, nous décrivons le cas classique, sans ce flou rétinien.

3.1.3.2 Sans pseudo flou rétinien

Pyramides d'intensité et de couleur

Comme nous l'avons vu dans le chapitre 2.3.3.2, 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 boites (DoB). En effet, les images intégrales ne permettant de calculer « que » des sommes de zone rectangulaire, il nous faut adapter l'ensemble des filtres à ce mode de calcul (figure 3.1.5).

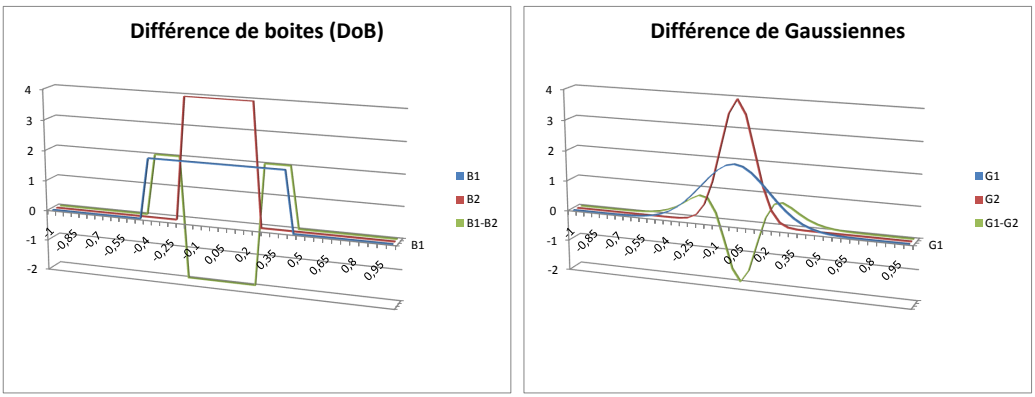


FIGURE 3.1.5: Différence de boites et de gaussiennes.

En utilisant le procédé décrit ci-dessus, on calcule les différences de boites 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 WxH, on aura au plus $N = 1 + log_2(min(W, H)$, arrondi à l'entier inférieur.

Pour chaque niveau r de résolution, on calcule alors la différence entre deux filtres boite. On ne fait pas varier la résolution de l'image source (qui est une image intégrale)

mais la taille du filtre. Dans notre implémentation, la taille des filtres $B_{1,r}$ et $B_{2,r}$ est :

$$S_{B_{1,r}} = \begin{cases} 1 & si\ r = 0 \\ 2^{r-1} \times 3 & sinon \end{cases} \tag{3.1.4}$$

$$S_{B_{2,r}} = \begin{cases} 3 & si\ r = 0 \\ 2^{r-1} \times 6 & sinon \end{cases} \tag{3.1.5}$$

avec $r \in \{1, N-3\}$. 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 (figure 3.1.6). 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).

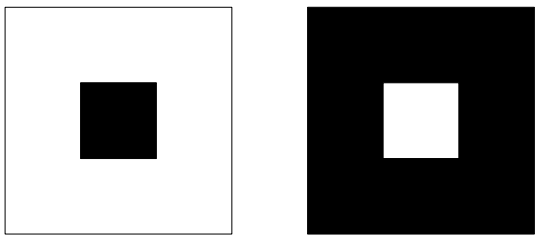


FIGURE 3.1.6: Combinaison de filtres boite. A gauche un filtre off-center / on-surround. A droite, un filtre on-center / off-surround.

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 (figure 3.1.7) :

$$P_{L_{on}}(x, y, r) = \begin{cases} CS_{L,r}(x, y) & si\ CS_{L,r}(x, y) > 0 \\ 0 & sinon \end{cases} \tag{3.1.6}$$

$$P_{L_{off}}(x, y, r) = \begin{cases} -CS_{L,r}(x, y) & si\ CS_{L,r}(x, y) \leq 0 \\ 0 & sinon \end{cases} \tag{3.1.7}$$

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

boite » [2].

FIGURE 3.1.7: Pyramides centre-périphérie du canal intensité de l'image « Parrots ». En haut, pyramide On-Off $P_{L_{on}}$. En bas, pyramide Off-On $P_{L_{off}}$.

On procède de même avec les couples de pyramides P_R, P_G et P_B ,P_Y (figure 3.1.8).

Pyramides d'orientation

Dans les modèles d'Itti [Itti 98] et Frintrop [Frintrop 05a], les pyramides orientées sont calculées à partir de filtres de Gabor appliqués sur une pyramide Laplacienne. Le modèle d'Itti applique sur cette pyramide une différence centre-périphérie afin d'obtenir les cartes de caractéristiques finales. Mais comme le souligne Frintrop, le filtrage centre-périphérie orienté est déjà réalisé par les filtres de Gabor. On peut donc se passer des filtres centre-périphérie supplémentaires sur les pyramides de Gabor.

Ces filtres étant assez lents à calculer, 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 (figure3.1.9).

On calcule pour les résolutions $r \in \{1, N-3\}$, quatre pyramides orientées P_{Ori0}, P_{Ori45}, P_{Ori90} et P_{Ori135} correspondant aux orientation 0°, 45°, 90° et 135° (figure 3.1.10). Celles-ci sont moins précises, mais plus rapides à calculer qu'avec les méthodes proposées par [Itti 98] et [Frintrop 05a] (cf. annexe D.3).

Pyramides de mouvement

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

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

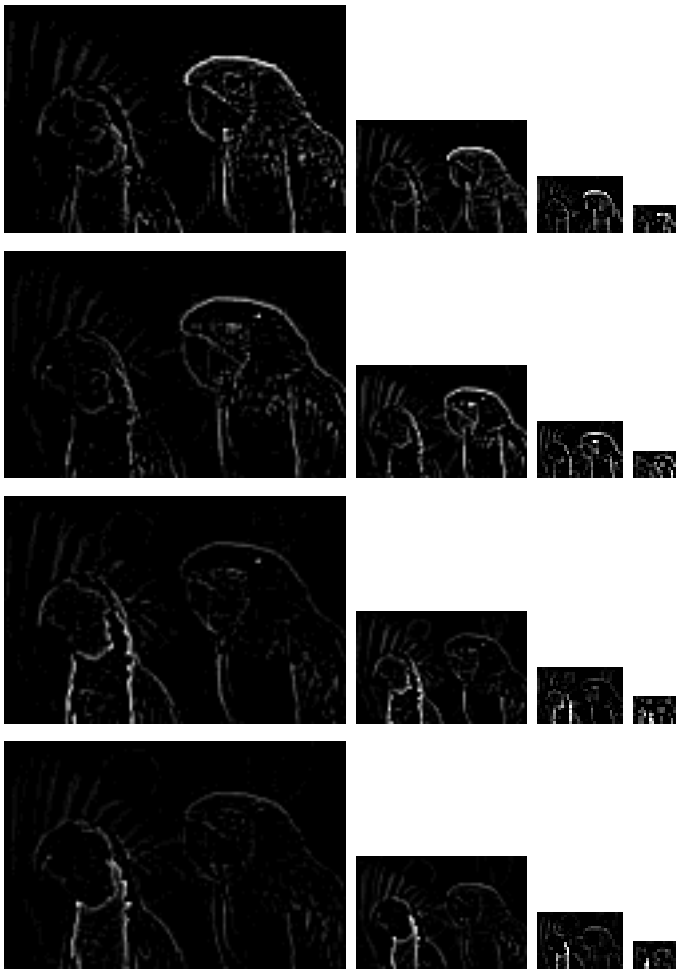


FIGURE 3.1.8: Pyramides centre-périphérie des canaux R/G et B/Y de l'image « Parrots ». De haut en bas, P_R, P_G, P_B et P_Y.

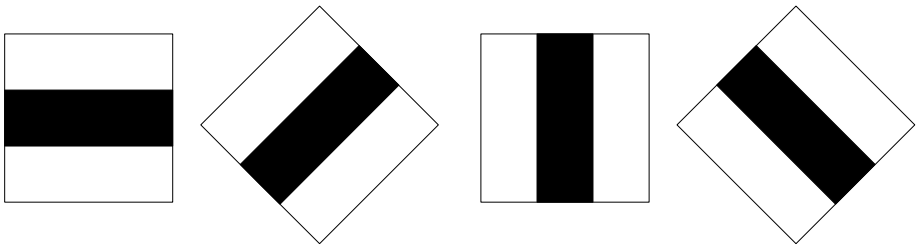


FIGURE 3.1.9: Les différents filtres sélectifs à l'orientation pouvant être générés à partir des images intégrales « standard »II_L et avec rotation IIR_L.

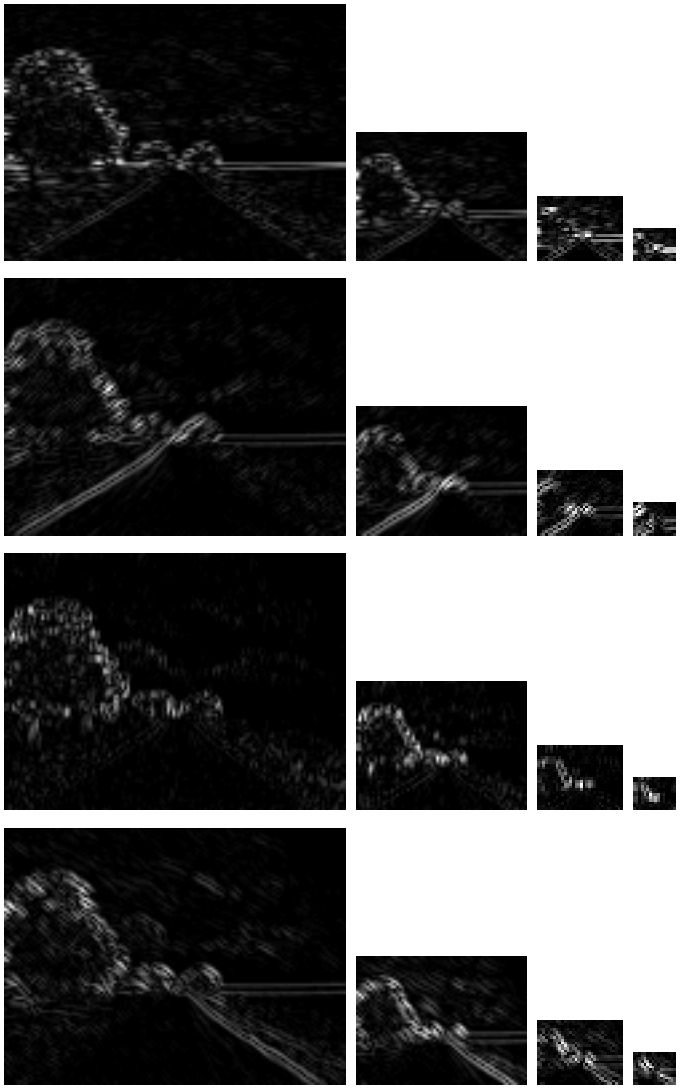


FIGURE 3.1.10: Pyramides orientées calculées sur l'image « Road ». De haut en bas, P_{Ori0}, P_{Ori45}, P_{Ori90} et P_{Ori135}.

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 [Zach 07], 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 connaitre précisement 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 d'images 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 la trame précédente :

$$I_{Diff_L} =\mid I_{L_t} - I_{L_{t-1}} \mid \tag{3.1.8}$$

On calcule ensuite les classiques pyramides centre-périphérie on/off $P_{M_{on}}$et off/on $P_{M_{off}}$. Comme nous l'évoquons en annexe D.3, 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 multi-résolutions plus classiquement, en remplaçant toutefois les filtres gaussien par des filtres boite.

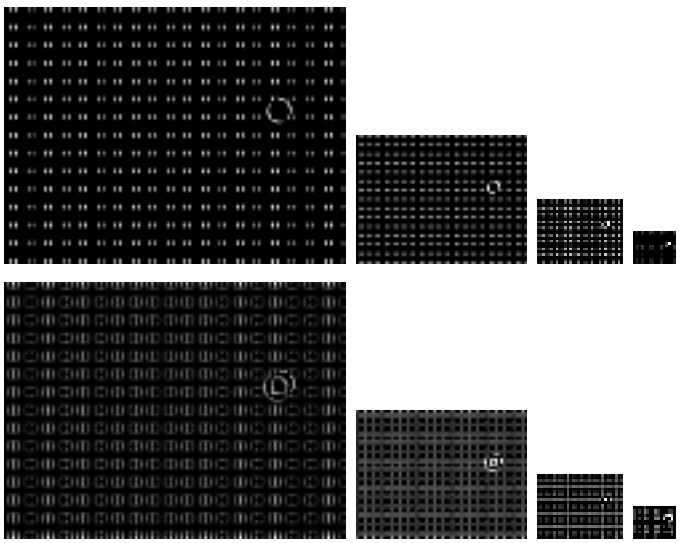


FIGURE 3.1.11: Pyramides de mouvement du canal intensité de la trame 20 de la séquence « Boule + Grille ». En haut, pyramide On-Off $P_{M_{on}}$. En bas, pyramide Off-On $P_{M_{off}}$.

Cartes de caractéristiques, cartes de singularité et normalisation

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 98]et de manière similaire à [Frintrop 05a], 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 aura par exemple pour $FM_{L_{on}}$:

$$FM_{L_{on}}(x,y) = \bigoplus_r P_{L_{on}}(x,y,r) \tag{3.1.9}$$

avec $R \in \{1, N-2\}$et $\bigoplus$ l'opérateur d'addition trans-échelles *(across-scale addition)* .

Les différentes résolutions sont respectivement sur échantillonnées et sous échantillonnées par interpolation bilinéaire et filtre boite afin de correspondre à la résolution demandée (généralement r_2). Chaque carte représente alors les éléments les plus saillants de l'image pour une caractéristique donnée, quelle que soit l'échelle d'observation.

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 (comme le font [Itti 98] et [Frintrop 05a]) à un regroupement des cartes de caractéristiques en 4 cartes de singularité *(conspicuity maps)* d'intensité SM_L, de couleur SM_{C}, d'orientation SM_O et de mouvement SM_M.

Comme l'ont montré [Itti 98, Frintrop 05a] ou [Bruce 03], 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. [Itti 98] introduit un opérateur $\mathcal{N}(X) = X * (M - \bar{m})^2$ avec M le maximum global et $\bar{m}$ la moyenne des maximums locaux. Cet opérateur normalise les cartes afin de favoriser celles ayant un unique pic d'activité. Malheureusement, comme le souligne [Itti 01a], si une carte contient de grand maxima de même valeur, alors la différence devient nulle et la carte n'est pas prise correctement en compte lors du calcul de la carte de singularité. Une solution dans ce même article est d'utiliser un calcul itératif basé sur l'opérateur $\mathcal{W}(X) = \frac{X}{\sqrt{n}}$ avec n le nombre de maxima locaux situés au-dessus d'un certain seuil. Cette solution donne des résultats corrects mais nécessite l'estimation toujours difficile d'un seuil pour calculer n.

Nous proposons un autre opérateur de normalisation inspiré de [Mancas 07] 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 :

$$SI_i(x, y) = -\log(p(FM_i^{'}(x, y))) \tag{3.1.10}$$

$FM^{'}$ étant la version quantifiée sur 16 valeurs de la carte de caractéristiques 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.

$$FM_i^{"}(x, y) = \frac{FM_i(x, y) \times SI_i(x, y)}{\log(Card(FM_i))} \tag{3.1.11}$$

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.

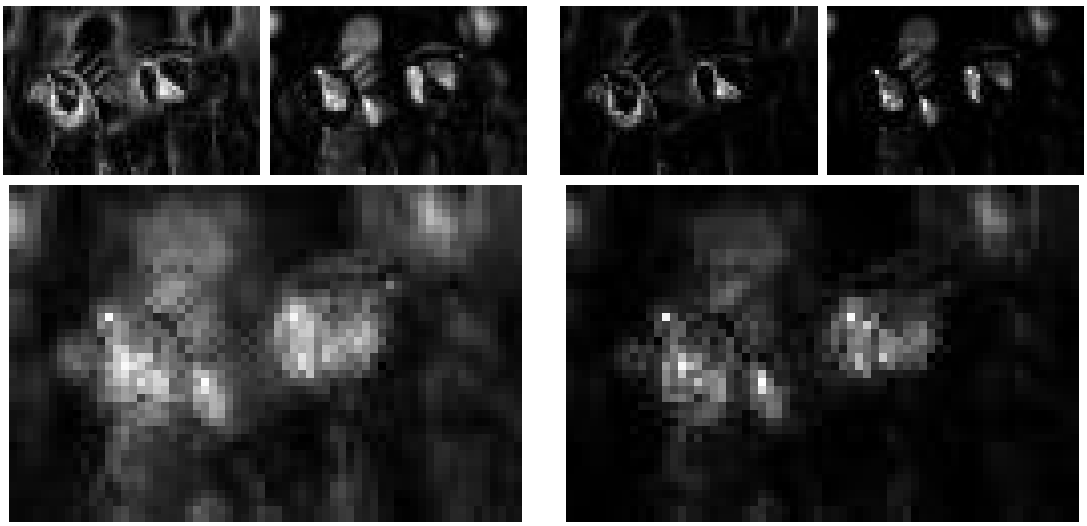
$$SM_L = \frac{FM_{L_{on}}^{"} + FM_{L_{off}}^{"}}{2} \tag{3.1.12}$$

$$SM_C = \frac{FM_R^{"} + FM_G^{"} + FM_B^{"} + FM_Y^{"}}{4} \tag{3.1.13}$$

$$SM_O = \frac{FM_{Ori0}^{"} + FM_{Ori45}^{"} + FM_{Ori90}^{"} + FM_{Ori135}^{"}}{4} \tag{3.1.14}$$

$$SM_{LM} = \frac{FM_{M_{on}}^{"} + FM_{M_{off}}^{"}}{2} \tag{3.1.15}$$

Les figures 3.1.12 à 3.1.15 illustrent le gain obtenu par l'utilisation de la normalisation pour les différentes cartes de caractéristiques ainsi que les cartes singularité résultantes.



Sans normalisation Avec normalisation favorisant la rareté

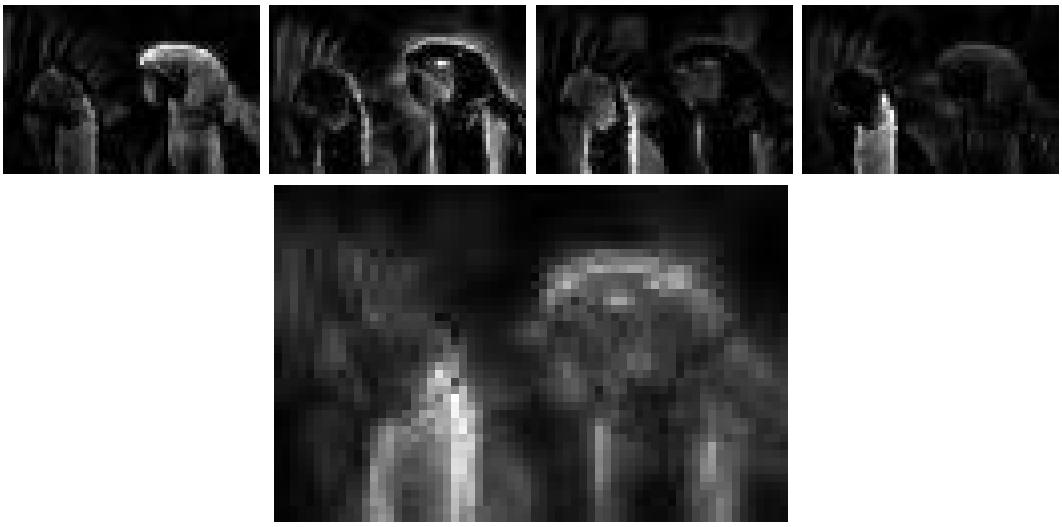
FIGURE 3.1.12: En haut, les cartes de caractéristiques d'intensité On/Off (gauche) et Off/On (droite) de l'image « Parrots ». En bas, la carte de singularité d'intensité.

3.1.3.3 Pseudo flou rétinien

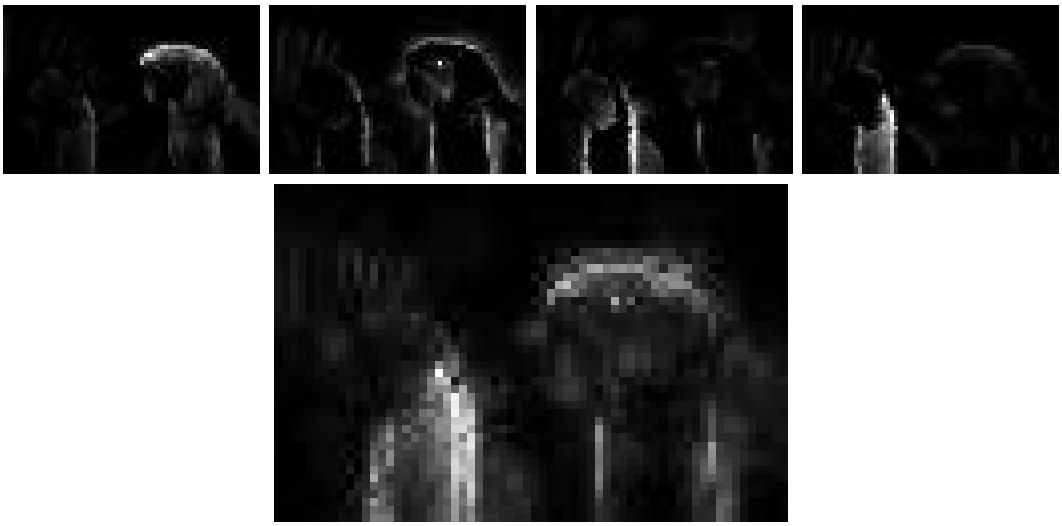
La grande majorité des modèles computationnels d'attention étudiés au chapitre 2 effectue les mêmes traitements sur tous les pixels des images qui leur sont fournies. Certains néanmoins, utilisent la transformation *log polaire* afin de simuler la structure interne de la rétine, dont la résolution est 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. Nous souhaitions cependant étudier l'impact d'une représentation à résolution variable sur la modélisation de l'attention. Une solution efficace et adaptée à notre architecture est l'utilisation non plus de pyramides multi-résolutions, mais de colonnes multi-résolutions (figure 3.1.16). Nous justifions cette structure dans le paragraphe suivant.

Nous décrivons le principe de calcul des colonnes mutli-résolutions en prenant pour exemple les cartes de caractéristiques d'intensité. Le principe est également appliqué aux cartes de couleur et d'orientation.

Dans la section 3.1.3.2 on calcule, à partir de l'image intégrale II_L et pour chaque résolution, l'ensemble des réponses aux filtres centre-périphérie. Dans le cas des colonnes multi-résolutions, on limite les calculs à une zone de taille constante (par exemple 16×16

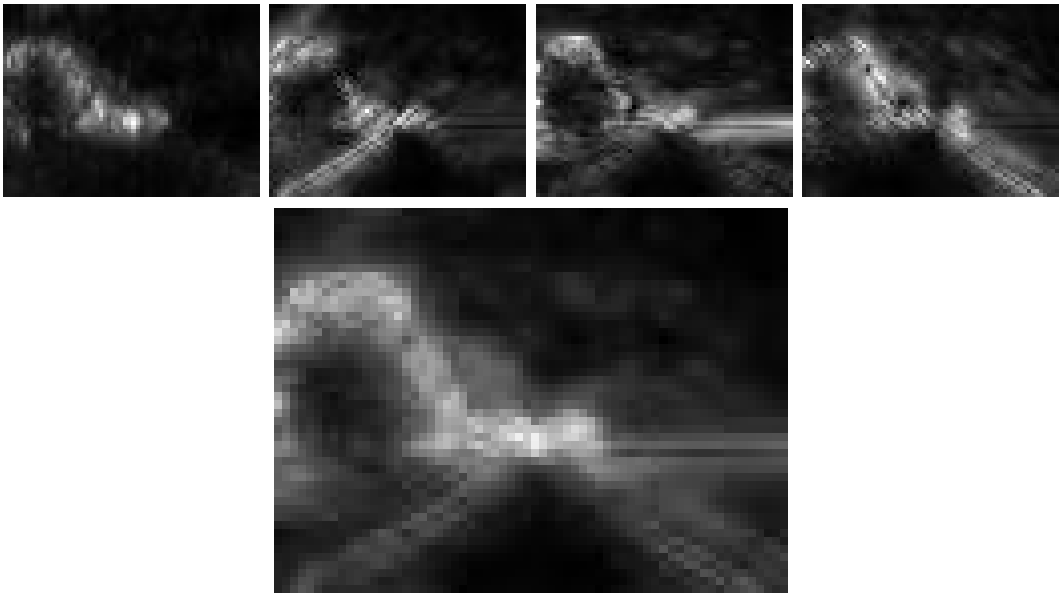


Sans normalisation

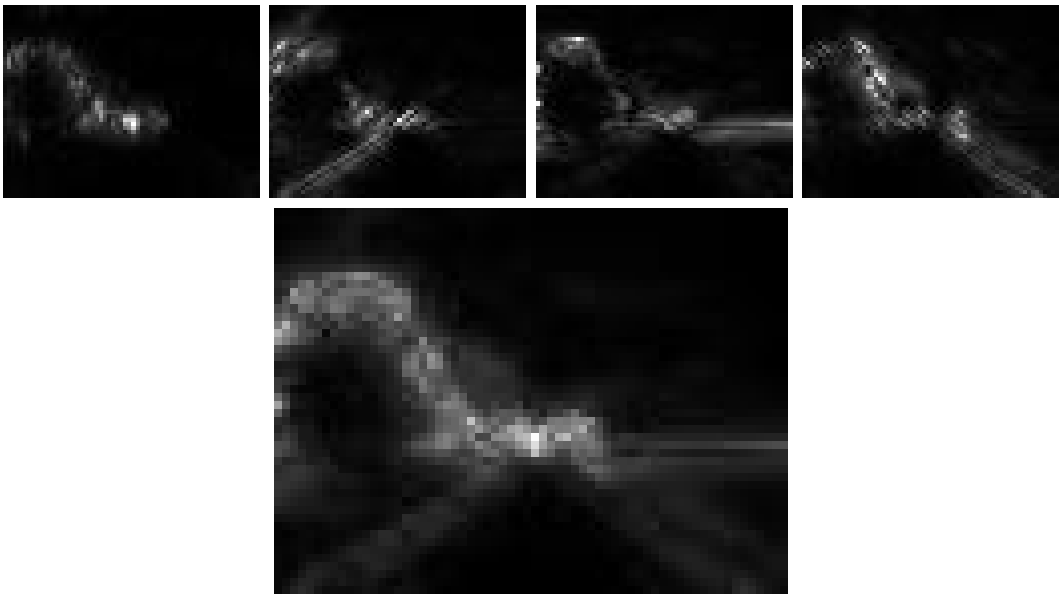


Avec normalisation favorisant la rareté

FIGURE 3.1.13: En haut, les cartes de caractéristiques rouge, vert, bleu et jaune (de gauche à droite) de l'image « Parrots ». En bas, la carte de singularité de couleur.

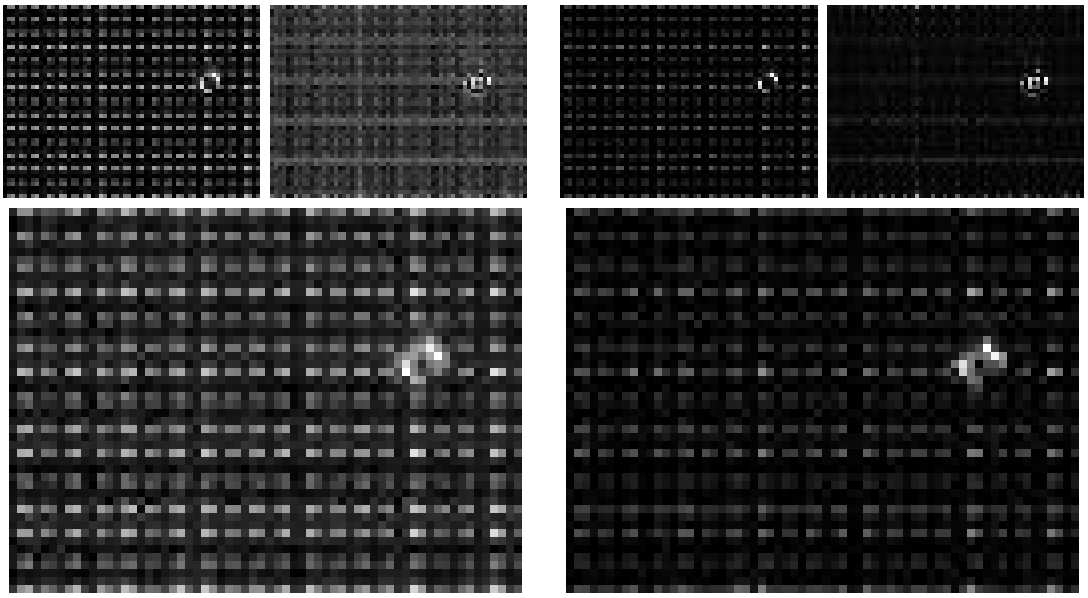


Sans normalisation



Avec normalisation favorisant la rareté

FIGURE 3.1.14: En haut, les cartes de caractéristiques des orientation 0°, 45°, 90°, 135° (de gauche à droite) de l'image « Road ». En bas, la carte de singularité d'orientation.



Sans normalisation | Avec normalisation favorisant la rareté

FIGURE 3.1.15: En haut, les cartes de caractéristiques de mouvement On/Off (gauche) et Off/On (droite) sur la trame 20 de la séquence « Boule + Grille ». En bas, la carte de singularité de mouvement.

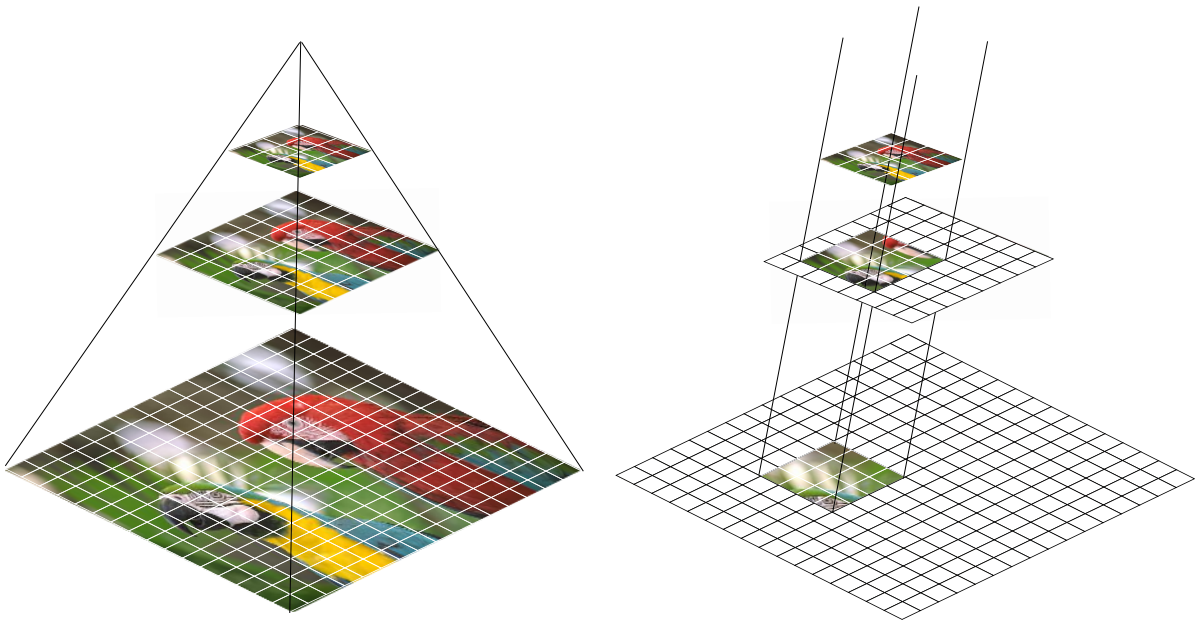


FIGURE 3.1.16: Pyramide multi-résolutions vs. colonne multi-résolutions. Dans la pyramide (à gauche), chaque résolution représente la totalité de l'image. Dans la colonne (à droite), l'image n'est représentée entierement qu'à la résolution la plus basse.

pixels), quelle que soit la résolution. Cette zone de taille constante est centrée sur la position du dernier point de focalisation calculé 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.

Cette représentation permet également de réduire considérablement les calculs nécessaires. Pour une image 256x256 calculée sur 5 niveaux de résolution on aura :

- $256 \times 256 + 128 \times 128 + 64 \times 64 + 32 \times 32 + 16 \times 16 = 87296$ réponses de filtres calculées dans le cas de la pyramide multi-résolution ;
- $16 \times 16 \times 5 = 256 \times 5 = 1280$ réponses de filtre calculées dans le cas de la colonne multi-résolution, soit 98.5% de calcul en moins !

Les gains computationnels sont en réalité moins importants car l'addition trans-échelles des différents éléments de la colonne est plus complexe. De plus, le calcul des images intégrales représente un coût constant non négligeable par rapport au filtrage. Malgré cela, en prenant en compte l'ensemble des calculs effectués, la version « colonne » est 33% plus rapide que la version « pyramide » (voir Annexe D).

Précisons enfin que le calcul des cartes de mouvement n'est pas affecté par les changements décrits ci-dessus et ceci pour deux raisons :

- bien que la rétine ait une résolution variable, sa sensibilité au mouvement est importante en périphérie ;
- la méthode décrite ci-dessus est computationnellement intéressante lorsque l'on utilise des images intégrales pour calculer les pyramides / colonnes. Ce n'est pas le cas pour la carte de mouvement qui est calculée à partir de pyramides classiques.

Dans cette sous-section, nous avons proposé une architecture permettant de calculer les différentes cartes de caractéristiques et de singularité nécessaires à notre système attentionnel en respectant un critère de rapidité de traitement. Cette architecture utilise des filtres approximés mais plus rapides que ceux utilisés dans [Itti 98, Frintrop 05a] (cf annexe D.3). En contrepartie de cette « perte de précision », le nombre de niveaux de résolution calculés est plus important (jusqu'à deux fois plus pour une image 800×600). La comparaison avec une vérité terrain humaine et avec d'autres modèles d'attention (présentée en section 3.2) montre que cette stratégie n'a pas d'impact négatif sur la plausibilité des prévisions de notre modèle.

Les cartes de singularité générées par le système visuel hiérachique décrit dans cette section ne sont que la première étape de notre modèle d'attention (cf figure 3.1.2). Celui-ci utilise ensuite le système attentionnel décrit dans la prochaine section pour mettre

ces cartes en compétition et ainsi calculer l'évolution du focus attentionnel.

3.1.4 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 présentés en section 2.3.3.1. Pour résoudre ce problème de compétition, de nombreuses solutions sont possibles. La plus courante est une approche neuromimétique, basée sur les réseaux de neurones [VanRullen 03, Spratling 04, Deco 04, Tsotsos 05b]. 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 95, Lesser 98]. On remplace alors les réseaux de neurones des modèles distribués par un système d'équations différentielles représentatives du comportement à reproduire [Vitay 05, Fix 08]. 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 3.1.17.

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.

3.1.4.1 Construction d'un système proies / prédateurs 2D

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 03a, Murray 03b]. Pour notre modèle, nous nous sommes inspirés de [Lesser 98] 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 [Volterra 28] 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 ;

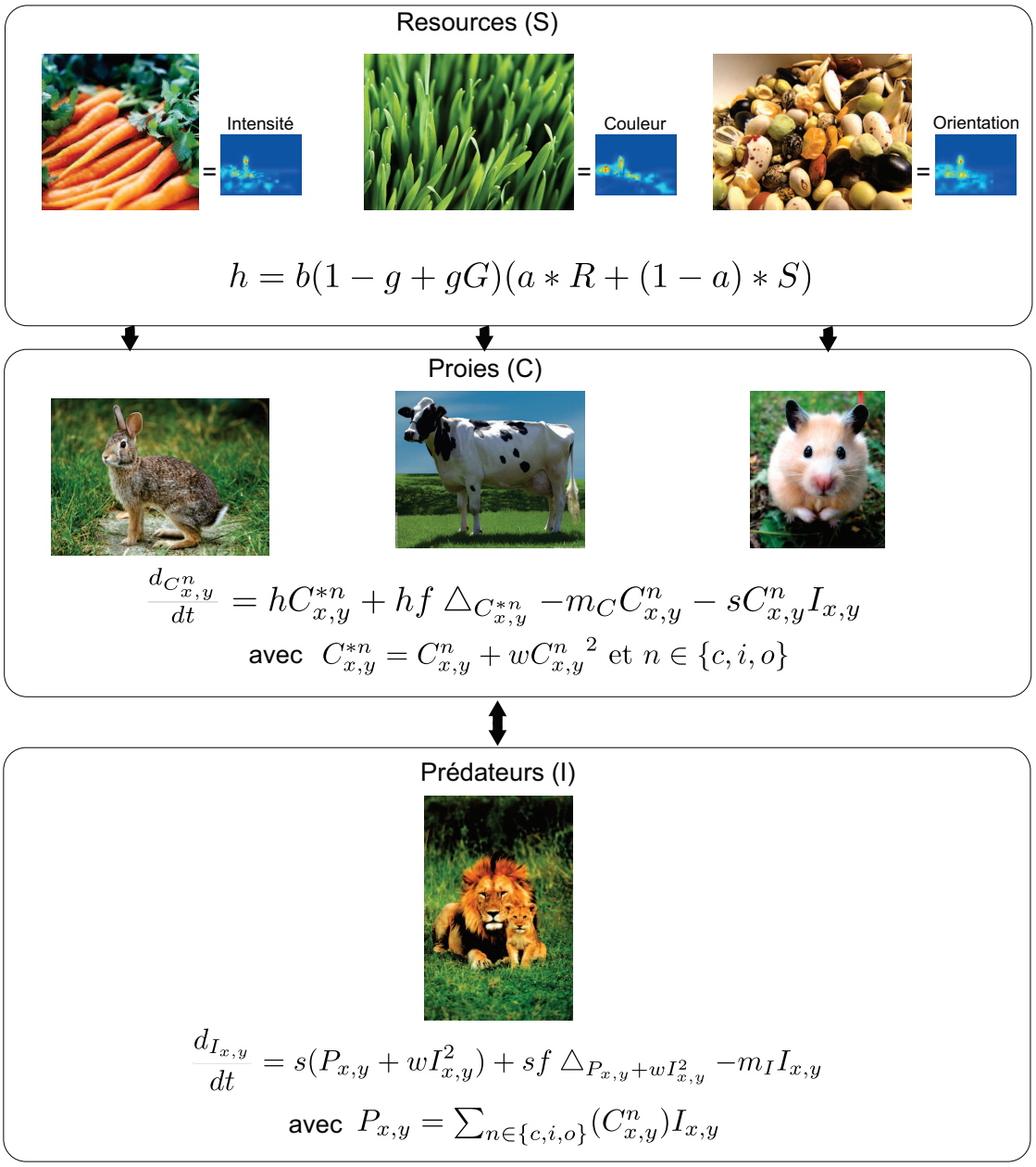


FIGURE 3.1.17: Architecture du système proies / prédateurs de fusion de cartes de singularité.

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 :

$$\begin{cases} \frac{dC}{dt} & = bC - s^{'}CI \\ \frac{dI}{dt} & = sCI - m_I I \end{cases} \tag{3.1.16}$$

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 :

$$\begin{cases} \frac{d_{C_{x,y}}}{dt} & = bC_{x,y} + f \triangle_{C_{x,y}} - m_C C_{x,y} - sC_{x,y}I_{x,y} \\ \frac{d_{I_{x,y}}}{dt} & = sC_{x,y}I_{x,y} + sf \triangle_{P_{x,y}} - m_I I_{x,y} \end{cases} \tag{3.1.17}$$

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 :

$$\begin{cases} \frac{d_{C_{x,y}}}{dt} & = b(C_{x,y} + w(C_{x,y})^2) + f \triangle_{C_{x,y}} - m_C C_{x,y} - sC_{x,y}I_{x,y} \\ \frac{d_{I_{x,y}}}{dt} & = s(C_{x,y}I_{x,y} + w(I_{x,y})^2) + sf \triangle_{P_{x,y}} - m_I I_{x,y} \end{cases} \tag{3.1.18}$$

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 sous-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 3.1.18 permet d'obtenir les équations décrites dans le paragraphe suivant.

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

Pour chacune des quatre cartes de singularité (couleur, intensité, orientation et mouvement) calculées, l'équation de la matrice des proies C est régie par l'équation 3.1.19, directement dérivée de l'équation 3.1.18 :

$$\frac{d_{C^n_{x,y}}}{dt} = hC^{*n}_{x,y} + hf \triangle_{C^{*n}_{x,y}} - m_C C^n_{x,y} - sC^n_{x,y} I_{x,y} \tag{3.1.19}$$

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 3 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 :

$$h = b(1 - g + gG)(a * R + (1 - a) * SM_n)(1 - e) \tag{3.1.20}$$

- *la singularité SM_n de l'image* (avec $n \in \{c, i, o, m\}$) calculée *via* le système visuel décrit en sous-section 3.1.3, et 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 07] 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 05a, Tatler 07]. 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 (regroupée 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 3.1.11.

L'évolution de la matrice I des prédateurs consommant ces 3 types de proies est régie par l'équation 3.1.21 :

$$\frac{d_{I_{x,y}}}{dt} = s(P_{x,y} + wI_{x,y}^2) + sf \triangle_{P_{x,y}+wI_{x,y}^2} - m_I I_{x,y} \tag{3.1.21}$$

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. On aura donc :

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

avec (x_{max}, y_{max}) les coordonnées du maximum de $P_{x,t}(t)$, $Seuil_{Hysteresis}$le seuil d'hystérésis et $Focus(t)$ les coordonnées du focus d'attention à l'itération courante.

3.1.4.3 Valeur par défaut des paramètres du système

Nous avons dans un premier temps déterminé empiriquement un jeu de paramètres par défaut (tableau 3.1), permettant un équilibre du système en l'absence de toute image.

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

TABLE 3.1: Paramètres par défaut du système proies / prédateurs

Ces paramètres représentent des valeurs « raisonnables » permettant d'obtenir un système à l'équilibre. Une étude plus détaillée de la stabilité du système est présentée en sous-section 3.3.1.

Afin de pouvoir ajuster plus précisément ces paramètres, il faut connaître leur influence sur le comportement du système. Pour cela, nous avons besoin :

- de propriétés à observer : que peut-on observer pour qualifier le comportement du système ? On pense en premier lieu à ses performances comparativement au modèle humain ; cette validation est l'objet de la prochaine section. D'autres propriétés, également intéressantes, sont abordées en section 3.3 ;
- un ensemble de mesures : que peut-on mesurer pour pouvoir étudier les propriétés du système ? Nous disposons en sortie du système des positions (x, y) du point de focalisation attentionnel pour différents instants, mais d'autres mesures (directes ou dérivées) peuvent être utiles.

Nous abordons ces points en section 3.3, en préambule de la présentation des résultats des mesures permettant de cerner le rôle de ces différents paramètres dans le comportement de notre système dynamique d'attention visuelle.

3.2 Fidélité au modèle humain

L'objectif de notre modèle est de piloter un système de vision afin qu'il puisse effectuer son analyse de scène plus efficacement. Nos contraintes consistent donc à fournir un système d'attention visuelle rapide, dont les propriétés peuvent être modifiées afin de satisfaire des contraintes particulières (exploration de toute la scène, focalisation sur les éléments les plus saillants, reproductibilité, etc.). Ainsi, nous ne sommes pas dans le cas d'un modèle essayant de reproduire fidèlement le comportement du système attentionnel humain : nous avons d'ailleurs pris des libertés sur le calcul des cartes de caractéristiques et de singularité afin d'en accélérer le traitement. Cependant, il parait nécessaire de comparer notre modèle à une vérité terrain humaine, afin de vérifier la cohérence de son comportement. De même, le comparer avec d'autres modèles computationnels permet de mesurer sa distance avec des approches destinées, elles, à être plus fidèles.

La comparaison avec le système attentionnel humain peut être réalisée de deux façons :

- en observant le parcours oculaire de sujets lorsqu'ils observent des images : c'est une évaluation objective, basée sur la mesure du comportement ;

– en demandant au sujet d'évaluer la qualité du résultat : c'est une évaluation subjective, basée sur la connaissance et les *a priori* du sujet concernant ce qui est intéressant dans une image.

Ces deux méthodologies d'évaluation sont complémentaires. Nous les avons appliquées à l'étude de notre modèle.

3.2.1 Évaluation objective

La propriété des systèmes d'attention visuelle la plus couramment étudiée est sa fidélité avec une vérité terrain humaine. Pour effectuer cette comparaison, il faut : capturer les fixations d'un certain nombre de sujets, obtenir des données à l'aide d'un modèle d'attention et enfin, appliquer une méthodologie de comparaison. Nous décrivons ces trois étapes dans la suite de cette sous-section.

3.2.1.1 Acquisition d'une vérité terrain

L'acquisition des mouvements oculaires des sujets est réalisée grâce à un oculomètre *(eye-tracker)*. Il en existe différentes sortes (figure 3.2.1), pouvant être installés sur la tête du sujet *(head-mounted)* ou « sur table » *(desktop),* de manière non invasive pour le sujet . Ils utilisent généralement le phénomène de réflexion cornéal, observé lorsque l'œil est illuminé par une source de lumière proche infrarouge, afin de capturer l'orientation du regard (à une fréquence située généralement entre 60 et 250 hertz). On pourra trouver plus de détails sur l'*eye-tracking* et ses différentes techniques dans [Ould Mohamed 07] et [Hansen 10].

FIGURE 3.2.1: Deux dispositifs d'eye-tracking. A gauche une version « *head-mouted* ». A droite une version « *desktop* ».

Le procédé d'*eye tracking* mesure en continu le mouvement des yeux : les données acquises contiennent les saccades et fixations oculaires effectuées par le sujet. Cependant, dans le cadre de l'étude de l'attention, seules les fixations sont intéressantes. Les données sont alors filtrées à l'aide de différentes techniques [Salvucci 00] afin de ne garder que ces dernières.

3.2.1.2 Données fournies par notre modèle d'attention

Contrairement à certains modèles ne produisant que des cartes de saillance [Ma 03, Le Meur 05a, Bruce 09], notre système fournit à chaque pas de temps la position 2D du focus d'attention. Bien que de même nature, ces données ne sont pas directement comparables aux positions brutes capturées *via* un dispositif d'*eye-tracking* (sans filtrage). En effet, ces dernières sont le résultat d'un système couplé : mécanisme attentionnel + mécanisme moteur. Les contraintes physiques liées au mouvement des yeux impliquent des déplacements continus (saccades) entre chaque fixation. Notre système attentionnel ne simulant pas le mouvement des yeux, les données qu'il génère ne sont pas aussi « continues » que celles observées par un *eye-tracker*.

L'étude de la dynamique du système faisant l'objet d'une section dédiée (sous-section 3.3.4) nous nous focalisons, pour la comparaison avec le modèle humain, sur une étude statistique de leur similarité.

3.2.1.3 Comparaison

La comparaison des modèles computationnels avec le comportement attentionnel humain dépend généralement du type de modèle utilisé.

Dans le cas des modèles centralisés, on effectue une étude statistique en comparant la carte de saillance produite par le modèle, avec une *heat-map* générée à partir de fixations humaines. Cette *heat-map* est un résumé statistique, représentant (tout comme une carte de saillance) la probabilité de focalisation de l'attention en chaque point de l'image observée. Celle-ci est générée, pour un observateur o, grâce à la formule suivante :

$$HM_o(x,y) = \left(\sum_{i=1}^{N} (d_i \times \delta_{x,y}^{x_i,y_i}) \right) * g_{\sigma_x,\sigma_y}(x,y) \tag{3.2.1}$$

avec N le nombre de fixations effectuées sur l'image, (x_i, y_i) les coordonnées, d_i la durée de la fixation i, g_{σ_x,σ_y} une gaussienne 2D d'écarts types $\sigma_x$et σ_y, et

$$\delta_j^i = \begin{cases} 1 & si\ i = j \\ 0 & sinon \end{cases}$$

Le filtrage par la gaussienne g_{σ_x,σ_y} est nécessaire pour prendre en compte deux phénomènes : l'œil ne focalise pas sur un point mais sur une zone plus large (déterminée par la taille de la fovéa) ; la précision des appareils de mesure est généralement limitée (au mieux 0.25° à 0.5°).

Pour un nombre O d'observateurs, on aura alors :

$$HM(x,y) = \frac{1}{O} \sum_{o=1}^{O} HM_o(x,y) \tag{3.2.2}$$

Notre modèle d'attention ne produit pas de carte de saillance. Il nous faut donc générer une *heatmap*. Cependant, comme nous l'avons précisé plus haut, la notion de saccades et de fixations n'est pas applicable à notre système. On génère donc la *heatmap* pour une simulation s de la manière suivante :

$$HM_s(x,y) = \left(\sum_{i=1}^{N}(\delta_{x,y}^{x_i,y_i})\right) * g_{\sigma_x,\sigma_y}(x,y) \tag{3.2.3}$$

avec N le nombre de focalisations effectuées sur l'image, (x_i, y_i) les coordonnées de la focalisation i, la signification des autres termes restant inchangée.

Si l'on effectue S simulations consécutives on aura alors :

$$HM(x,y) = \frac{1}{S}\sum_{s=1}^{s} HM_s(x,y) \tag{3.2.4}$$

On peut s'interroger sur l'intérêt d'effectuer plusieurs simulations pour une même image et un même paramétrage du système. Cela est cependant nécessaire dans le cas de notre modèle car celui-ci n'est pas déterministe. De même que chaque observation d'un même observateur donnera des résultats différents, chaque simulation donnera un résultat dont la variabilité dépendra du paramétrage du système. Ce phénomène est étudié en section 3.3.2.

Dans toutes les expérimentations liées à la fidélité au modèle humain (objective ou subjective), les *heatmaps* générées à partir de notre modèle ont été filtrées avec une gaussienne g_{σ_x,σ_y}dont les écart types ont été définis de la manière suivante :

$$\sigma_x = \sigma_y = 0.3 \times foveaSize \times \max(W,H)$$

avec W et H la largeur et la hauteur de l'image traitée et $foveaSize = 0.15$. On filtre donc les *heatmaps* avec une gaussienne aux valeurs non nulles sur environ 15% du plus grand coté de l'image (puisqu'une gaussienne s'étale sur environ 3 σ).

3.2.1.4 Mesures

Nous pouvons disposer pour les expérimentations humaines et la modélisation computationnelle, d'un ensemble de fixations ou de focalisations ainsi que leur *heatmaps* associées. Pour les comparer, nous avons choisi les trois méthodes les plus courantes parmi le panel des solutions proposées dans la littérature [Shic 06, Aziz 09b, Carmi 06, Koch 09, Peters 05].

Corrélation croisée

La corrélation croisée *(cross-correlation)* normalisée entre les *heatmaps* est une méthode simple permettant de caractériser la similitude entre deux images. Elle est définie

pour deux *heatmaps* HM_1 et HM_2 par la relation suivante :

$$CrossCor_{HM_1,HM_2} = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \frac{\left(HM_1(x,y) - H\bar{M}_1\right)\left(HM_2(x,y) * H\bar{M}_2\right)}{\sigma_{I_1}\sigma_{I_2}} \tag{3.2.5}$$

avec W et H les largeur et hauteur des images, $H\bar{M}_1$ et $H\bar{M}_2$ leur moyenne et σ_{I_1}et σ_{I_2}leur écart type.

Plus la valeur de corrélation sera élevée, meilleur sera le modèle considéré.

Divergence de Kullback Leiber

Les *heatmaps* représentant des probabilités de focalisation, Tatler [Tatler 05] et Le Meur [Le Meur 05a] proposent d'utiliser la théorie de l'information afin de caractériser la dissimilarité entre les deux distributions de probabilité P_M et P_G telles que :

$$P(x,y) = \frac{HM(x,y)}{\sum_{i,j} HM(i,j)} \tag{3.2.6}$$

On utilise alors la divergence de Kullback Leibler entre la distribution de la vérité terrain P_G et celle du modèle considéré P_M, définie par :

$$KL = \sum_{x=1}^{W} \sum_{y=1}^{H} P_G(x,y) \log_2 \left(\frac{P_G(x,y)}{P_M(x,y)}\right) \tag{3.2.7}$$

$$KL = \sum_{x=1}^{W} \sum_{y=1}^{H} P_G(x,y) \log_2 \left(P_G(x,y)\right) - P_G(x,y) \log \left(P_M(x,y)\right) \tag{3.2.8}$$

D'un point de vue théorie de l'information, cette mesure définit le nombre de bits additionnels nécessaires pour coder un message répondant à la distribution P_M avec un code ayant été défini pour la distribution de référence P_G. Une faible divergence de Kullback Leibler indiquera un modèle proche du comportement humain.

Notons que le calcul de la divergence de Kullback Leibler pose un problème lorsque $P_G(x,y) = 0$ ou $P_M(x,y)$. Une technique couramment utilisée dans ce cas est d'ajouter une très faible valeur ϵ au calcul des logarithmes. On aura alors :

$$KL = \sum_{x=1}^{W} \sum_{y=1}^{H} P_G(x,y) \log_2 \left(P_G(x,y) + \epsilon\right) - P_G(x,y) \log \left(P_M(x,y) + \epsilon\right)$$

Compte tenu de la taille des *heatmaps*, dont la dimension n'excède jamais 1000×1000, la valeur de ϵ a été définie à $\epsilon = 10^{-6}$.

Normalized scanpath salience

Une mesure plus spécifique à l'étude de l'attention est introduite par Peters [Peters 05] afin de caractériser la similitude entre un ensemble de fixations et une carte de saillance / *heatmap*. La *normalized scanpath salience (NSS)* est définie comme suit :

$$NSS = \frac{1}{N} \sum_{i=1}^{N} HM^{'}(x_i, y_i)$$

avec N le nombre de fixations, (x_i, y_i) les coordonnées de la fixation i, et $HM^{'}(x, y) = \frac{HM(x,y) - \bar{HM}}{\sigma_{HM}}$ la *heatmap* normalisée à une moyenne de 0 et un écart type de 1.

Cette mesure représente la moyenne des valeurs de la carte de saillance normalisée pour tous les points de fixations de la vérité terrain. Compte tenu de la normalisation de la carte de saillance, une valeur de NSS supérieure à zéro suggère une prédiction meilleure qu'un modèle aléatoire. Une valeur inférieure à zéro indique que le modèle effectue des prédictions inverses à celles attendues.

3.2.1.5 Résultats et interprétation

Nous avons étudié deux aspects distincts du modèle :

- l'influence de différents paramétrages sur la fidélité des résultats. En faisant varier la valeur d'un paramètre différent à chaque expérimentation, nous pouvons caractériser l'influence de celui-ci sur les performances. Nous avons ainsi testé plusieurs valeurs pour 6 paramètres différents, produisant 16 configurations différentes présentées dans le tableau 3.2.
- la comparaison des performances de notre système avec celles de modèles naïfs (saillance constante ou gaussienne centrée) ou de l'état de l'art. Pour cela nous avons utilisé les cartes de saillance générées à partir de modèles mis à disposition librement (NVT de Laurent Itti [Itti 98] et AIM de Neil Bruce et John Tsotsos [Bruce 09]) ou fournies par les auteurs (pour le modèle d'Oliver Le Meur [Le Meur 05a] [3]).

Bases d'images et de vérité terrain

Tous les résultats présentés dans cette section ont été générés à partir des images et de la vérité terrain issues de deux bases :

- Le Meur : fournie par Olivier Le Meur, cette base est composée de 27 images naturelles couleur, dont un sous-ensemble provient d'une base utilisée par Laurent Itti. Les différentes images de cette base ainsi que des informations complémentaires sont présentées en annexe E.2.

3. Nous remercions Olivier Le Meur d'avoir accepté de nous fournir les résultats de son modèle pour les différentes bases utilisées lors des expérimentation.

– Bruce : fournie par Neil D.B. Bruce, cette base comporte 120 images naturelles couleur. Des exemples d'images et plus de précisions sur cette base sont présentés en annexe E.1.

La configuration par défaut utilisée pour notre algorithme lors des expérimentations correspond aux paramètrages définis dans le tableau 3.1, sans détermination du focus d'attention par hystérésis ni pseudo filtre rétinien. Toutes les images ont été redimensionnées à une largeur de 256 pixels (la hauteur étant adaptée proportionnellement). Enfin, les simulations utilisant notre modèle ont été effectuées 20 fois sur chaque image (l'équivalent de 20 observateurs [4]), pendant une durée de 300 itérations (correspondant à 10 secondes de visualisation).

Pour le calcul de la corrélation croisée et de la divergence de Kullback Leiber, nous avons utilisé les *heatmaps* fournies par les auteurs (elles n'ont pas été recalculées à partir des données de fixation brutes). Pour le calcul de la *normalized scanpath salience* nous avons utilisé les données de fixations brutes.

Influence des différents paramètres

Le tableau 3.2 résume les performances de notre modèle pour les différents jeux de paramètres. On peut en tirer les conclusions et interprétations suivantes :

– L'utilisation du pseudo-flou rétinien améliore les résultats de 22%, confirmant l'intuition que la perception à résolution variable de la scène engendrée par la structure de la rétine, a une influence sur le processus attentionnel et qu'elle doit donc être prise en compte dans les modèles computationnels.
– L'introduction d'un biais central est un facteur déterminant pour les performances du modèle (+45% en faisant passer sa valeur de 0.25 à 0.5). Il permet en effet de prendre en compte différents phénomènes :
  – le biais dû au protocole d'acquisition de la vérité terrain (centrage du regard au centre de l'écran entre deux images).
  – le biais dû aux images utilisées pour les expérimentations : lors d'une prise de vue photographique, le sujet est généralement au centre de la photographie.
– La diffusion est nécessaire au bon fonctionnement du système proies / prédateurs (-14% si on la supprime), mais l'augmenter significativement n'améliore pas les performances (+1%).
– Le filtrage par hystérésis pour la sélection du maximum de la carte des prédateurs (et donc du focus d'attention) n'apporte aucun gain significatif.
– Le système proies / prédateurs doit être bruité pour fonctionner correctement. Peu ou pas de bruit (0 ou 0.25) fait chuter les performances en moyenne de plus de 90%, car dans ce cas, seuls les éléments très saillants sont visités par le système. On obtient alors un comportement type d'une allocation monotropique de l'attention

4. Notre modèle n'étant pas déterministe, chaque simulation génère un résultat différent. Cependant, puisque le paramétrage reste identique pour les 20 simulations, on pourrait également considérer qu'il s'agit de 20 observations d'un même observateur.

[Murray 05] [5]. L'ajout de bruit permet au contraire de réduire l'importance des éléments les plus saillants et ainsi de permettre à d'autres zones de l'image d'être parcourues. Le comportement attentionnel devient polytropique [Murray 05].
– Le terme quadratique (*feedback* positif) est à manipuler avec précaution. Une trop forte valeur a le même effet que l'absence de bruit : seuls les éléments les plus saillants sont visités, interdisant ainsi une exploration plus complète de l'image.

	CrossCorrelation		KullbackLeiblerDivergence		NormalizedScanpathSaliency		Gain moyen	Gain moyen
	Bruce	LeMeur	Bruce	LeMeur	Bruce	LeMeur	CC +NSS	KLD
Default	0,35	0,30	1,80	1,76	0,95	0,56	0%	0%
Default+RetinalFilter	0,43	0,38	1,61	1,40	1,17	0,73	26%	16%
Default+CentralBias=0	0,20	0,14	2,33	2,29	0,57	0,27	-48%	-29%
Default+CentralBias=0_25	0,48	0,44	1,57	1,49	1,29	0,82	41%	14%
Default+CentralBias=0_5	0,55	0,53	1,92	1,66	1,49	1,01	67%	0%
Default+Diffusion=0	0,33	0,23	2,06	2,26	0,96	0,47	-11%	-21%
Default+Diffusion=0_5	0,35	0,31	1,83	1,72	0,94	0,59	2%	0%
Default+Diffusion=0_125	0,35	0,29	1,77	1,70	0,95	0,55	-1%	3%
Default+Hysteris=0_1	0,36	0,30	1,81	1,75	0,96	0,57	1%	0%
Default+Hysteris=0_5	0,36	0,31	1,86	1,84	0,95	0,57	1%	-4%
Default+Hysteris=0_25	0,36	0,31	1,83	1,77	0,95	0,58	1%	-1%
Default+Noise=0	0,17	0,06	4,32	4,77	0,49	0,13	-65%	-155%
Default+Noise=0_25	0,16	0,07	4,21	4,49	0,48	0,15	-63%	-144%
Default+Noise=0_75	0,46	0,44	1,61	1,17	1,25	0,83	39%	22%
Default+Noise=1	0,27	0,35	1,89	1,30	0,68	0,64	-5%	11%
Default+QuadraticTerm=0	0,35	0,28	1,85	1,80	0,96	0,54	-2%	-2%
Default+QuadraticTerm=0_003	0,16	0,04	2,99	3,52	0,45	0,08	-70%	-83%

TABLE 3.2: Comparaison avec la vérité terrain : influence des différents paramètres.

Comparaison avec d'autres modèles

Nous avons calculé les scores de corrélation, divergence de Kullback Leiber et NSS pour 7 algorithmes, dont :
– deux modèles naïfs : « AllEqual » correspondant à une carte de saillance à la valeur moyenne constante (ici 127). Ce modèle considère que tous les points de l'image ont une saillance égale. « Gaussian » correspond à une gaussienne ajustée pour couvrir l'ensemble de l'image. Dans ce modèle, on considère le centre de l'image comme plus saillant que la périphérie.
– le modèle d'Olivier Le Meur, avec « normalisation cohérente » des cartes.
– le modèle AIM de Neil Bruce, dont le code Matlab a été légèrement modifié afin d'obtenir des cartes de saillance normalisées entre 0 et 255.
– le modèle NVT de Laurent Itti, dont les cartes de saillance ont été utilisées telles quelles, ou agrandies et filtrées avec un filtre gaussien. En effet, les cartes de saillance générées par le NVT sont quatre fois plus petites que l'image d'origine et présentent des pics possédant un faible étalement. Leur allure est alors bien différente des

5. Murray suggère que l'allocation monotropique de l'attention est un mécanisme permettant d'expliquer certains troubles autistiques.

heatmaps de la vérité terrain. Pour les en rapprocher, nous les avons agrandies puis filtrées avec un filtre gaussien afin d'étaler les pics de saillance sur une zone plus large de l'image.
- un modèle, « OurDirectMaps », n'utilisant que la partie hiérarchique de notre modèle d'attention. Les cartes de singularité obtenues en section 3.1.3.2 sont normalisées avec la même méthode que les cartes de caractéristiques (information propre) puis additionnées. On obtient alors une carte de saillance permettant d'évaluer la qualité intrinsèque de la partie hiérarchique de notre modèle.
- un modèle « Default + Retina », qui est notre modèle avec ses paramètres par défaut et le pseudo filtre rétinien activé. Cette configuration a été choisie car c'est un bon compromis entre neutralité des paramètres et fidélité au modèle humain.
- un modèle « Optimal », qui est notre modèle proies-prédateurs avec la combinaison de paramètres la plus optimale au sens des résultats obtenus lors de l'estimation de l'influence des différents paramètres. Ce n'est en aucun cas un optimal global sur l'ensemble des combinaisons possibles. Le modèle « Optimal » correspond à la configuration « Default+RetinalFilter+CentralBias=0.5+Noise=0.75 ».

Les cartes de saillance des modèles « AIM », « NVT » et « Le Meur » ont été calculées avec leurs paramètres par défaut.

La figure 3.2.2, permet d'observer sur l'exemple de l'image « Parrots », les cartes de saillance générées ainsi que leur score respectif pour les différentes mesures utilisées. Le tableau 3.2.3 permet quant à lui, d'avoir un aperçu des performances des différents algorithmes sur l'ensemble des deux bases de test. On peut effectuer le bilan suivant :
- bien que très simple, la corrélation croisée est un indicateur qui semble fiable. En effet, les scores de corrélation croisée semblent cohérents avec la normalised scanpath salience, spécialement conçue pour comparer les cartes de saillance avec une vérité terrain ;
- la divergence de Kullbach-Leibler est sensible à la normalisation des cartes : la carte allEqual obtient de meilleurs scores que le modèle d'Itti, alors que celui-ci obtient des scores supérieurs avec les deux autres indicateurs ;
- la divergence de Kullback-Leibler est très sensible à la taille des cartes de saillance et *heatmaps*, ainsi qu'à la distribution des niveaux de gris. Ainsi le modèle d'Itti obtient de mauvais scores car ses petites cartes de saillance ne mettent en valeur que quelques pics très localisés ;
- le modèle « All Equal » donne logiquement de mauvais résultats pour tous les indicateurs ;
- le modèle gaussien obtient de très bons scores. Cela peut être dû, comme nous l'avons déjà évoqué, au biais lié aux types d'images utilisés, aux protocoles d'expérimentation mais également peut-être à un réel biais attentionnel en faveur du centre des scènes observées.
- Notre modèle de calcul direct « OurDirectMaps » obtient des performances semblables à celles du modèle d'Itti (filtré). Les simplifications effectuées sur les filtres

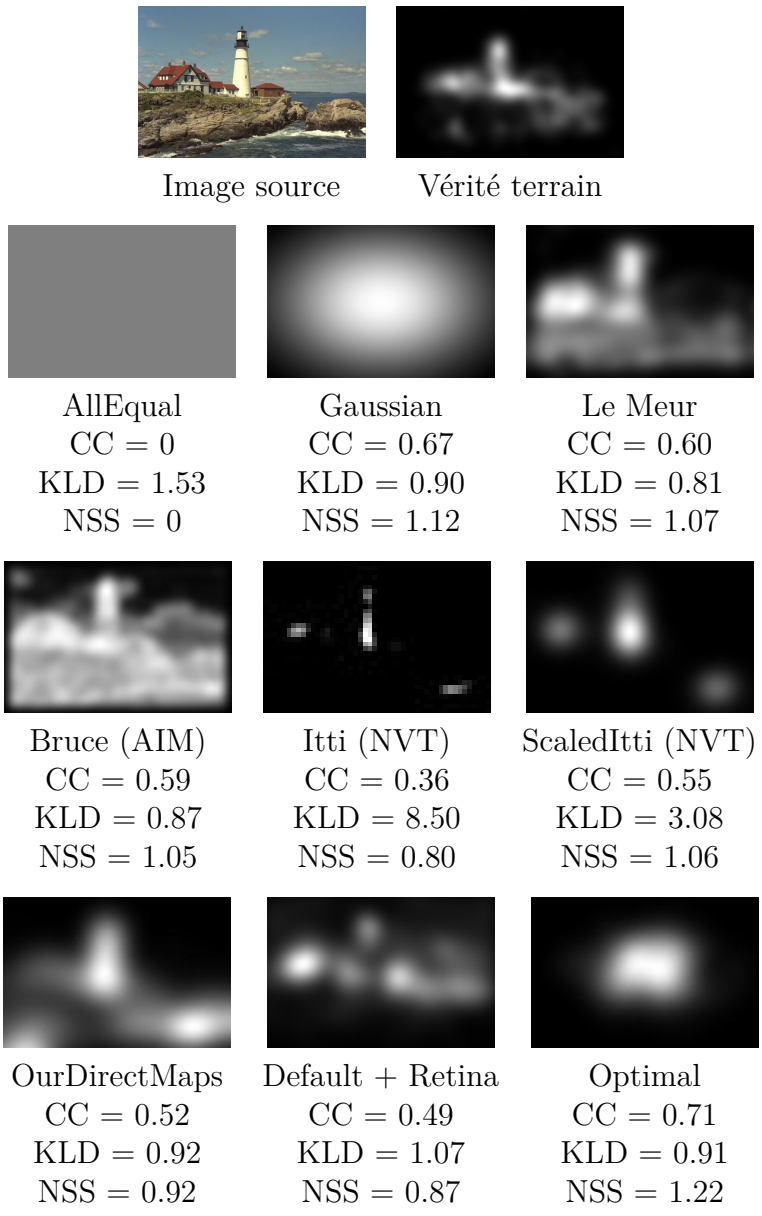


FIGURE 3.2.2: Exemples de cartes de saillance / *heatmaps* générées par les différents modèles testés pour l'image « Parrots ». Les scores affichés en-dessous des cartes correspondent à la corrélation croisée (CC), la divergence de Kullback Leibler (KLD) et la *normalized scanpath salience* (NSS).

pour accélérer les calculs sont donc compensées par le plus grand nombre de résolutions pris en compte ainsi que la plus grande précision du filtrage centre-périphérie (effectué au sein d'un même niveau de pyramide et non entre les niveaux).
- Notre modèle « Default + Retina » se comporte très bien et obtient même de très bons scores de corrélation et NSS pour la base Bruce.
- Notre modèle optimal permet d'obtenir les scores les plus élevés. Cependant, la *heatmap* alors obtenue est proche d'un modèle gaussien centré, légèrement modulé par la saillance de différents éléments de la scène. Bien que performant (en terme d'évaluation), ce paramétrage ne sera certainement pas le plus adapté si l'on connecte notre modèle à un système de vision.

3.2.2 Évaluation subjective [6]

Comme nous l'avons vu plus haut, la comparaison de différents algorithmes permettant de générer des cartes d'attention visuelle (*heatmaps*), est réalisée en comparant ces cartes aux données de vérité terrain obtenues à partir d'expériences oculométriques. Cependant, cette méthode d'évaluation est complexe à mettre en place et souffre de divers biais et défauts :
- Le biais sémantique dû à la nécessité de présenter les images plusieurs secondes au sujet afin d'obtenir suffisamment de fixations pour que la *heatmap* obtenue soit statistiquement représentative.
- Un biais de préférence centrale dû en partie à la nécessité de refocaliser l'attention du sujet sur un point déterminé (généralement le centre de l'écran) avant de présenter une nouvelle image.
- La difficulté de collecte des données. Le matériel nécessaire est généralement coûteux et il doit être calibré fréquemment et correctement. La réalisation des mesures nécessite un nombre de participants suffisants, devant se déplacer dans la salle d'expérimentations.

Puisque ces biais ne peuvent être totalement évités, nous avons expérimenté une méthode alternative ne nécessitant pas l'utilisation d'un oculomètre. La méthode proposée fait appel à une évaluation subjective de la qualité des cartes de saillance générées par différents algorithmes. Comme nous l'avons déjà évoqué, les deux types d'expérience ne sont pas de même nature : dans le cas de l'étude subjective, on effectue une observation d'un processus conscient (caractériser la qualité prédictive des cartes), alors que dans le cas des évaluations objectives, le processus observé est inconscient (on demande seulement aux sujets d'observer les images).

La mesure subjective est donc une mesure complémentaire permettant de vérifier que les modèles testés correspondent à ce qu'attendent les utilisateurs. Aucun modèle théorique ne reproduit correctement le fonctionnement de notre système attentionnel. Il peut donc être intéressant de voir si parmi ces modèles imparfaits, certains seront mieux

6. Ces travaux ont fait l'objet d'une publication dans [Perreira Da Silva 10a]

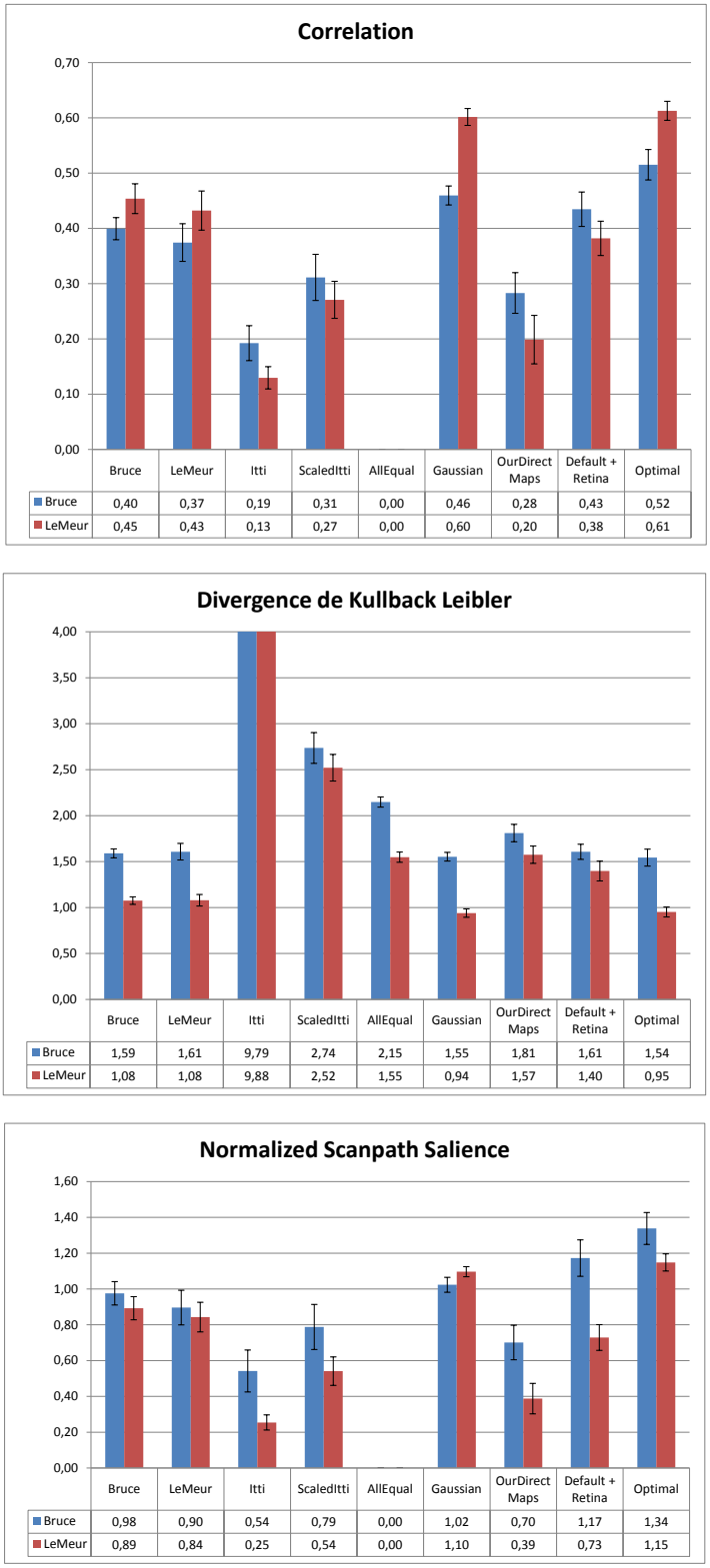


FIGURE 3.2.3: Comparaison avec la vérité terrain : performances des différents modèles.

acceptés que d'autres.

3.2.2.1 Mesures

Algorithmes

Nous avons comparé les 6 modèles d'attention visuelle suivants :

- Le modèle de référence de Laurent Itti [Itti 98] avec ses paramètres par défaut (dont la normalisation itérative). L'implémentation utilisée est celle fournie au travers du *NVT*. Les tests étant effectués sur des images fixes, les seules cartes de singularité utilisées sont celles d'intensité, couleur et orientation.
- Notre modèle d'attention proies / prédateurs avec ses paramètres par défaut (sauf le taux de natalité des proies, fixé à 0,05 au lieu des 0,1 afin de prendre en compte le plus fort contraste des cartes générées par le modèle d'Itti) utilisant les cartes de singularité de [Itti 98] sans normalisation (afin de conserver un maximum d'information)
- Notre modèle d'attention proies / prédateurs avec ses paramètres par défaut, utilisant nos cartes de singularité.
- Notre modèle d'attention / prédateurs avec ses paramètres par défaut, utilisant nos cartes de singularité avec simulation du flou rétinien.
- Un modèle aléatoire de génération de fixations avec préférence centrale. Ce modèle est en fait notre système proies / prédateurs avec un terme a aléatoire égal à 1 et donc un terme $1 - a$ d'attache aux données nul.
- Un modèle aléatoire de génération de fixation sans préférence centrale. Ce modèle est le même que celui cité précédemment, mais avec également un terme g de préférence centrale nul.

Un exemple des cartes générées par les différents algorithmes est présenté en figure 3.2.4.

Base d'images et participants

Les bases « Bruce » et « Le Meur » fournissent une vérité terrain importante pour l'évaluation objective des algorithmes, mais peu utile pour leur évaluation subjective. De plus, elle ne permettent pas d'étudier le comportement des algorithmes en fonction de différentes catégories d'images. Ainsi, une nouvelle base de 48 images a été créée afin de réaliser les expérimentations subjectives. Celles-ci ont été collectées parmi les images respectant la licence *« creative commons attribution »* du site de partage de photos en ligne Flickr. Six catégories de 8 images ont été réalisées :

- Abstrait
- Animaux
- Ville
- Fleurs
- Paysages
- Portraits

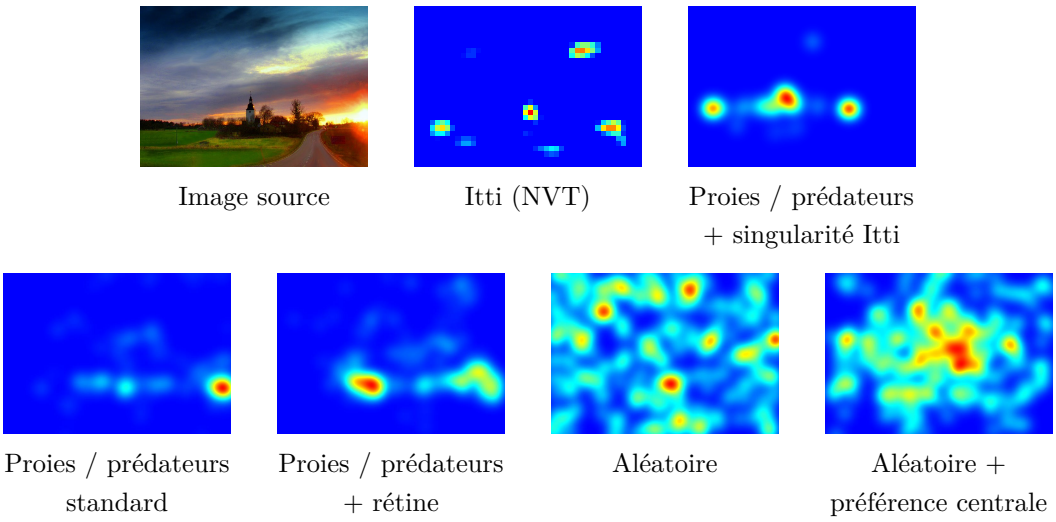


FIGURE 3.2.4: Exemples de cartes de saillance et *heatmaps* générées pour une image de la catégorie paysages.

Des exemples d'images de cette base sont consultables en annexe E.3. Les expérimentations ont été réalisées auprès de 16 personnes dont 11 hommes et 5 femmes, âgés de 21 à 57 ans.

Protocole

Nous avons demandé aux participants de l'expérimentation de visionner un ensemble de 288 couples d'images. Chaque couple était composé d'une image de référence, choisie au hasard parmi les 48 images de la base d'expérimentation, ainsi qu'une carte d'attention visuelle générée pour cette image par l'un des 6 algorithmes testés. Pour chacun de ces couples, le participant devait donner une note entre 0 (pas du tout fidèle) et 3 (très fidèle) afin d'évaluer le potentiel de la carte d'attention à représenter les zones de l'image qui pourraient attirer l'attention de la majorité des observateurs (figure 3.2.5). Aucune limite de temps n'était imposée, il était cependant conseillé de ne passer que 2 à 3 secondes sur chaque image afin que l'expérience ne dure pas plus d'une quinzaine de minutes. Enfin, les dix premiers couples présentés n'étaient pas pris en compte pour l'exploitation des résultats, permettant à chaque participant de se familiariser avec l'exercice demandé.

La notation des participants étant assez hétérogène, nous avons normalisé les 288 notes de chacun d'eux à une moyenne et un écart type fixes (de valeur respective 1,5 et 1). Ainsi la notation de chacun est respectée, tout en garantissant un ensemble de données plus homogènes.

FIGURE 3.2.5: Interface de l'application de notation des cartes d'attention.

3.2.2.2 Résultats et interprétation

Lorsque l'on compare les performances moyennes des différents algorithmes indépendammment des différentes catégories (figure 3.2.6), on constate que l'utilisation de notre système de fusion proies / prédateurs permet d'améliorer sensiblement la plausibilité des cartes d'attention générées. Nous notons également que, couplée avec notre système proies / prédateurs, l'utilisation de nos cartes de singularité apporte un gain de performance intéressant, malgré les nombreuses simplifications effectuées lors de leur génération. Par contre (et contrairement aux résultats obtenus lors de l'évaluation objective), aucune différence notable de performance n'est observée lors de l'utilisation du système de simulation de flou rétinien. Il est également surprenant de constater que le modèle aléatoire avec préférence centrale semble obtenir des performances de même niveau que le modèle proies / prédateurs appliqué aux cartes de singularité de l'algorithme d'Itti. En réalité ces résultats sont très variables en fonction de la catégorie des images observées. Enfin, le modèle purement aléatoire obtient les plus mauvaises performances.

Sur les 6 catégories de notre base de tests (figure 3.2.7), on constate que les résultats sont assez variables :

- sur les images des classes *portraits*, *abstrait* et *animaux*, le modèle aléatoire à préférence centrale obtient, soit les meilleures performances (*portrait*s), soit des performances tout à fait correctes (*animaux*). Ceci s'explique en partie par le biais dû au photographe, qui a tendance à centrer son sujet dans l'image (pour les catégories *portraits* et *animaux*). Pour la catégorie *abstrait* ;
- les images de la catégorie *abstrait*, étant particulièrement difficiles à noter, aucune différence significative n'apparaît entre les différents modèles. Le modèle aléatoire à préférence centrale obtient toutefois le meilleur score : il semble que les participants ont accepté ce choix central comme la moins mauvaise solution ;
- l'utilisation des systèmes proies / prédateurs apporte un gain significatif pour toutes

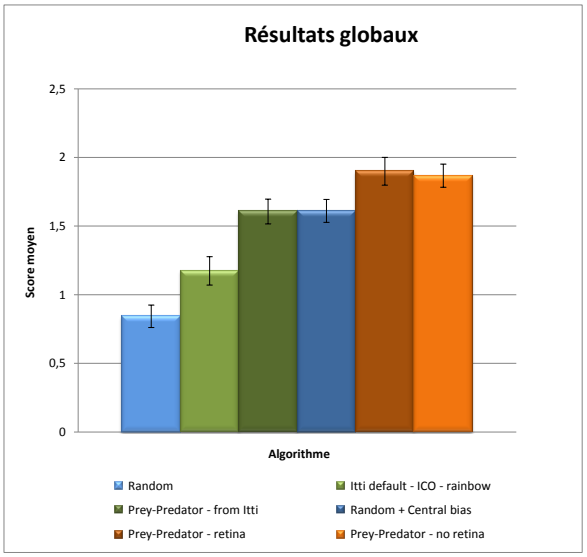


FIGURE 3.2.6: Résultats globaux de l'évaluation subjective. Les barres noires représentent l'intervalle de confiance à 95%.

les catégories, sauf pour les images de paysage où l'algorithme d'Itti se comporte particulièrement bien. Dans ce cas, les performances obtenues par nos modèles sont légèrement supérieures à celles du modèle d'Itti, mais il est difficile de conclure avec certitude compte tenu des intervalles de confiance relativement importants pour cette catégorie.

3.2.3 Bilan

Les expériementations menées permettent de constater que le paramétrage du modèle peut influer de manière importante sur son comportement et donc sa fidélité avec le système visuel humain. Les principaux paramètres permettant d'améliorer la plausibilité du modèle sont : le biais central, le bruit, et l'activation du pseudo-flou rétinien. Ainsi, si l'on accepte de rendre le modèle très peu reproductible et particulièrement dépendant du biais central, on peut alors dépasser les performances de fidélité de tous les modèles testés. Il faut cependant relativiser ces bon scores, car le modèle ainsi obtenu, très proche du modèle gaussien, fournit des résultats peu spécifiques à chaque image.

En utilisant un paramétrage plus « raisonable », les résultats des études objectives et subjectives montrent que notre système est encore capable de rivaliser avec des modèles aux performances établies (Itti, Le Meur et Bruce). Ces résultats valident également les simplifications effectuées pour améliorer l'efficacité computationnelle de la partie hiérarchique de notre modèle puisque celles-ci ne semblent pas réduire sa plausibilité. Ils confirment également que notre système d'attention proies / prédateurs permet une compétition efficace entre les différentes cartes de singularité.

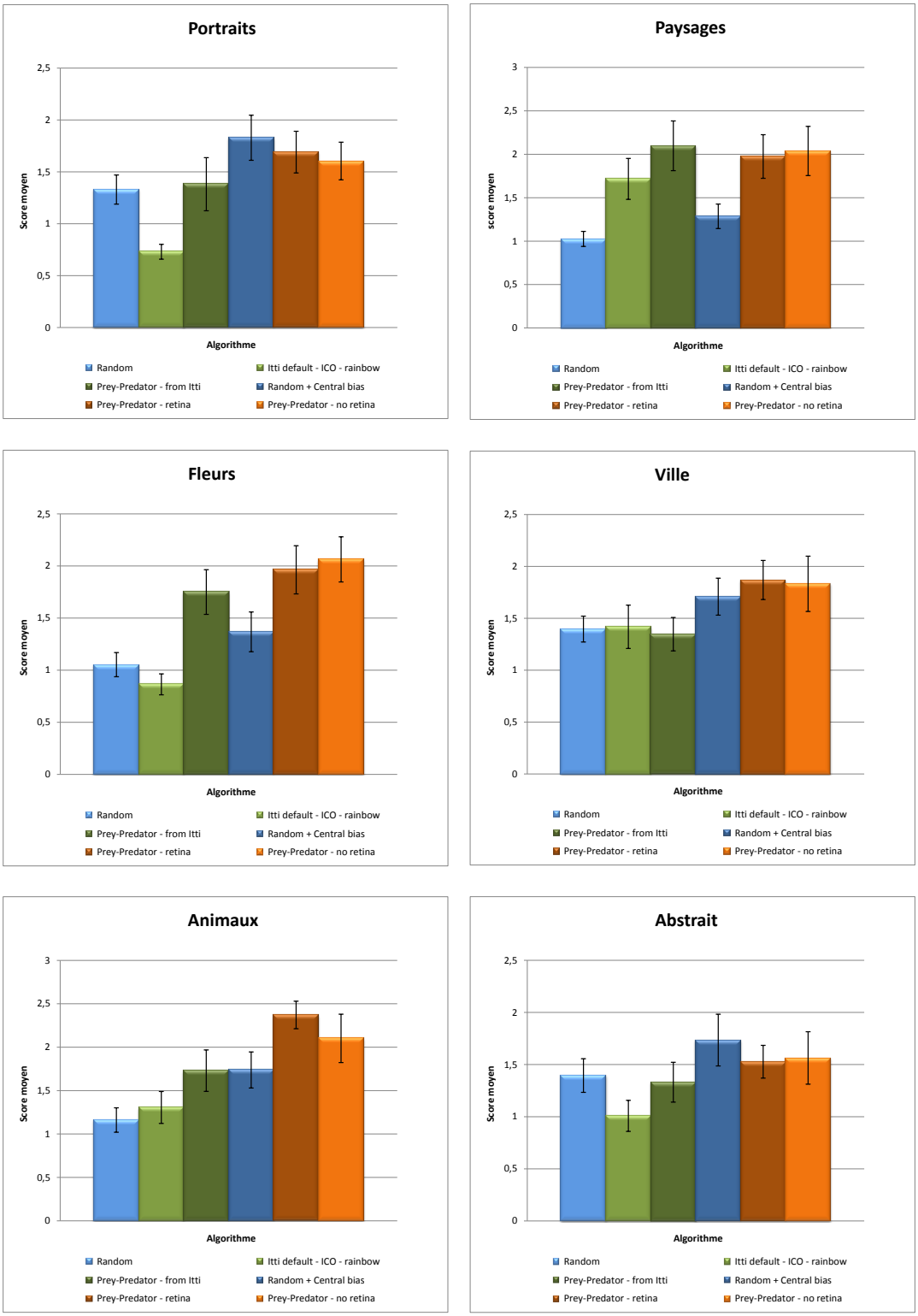


FIGURE 3.2.7: Évaluation subjective : résultats par catégories.

On pourra regretter que nos expérimentations soient limitées à l'étude des performances de notre modèles sur des images statiques. Celui-ci est tout à fait capable de gérer des flux vidéo (nous en verrons quelques exemples dans la partie 4.4, dédiée aux applications), mais l'étude statique était un sujet suffisamment riche pour laisser l'étude dynamique en perspective de nos travaux.

Ces perspectives incluent également l'application du système proies / prédateurs non plus sur les cartes de singularité, mais sur les cartes de caractéristiques. Comme nous l'avons déjà évoqué, le comportement obtenu serait alors certainement plus riche. Néanmoins, une telle mise en œuvre complexifie beaucoup le système : son exécution temps réel serait alors difficile à obtenir. Nous avons donc dans un premier temps privilégié l'étude d'un système plus simple.

Dans sa version actuelle, notre modèle dispose déjà de nombreux paramètres dont l'étude sur le comportement du système fait l'objet de la prochaine section.

3.3 Propriétés et mesures

L'étude d'un système dynamique complexe (possédant généralement de multiples paramètres) est une tâche délicate. Pour analyser le comportement de notre système proies / prédateurs, nous nous sommes inspirés de la démarche proposée par Brodu [Brodu 07], consistant en l'utilisation de la « méthode scientifique expérimentale », que l'on peut résumer en trois points :

1. émettre une hypothèse. Par exemple : le paramètres X de notre modèle à une influence sur une propriété Y ;
2. définir une ou des expériences permettant de vérifier cette hypothèse. Celà nécessite entre autre, de choisir les mesures à effectuer sur le système étudié ;
3. analyser les résultats obtenus afin de réaliser des prédictions sur le comportement du système.

On peut ainsi étudier l'influence des différents paramètres du système en fonction des propriétés que l'on souhaite observer.

Pour mener à bien cette étude, il convient de définir un ou des niveaux d'observation (microscopique ou macroscopique) ainsi que les propriétés à observer. Dans notre cas, nous étudions des propriétés macroscopiques, puisque c'est le comportement global du système (compétition entre les différentes sources d'attention) qui nous intéresse. Les propriétés étudiées, dérivées de notre cahier des charges et de l'étude des modèles existants et de leurs applications, sont les suivantes :

– la stabilité : les valeurs calculées par le système dynamique restent-elles bornées lorsque l'on fait varier ses différents paramètres ?

- la reproductibilité : les systèmes dynamiques discrets pouvant avoir un comportement chaotique, quelle est l'influence des différents paramètres (et en particulier le bruit) sur la variabilité des résultats générés ?
- l'exploration de l'espace : dans quelle mesure une variation des paramètres peut-elle influer la stratégie d'exploration de la scène de notre modèle attentionnel ?
- la dynamique du système : comment peut-on influer sur la réactivité du système, et en particulier sur son inertie lors de changements rapides dans la scène ?

Dans cette section, nous décrivons l'influence des principaux paramètres de notre modèle sur ces propriétés. L'ensemble des expérimentations réalisées a été effectué sur les deux bases « Le Meur » et « Bruce » présentées lors de la description de l'évaluation objective de notre modèle (section 3.2.1.5). Sauf indications contraires, le paramétrage utilisé est celui décrit en section 3.2.1.5.

3.3.1 Stabilité

Les équations de Volterra-Lotka ne produisent un modèle stable [7] que dans une plage de paramètres déterminée [Idema 05]. 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 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.

3.3.1.1 Mesures

A chaque pas de temps, le système dynamique met à jour l'ensemble des valeurs des cartes de proies et de prédateurs. Compte tenu de leur nombre, il n'est pas raisonnable d'étudier chacune de ces variations locales individuellement. Nous proposons alors d'observer la valeure de la moyenne de la carte des prédateurs :

$$Moy_I(t) = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} I_{x,y}(t) \tag{3.3.1}$$

et des différentes cartes de proies :

$$Moy_{C^n}(t) = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} C^n_{x,y}(t) \tag{3.3.2}$$

avec $n \in \{c, i, o, m\}$.

Si ces mesures oscillent entre des valeurs bornées, alors notre système sera stable.

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

3.3.1.2 Résultats et interprétation

Pour que le système attentionnel ait un comportement correct tout au long de son fonctionnement, il est nécessaire que le système d'équations proies / prédateurs sous-jacent soit stable. Pour cela, il nous faut trouver une plage de valeurs des principaux paramètres du système permettant d'obtenir ce régime. Pour y parvenir, nous avons dans un premier temps déterminé empiriquement un jeu de valeurs permettant d'obtenir un système oscillant lorsque une image noire est présentée au système (fonctionnement « à vide »). Ce jeu de paramètres « par défaut » (présenté en sous-section 3.1.4.3) fournit la base des valeurs autour desquelles nous avons étudié la stabilité du système dynamique.

Les paramètres étudiés sont au nombre de quatre :
- b le facteur de natalité des proies ;
- m_C le facteur de mortalité des proies ;
- s le facteur de prédation (qui est aussi le facteur de natalité des prédateurs) ;
- m_I le facteur de mortalité des prédateurs et w le facteur de *feedback* positif.

Nous avons vérifié que les autres paramètres permettent d'affiner le comportement du système, sans rôle majeur sur sa stabilité.

En faisant varier individuellement chacun des quatre paramètres autour des valeurs d'équilibre par défaut, nous révélons les comportements « limites » du système. Le tableau 3.3 résume les plages de valeurs pour une stabilité optimale et les comportements limites associés. Notons que les valeurs données sont approximatives, le changement de comportement du système étant progressif.

En observant ce tableau et les figures 3.3.1 et 3.3.2, on constate une multitude de comportements limites différents. Cependant, en théorie seuls deux comportements limites existent :
- l'accroissement vers l'infini des proies et la mort des prédateurs ;
- la mort des proies et prédateurs

Il ne peut y avoir de situation où toutes les proies sont mortes et seuls les prédateurs survivent, puisque ceux-ci se nourrissent des proies. On peut alors s'interroger sur l'origine de ces comportements additionnels ? Ils sont en réalité générés par quelques contraintes ajoutées lors de l'implémentation afin de renforcer la stabilité du système :
- 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 (figure 3.3.1c).
- 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 données en section

Paramètre X	Valeur par défaut	Limite min	Limite max	Comportement pour X=zéro	Comportement pour X très grand ($X \geq 10$)
b	0.007	0.006	0.013	Extinction (b)	Saturation oscillante double (h)
m_C	0.3	0.3	0.36	Ecrêtage(c)	Extinction (b)
s	0.025	0.017	0.05	Saturation des proies / mort des prédateurs (d)	Saturation oscillante des prédateurs (e)
m_I	0.5	0.1	1.5	Saturation des prédateurs (f)	Saturation double (g)
w	0.001	0	0.003	Stable (a)	Saturation oscillante double (h)

TABLE 3.3: Paramètres influant sur la stabilité du système : plages de valeurs optimales et comportements limites. Les lettres entre parenthèses renvoient au sous-figure de la figure 3.3.1.

3.1.4 deviennent alors :

$$\begin{aligned}
\frac{d_{C^n_{x,y}}}{dt} &= \left(1 - \frac{C^n_{x,y}}{Max_{population}}\right)\left(hC^{*n}_{x,y} + hf\triangle_{C^{*n}_{x,y}}\right) - m_C C^n_{x,y} - sC^n_{x,y} I_{x,y} \\
\frac{d_{I_{x,y}}}{dt} &= \left(1 - \frac{I_{x,y}}{Max_{population}}\right)\left(s(P_{x,y} + wI^2_{x,y}) + sf\triangle_{P_{x,y}+wI^2_{x,y}}\right) - m_I I_{x,y}
\end{aligned}$$

On peut alors expliquer l'ensemble des comportements limites :
- la saturation des proies et la mort des prédateurs si le taux de prédation n'est pas suffisant (figure 3.3.1d) ;
- la saturation oscillante des prédateurs lorsque le taux de prédation est élevé. Les proies à peine nées sont consommées, faisant grimper le taux de prédateurs temporairement, celui-ci redescend ensuite également temporairement à cause du nombre élevé de prédateurs dont la mortalité devient alors élevée (figure 3.3.2e).
- la saturation des prédateurs, maintenant possible lorsque le taux de mortalité des prédateurs est nul, puisque le nombre de proies n'est jamais nul (figure 3.3.2f).
- la saturation double : lorsque le taux de mortalité des prédateurs est très haut, la population de proies augmente localement jusqu'à saturation aux endroits où il n'y a plus de prédateurs. Parallèlement, grâce à la composante aléatoire du système la population des prédateurs réussit également à atteindre la valeur de saturation à d'autres endroits (complémentaires des premiers). Le système ainsi

formé n'évolue plus.
- la saturation oscillante double : grâce à une forte hausse du nombre de proies à chaque pas de temps, on observe le même phénomène de saturation locale que pour la saturation double. Cependant, au lieu de se stabiliser, le système oscille globalement entre deux conformations complémentaires localement saturées .

Le système est dans l'ensemble assez tolérant sur la variation de ses paramètres. Même en dehors des plages limites indiquées, le comportement est dégradé mais le système est toujours utilisable. Pour tous les paramètres, seules des valeurs extrêmes ($\geq$10) permettent d'obtenir les comportements limites présentés ci-dessus. La seule exception est le paramètre de *feeback* positif w ; celui-ci intervenant sur le carré de la population de prédateurs $I_{x,y}$ une valeur importante entraîne très rapidement la saturation du système.

3.3.2 Reproductibilité

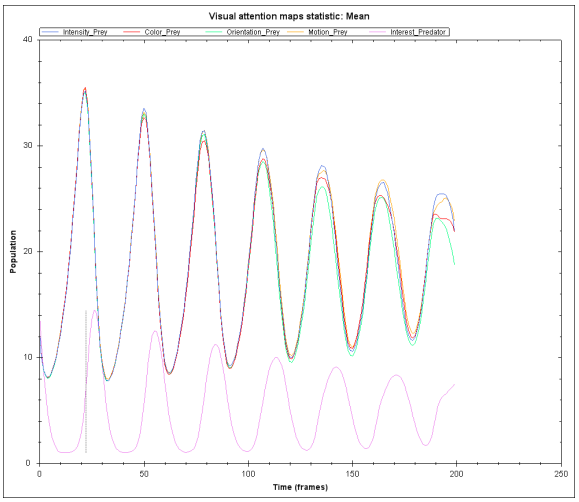
De par sa discrétisation et l'introduction d'une carte aléatoire lors du calcul du facteur de croissance globale h, notre système attentionnel est non déterministe. Ce mode de fonctionnement est intéressant car il permet de simuler la variabilité naturelle observée lorsque l'on mesure plusieurs fois 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.

3.3.2.1 Mesures

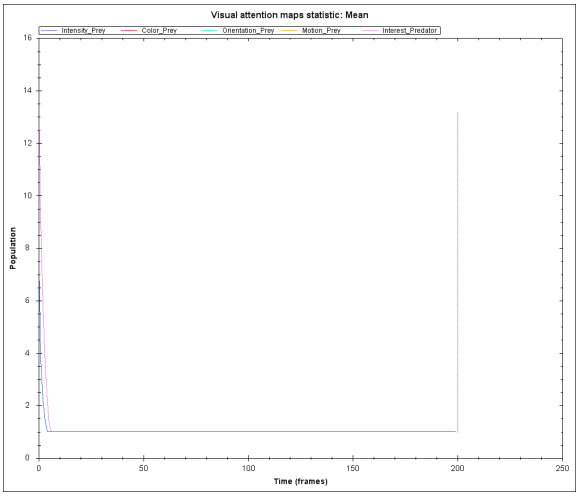
Quelle est l'influence de l'introduction de bruit dans le système sur sa reproductibilité ? Quelle est la part de variation, en fonction de la valeur du facteur de curiosité a ? Quels autres paramètres jouent un rôle dans la variabilité du système ? Pour étudier ce phénomène, nous proposons d'utiliser les mêmes mesures que pour la comparaison de notre modèle avec la vérité terrain : la corrélation croisée normalisée, la divergence de Kullback-Leibler et la *normalized scanpath salience*. Ces mesures de similarité / dissimilarité seront alors calculées entre plusieurs simulations effectuées avec les mêmes paramètres et sur la même image.

3.3.2.2 Résultats et interprétation

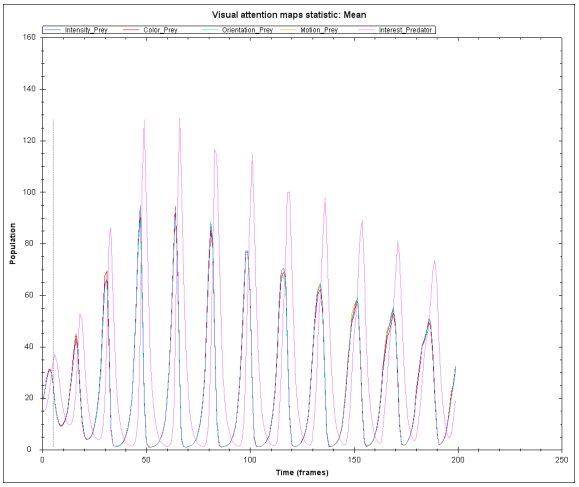
Les résultats présentés dans le tableau 3.4 permettent de confirmer que le bruit a est un facteur clé dans le contrôle de la reproductibilité du système. En présence d'un bruit nul, le système reproduit le même comportement à chaque simulation. Cependant, ce comportement est alors très peu varié et le système ne visite qu'une petite partie



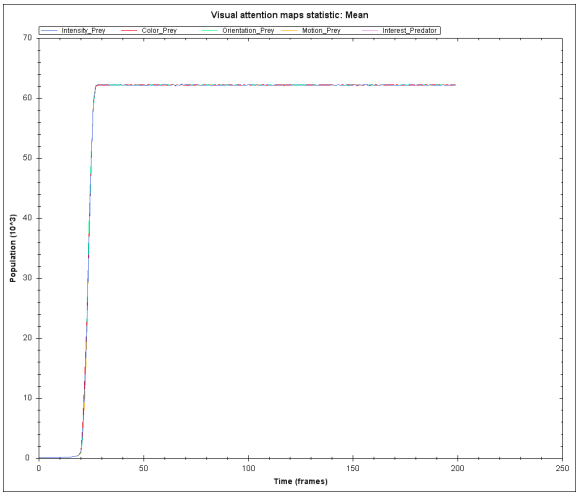
(a) Régime stable : différentes populations de proies et prédateurs oscillent sans saturation.



(b) Extinction des proies et prédateurs.

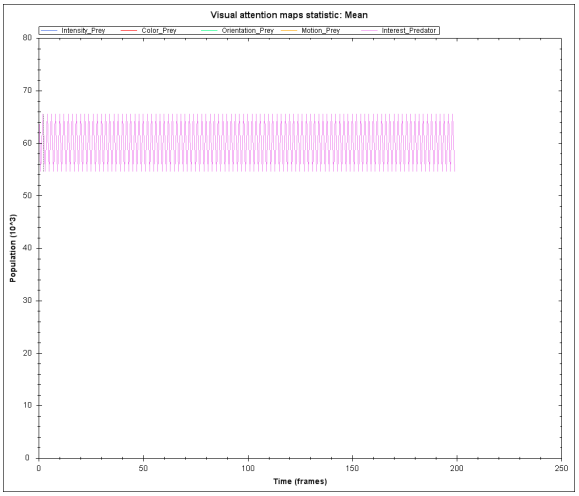


(c) Ecrêtage. Les populations de proies et prédateurs atteignent périodiquement le seuil minimal de population (ici fixé à 1). Les courbes ne sont alors plus symétriques.

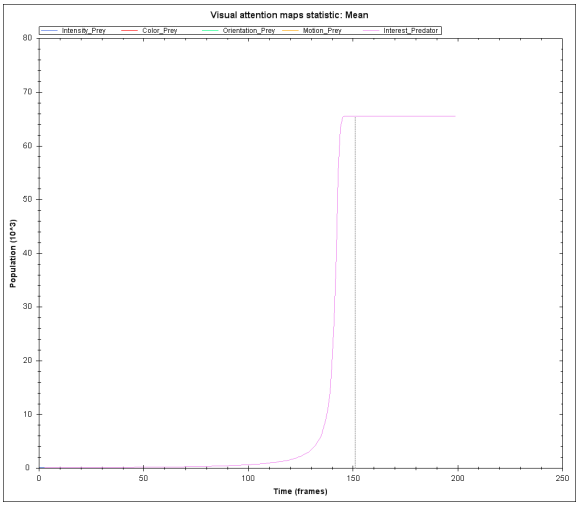


(d) Saturation des proies, mort des prédateurs.

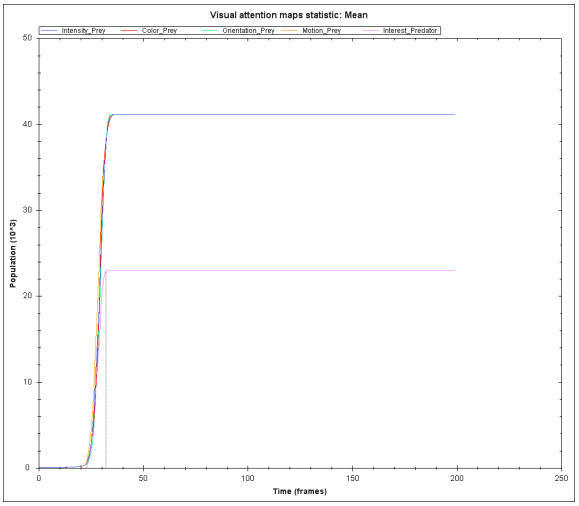
FIGURE 3.3.1: Comportement du système pour différents paramétrages (partie 1).



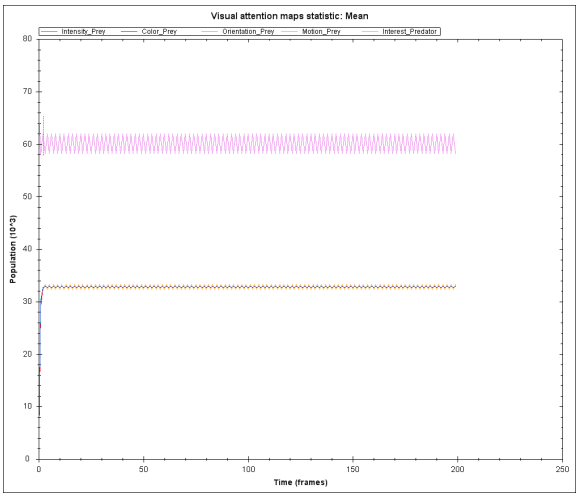
(e) Saturation oscillante des prédateurs.



(f) Saturation des prédateurs.



(g) Saturation double.



(h) Saturation oscillante double.

FIGURE 3.3.2: Comportement du système pour différents paramétrages (partie 2).

de l'image, ce qui le rend peu cohérent avec le modèle humain (c.f. tableau 3.2). En augmentant le bruit raisonnablement (par exemple à sa valeur par défaut de 0.5), le système reste globalement cohérent d'une simulation à l'autre (corrélation d'au moins 0.85) tout en ayant un comportement beaucoup plus riche. Au-delà le système explore très bien tout l'espace de la scène (nous le verrons dans la sous-section suivante), mais son comportement devient plus variable d'une simulation à l'autre. Cette variabilité est cependant toute relative : même pour un bruit maximal, la différence moyenne entre les différentes *heatmaps* générées reste faible (KLD moyen d'au maximum 0.40) ; par contre les *heatmaps* sont peu corrélées entre elles (corrélation croisée minimum de 0.31 et NSS minimum de 0.13).

L'utilisation du pseudo filtre rétinien diminue légèrement (ou plus fortement si l'on considère le NSS moyen) la reproductibilité. Ceci s'explique par le fait que lorsque ce filtre est utilisé, le système a tendance à effectuer des focalisations plus proches les unes des autres (les zones éloignées du point de focalisation étant plus floues). Ceci a pour conséquences une plus grande sensibilité aux conditions de départ : si le premier point de focalisation varie (ce qui est possible en fonction du facteur de bruit) l'ensemble des focalisations suivantes sera affecté.

A noter les scores incohérents du système pour une valeur du facteur de *feedback* positif (*quadratic-term*) importante. Cela s'explique par le fait qu'avec cette valeur de *feedback*, on assiste à une saturation double oscillante (voir section 3.3.1.2). La *heatmap* ainsi générée a une distribution assez atypique à laquelle réagissent différemment la corrélation croisée, la divergence de Kullback Leibler et la *normalized scanpath salience*.

D'autres paramètres comme l'hystérésis ou la diffusion ont également une influence sur la reproductibilité, mais celle-ci reste limitée face à celle du bruit.

3.3.3 Exploration de l'espace

Comme nous l'avons vu au chapitre 2, le rôle du système d'attention visuelle humain est 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, présentée dans la sous-section suivante.

	Cross Correlation		Kullback LeiblerDivergence		Normalized Scanpath Saliency		Gain moyen	Gain moyen
	Bruce	LeMeur	Bruce	LeMeur	Bruce	LeMeur	CC +NSS	KLD
Default	0,85	0,89	0,26	0,25	1,82	2,35	0%	0%
Default+RetinalFilter	0,82	0,84	0,29	0,30	1,35	1,74	-16%	-16%
Defautl+CentralBias=0	0,81	0,86	0,27	0,27	1,72	2,24	-5%	-5%
Defautl+CentralBias=0_25	0,92	0,93	0,20	0,20	2,16	2,57	10%	22%
Defautl+CentralBias=0_5	0,95	0,95	0,14	0,14	2,64	2,83	21%	46%
Defautl+Diffusion=0	0,97	0,97	0,12	0,12	3,42	3,69	42%	54%
Defautl+Diffusion=0_5	0,86	0,89	0,28	0,27	1,76	2,20	-3%	-7%
Defautl+Diffusion=0_125	0,87	0,90	0,23	0,23	2,00	2,52	5%	9%
Defautl+Hysteris=0_1	0,85	0,89	0,28	0,27	1,84	2,39	0%	-7%
Defautl+Hysteris=0_5	0,83	0,88	0,36	0,33	1,90	2,52	2%	-36%
Defautl+Hysteris=0_25	0,84	0,88	0,31	0,29	1,88	2,43	1%	-19%
Defautl+Noise=0	1,00	1,00	0,00	0,00	5,05	5,38	84%	100%
Defautl+Noise=0_25	0,98	0,98	0,07	0,09	5,07	5,05	80%	68%
Defautl+Noise=0_75	0,58	0,64	0,29	0,27	0,44	0,57	-53%	-10%
Defautl+Noise=1	0,31	0,32	0,40	0,37	0,13	0,13	-79%	-52%
Defautl+QuadraticTerm=0	0,84	0,88	0,34	0,32	2,12	2,72	7%	-30%
Defautl+QuadraticTerm=0_003	0,77	0,83	0,80	0,75	7,15	10,23	153%	-204%

TABLE 3.4: Reproductibilité des fixations pour différentes simulations : influence des paramètres.

3.3.3.1 Mesures

A chaque pas de temps de la simulation, lorsque le focus d'attention change, nous « reconstruisons » incrémentallement 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 (figure 3.3.3).

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. On a alors :

$$M_{flou}(x,y,t) = \max(M_{flou}(x,y,t-1), N_{Levels} - \min\left(\frac{dist(x,y,x_f,y_f)}{BlurSize}, N_{Levels}\right) \quad (3.3.3)$$

avec (x_f, y_f), les coordonnées de la focalisation courante, N_{Levels} la valeur de flou maximum et $dist(x_1,y_1,x_2,y_2)$ la distance euclidienne entre les points $(x1, y1)$ et (x_2, y_2). N_{Levels} est déterminée en fonction de la largeur W et de la hauteur H de l'image source :

$$NbLevels = ceiling\left(\log_2\left(\min(W,H)\right)\right) \quad (3.3.4)$$

$ceiling(x)$ représentant l'arrondi de la valeur x à l'entier supérieur.

$BlurSize$ est la taille de la zone de netteté maximale. Pour nos expérimentations, celle-ci a été fixée à 10% du plus grand côté de l'image. Cette valeur correspond à la taille de la fovéa (environ 2 degrés) si l'on considère que l'image observée par notre modèle occupe 20 degrés de son champ visuel (image d'un mètre de large, observée à trois mètres de distance).

L'image « reconstruite » I_R est ensuite générée à partir de l'image source I_S par convolution avec un filtre boite (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 boite (moyenneur) carré de taille s, et $s(x, y) = 2^{N_{Levels} - M_{flou}(x,y)}$ la fonction calculant la taille du filtre en fonction du masque de flou.

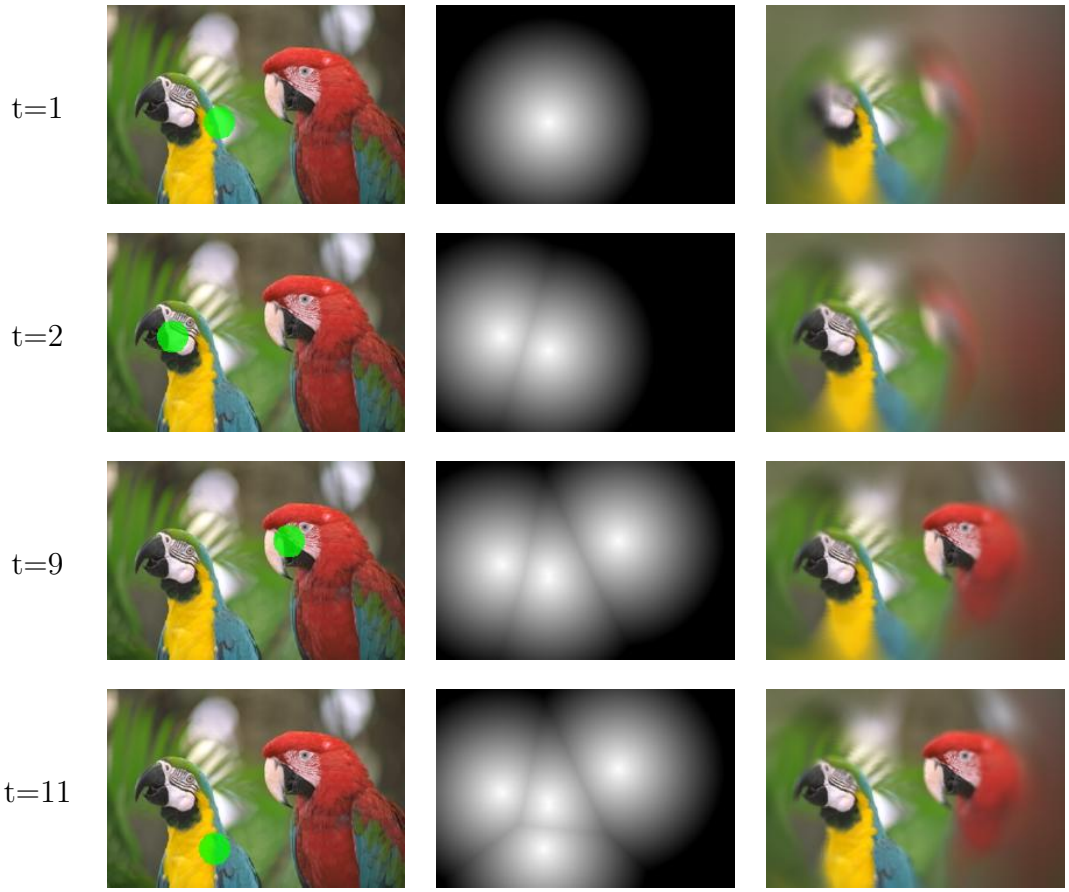


FIGURE 3.3.3: Exemple de « reconstruction » de l'image d'origine à partir des différentes focalisations. Colonne de gauche, image d'origine I_S et point de focalisation courant (disque vert). Colonne centrale, masque de flou M_{flou} appliqué à l'image (blanc : non flouté, noir : très flouté). Colonne de droite, image « reconstruite » I_R à partir 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 prin-

cipes 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 78]. 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 :

$$InformationRatio_{JPEG} = \frac{size\left(compress_{JPEG}(I_S)\right)}{size\left(compress_{JPEG}(I_R)\right)} \tag{3.3.5}$$

$$InformationRatio_{PNG} = \frac{size\left(compress_{PNG}(I_S)\right)}{size\left(compress_{PNG}(I_R)\right)} \tag{3.3.6}$$

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

Nous estimons également la quantité d'information restant à acquérir par une simple moyenne de la valeur absolue des différences entre les deux images.

$$MissingInformation = \frac{1}{W \times H} \sum_{x-1}^{W} \sum_{y=1}^{H} \mid I_S(x,y) * I_R(x,y) \mid \tag{3.3.7}$$

Ces trois méthodes permettent de mesurer l'exhaustivité du parcours de l'image, ainsi que le rythme d'acquisition de l'information.

3.3.3.2 Résultats et interprétation

Les tableaux de la figure 3.3.4 permettent de constater l'influence des différents paramètres du système sur ses capacités d'exploration de l'espace. Les trois indicateurs calculés donnent des résultats cohérents.

Contrairement à la section précédente, le facteur de bruit apporte ici un effet bénéfique : plus le bruit est important, plus l'espace est exploré rapidement. Il n'est cependant pas nécessaire d'utiliser une quantité de bruit importante pour bien explorer l'espace. La valeur par défaut ($a = 0.5$) donne des résultats très proches de ceux obtenus avec un bruit de 1 ; ceci est valable en début de simulation ($t = 50$ [8]) au milieu ($t = 150$) ou en

8. Afin d'éviter des lourdeurs dans la formulation du temps de simulation, nous utilisons le raccourci « t=X » pour signifier « après X itérations de notre modèle ».

fin de simulation ($t = 300$).

De fortes valeurs de préférence centrale (≥ 0.5) ont également un rôle néfaste sur les capacités d'exploration de l'espace : un biais central fort interdit la visite des zones situées à la périphérie, rendant impossible le parcours de toute l'image.

Les autres paramètres ont une influence relativement faible sur les capacités d'exploration, qui restent proches de celles obtenues avec le paramétrage par défaut.

3.3.4 Réactivité / Dynamique

Si l'on veut pouvoir utiliser efficacement notre système pour traiter des flux vidéo en temps réel, on doit pouvoir caractériser le comportement dynamique de celui-ci. Nous proposons d'étudier deux propriétés importantes : son « temps de démarrage », et la durée moyenne entre deux « fixations ».

3.3.4.1 Mesures

Pour calculer des indicateurs liés à ces deux propriétés, nous disposons :
- des statistiques des différentes cartes du système proies / prédateurs (cf 3.3.1.2) ;
- de la position du focus d'attention à chaque pas de temps.

« Temps de démarrage » de l'algorithme

Puisque nous utilisons des équations différentielles pour faire évoluer notre système, son fonctionnement est itératif. A chaque pas de simulation, l'état suivant du système est calculé à partir du précédent. Lorsque le système est en fonctionnement « normal » (image fixe présentée depuis un temps suffisant ou séquence d'images sans changement de plan) la différence entre deux images sources consécutives est faible ou nulle. Le système n'a alors aucun mal à s'adapter à ces faibles fluctuations. A l'inverse, lorsque le système démarre (présentation initiale d'une image) ou est sujet à un changement important (changement de plan dans une vidéo) de nombreuses itérations seront nécessaires pour s'adapter aux modifications.

On définit le « temps de démarrage » du système $TpsDemarrage$ comme le temps nécessaire à l'obtention d'un régime oscillant stable. Celui-ci est estimé en mesurant le temps nécessaire pour atteindre la valeur moyenne $MoyDeMoy_I$ de la moyenne $Moy_I(t)$ de la carte des prédateurs (figure 3.3.5). Cette carte a été préférée aux cartes de proies, car elle constitue la dernière étape avant le calcul du focus d'attention.

Pour ne pas fausser le calcul, la moyenne est calculée à partir du point situé à $t = TpsDemarrage$. On utilise donc un algorithme itératif (algorithme 3.1) pour estimer

	Bruce			Le Meur		
	t=50	t=150	t=300	t=50	t=150	t=300
Default	0,732	0,881	0,931	0,696	0,840	0,891
Default+RetinalFilter	0,720	0,888	0,944	0,688	0,856	0,913
Defautl+CentralBias=0	0,722	0,894	0,946	0,692	0,863	0,909
Defautl+CentralBias=0_25	0,681	0,794	0,831	0,641	0,750	0,793
Defautl+CentralBias=0_5	0,606	0,679	0,708	0,594	0,663	0,697
Defautl+Diffusion=0	0,715	0,817	0,862	0,672	0,769	0,810
Defautl+Diffusion=0_5	0,717	0,876	0,928	0,670	0,825	0,886
Defautl+Diffusion=0_125	0,737	0,884	0,932	0,697	0,837	0,887
Defautl+Hysteris=0_1	0,726	0,870	0,925	0,682	0,821	0,877
Defautl+Hysteris=0_5	0,679	0,835	0,900	0,637	0,784	0,847
Defautl+Hysteris=0_25	0,700	0,859	0,915	0,664	0,811	0,868
Defautl+Noise=0	0,489	0,502	0,506	0,582	0,610	0,618
Defautl+Noise=0_25	0,632	0,664	0,673	0,586	0,625	0,638
Defautl+Noise=0_75	0,760	0,929	0,979	0,746	0,927	0,977
Defautl+Noise=1	0,791	0,924	0,977	0,781	0,924	0,979
Defautl+QuadraticTerm=0	0,705	0,858	0,903	0,672	0,817	0,867
Defautl+QuadraticTerm=0_003	0,713	0,817	0,849	0,674	0,739	0,760

(a) Ratio de compression JPEG.

	Bruce			Le Meur		
	t=50	t=150	t=300	t=50	t=150	t=300
Default	0,906	0,972	0,987	0,893	0,954	0,969
Default+RetinalFilter	0,902	0,976	0,991	0,891	0,959	0,975
Defautl+CentralBias=0	0,903	0,977	0,991	0,896	0,963	0,977
Defautl+CentralBias=0_25	0,868	0,925	0,942	0,863	0,914	0,931
Defautl+CentralBias=0_5	0,814	0,866	0,884	0,826	0,871	0,889
Defautl+Diffusion=0	0,873	0,925	0,945	0,878	0,917	0,935
Defautl+Diffusion=0_5	0,900	0,971	0,986	0,887	0,948	0,967
Defautl+Diffusion=0_125	0,908	0,971	0,986	0,895	0,953	0,968
Defautl+Hysteris=0_1	0,903	0,969	0,985	0,889	0,948	0,966
Defautl+Hysteris=0_5	0,875	0,954	0,978	0,862	0,933	0,956
Defautl+Hysteris=0_25	0,887	0,965	0,982	0,877	0,943	0,963
Defautl+Noise=0	0,803	0,822	0,826	0,810	0,823	0,827
Defautl+Noise=0_25	0,810	0,831	0,836	0,809	0,833	0,842
Defautl+Noise=0_75	0,934	0,989	0,998	0,934	0,985	0,995
Defautl+Noise=1	0,944	0,986	0,997	0,943	0,984	0,996
Defautl+QuadraticTerm=0	0,879	0,961	0,976	0,882	0,945	0,961
Defautl+QuadraticTerm=0_003	0,894	0,935	0,946	0,883	0,908	0,917

(b) Ratio de compression PNG.

	Bruce			Le Meur		
	t=50	t=150	t=300	t=50	t=150	t=300
Default	0,030	0,013	0,008	0,037	0,019	0,014
Default+RetinalFilter	0,031	0,013	0,007	0,038	0,018	0,011
Defautl+CentralBias=0	0,030	0,012	0,007	0,037	0,017	0,012
Defautl+CentralBias=0_25	0,037	0,022	0,017	0,047	0,030	0,025
Defautl+CentralBias=0_5	0,053	0,039	0,034	0,057	0,045	0,039
Defautl+Diffusion=0	0,032	0,019	0,014	0,040	0,026	0,021
Defautl+Diffusion=0_5	0,031	0,014	0,009	0,043	0,022	0,014
Defautl+Diffusion=0_125	0,029	0,013	0,008	0,037	0,019	0,014
Defautl+Hysteris=0_1	0,031	0,015	0,009	0,041	0,022	0,015
Defautl+Hysteris=0_5	0,038	0,018	0,012	0,047	0,025	0,018
Defautl+Hysteris=0_25	0,034	0,016	0,010	0,041	0,023	0,016
Defautl+Noise=0	0,048	0,041	0,040	0,053	0,049	0,047
Defautl+Noise=0_25	0,046	0,039	0,038	0,052	0,045	0,043
Defautl+Noise=0_75	0,026	0,009	0,003	0,032	0,011	0,004
Defautl+Noise=1	0,023	0,010	0,004	0,027	0,011	0,004
Defautl+QuadraticTerm=0	0,034	0,016	0,011	0,041	0,022	0,016
Defautl+QuadraticTerm=0_003	0,032	0,021	0,017	0,040	0,031	0,029

(c) Somme normalisée de la valeur absolue des différences.

FIGURE 3.3.4: Estimation des capacités d'exploration de l'espace de notre algorithme en fonction de ses paramètres. Trois mesures de la quantité d'espace couvert sont évaluées à différents moments de la simulation (après respectivement 50, 150 et 300 itérations).

simultanément moyenne et temps de démarrage.

Algorithm 3.1 Calcul du temps de démarrage du système

```
// ENTREES
// Moy_I : le tableau des moyennes de la carte des prédateurs
// NbPoints : nombre de points dans meanI
// SORTIES
// TpsDemarrage : temps de démarrage du système
TpsDemarrage = 0
j = 0
répéter
  TpsDemarrage = j
  // calcul de la moyenne à partir du point de démarrage courant
  MoyDeMoy_I = 0
  pour i de TpsDemarrage à NbPoints faire
    MoyDeMoy_I = MoyDeMoy_I + Moy_I[i]
  fin pour
  MoyDeMoy_I = MoyDeMoy_I/(NbPoints − TpsDemarrage)
  j = 0
  // si la moyenne de départ mean[0] est plus grande que la moyenne globale
  // on avance jusqu'à avoir une moyenne inférieure à la moyenne globale
  si Moy_I[0] > MoyDeMoy_I alors
    répéter
      j = j + 1
    jusqu'à Moy_I[j] < MoyDeMoy_I
  fin si
  // On cherche ensuite la première moyenne supérieure à la moyenne globale
  répéter
    j = j + 1
  jusqu'à Moy_I[j] > MoyDeMoy_I
jusqu'à j == TpsDemarrage
renvoyer TpsDemarrage
```

Statistiques liées aux « fixations »

Bien que notre algorithme ne génère pas de fixations et saccades comparables au modèle humain, on peut estimer le temps moyen entre deux changements de position du focus d'attention. On en calcule alors la moyenne $Temps_{Fixations}$. La détermination d'une nouvelle fixation peut être effectuée de deux façons :

– dès que la position du focus d'attention change, indépendamment de la distance avec la position du prochain focus ;

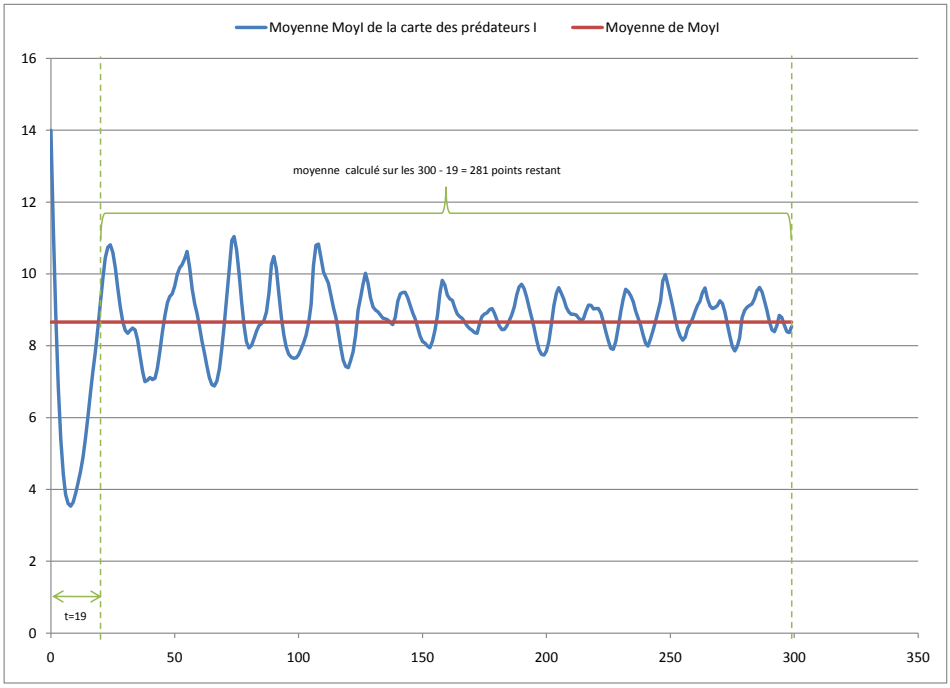


FIGURE 3.3.5: Calcul du temps de démarrage du système. Temps nécessaire pour atteindre la valeur moyenne de la moyenne de la carte des prédateurs I.

– si la distance entre le focus actuel et le suivant est supérieure à un seuil $S_{Fixation}$.

Nous avons opté pour la seconde solution car elle permet de gommer les effets des petits déplacements qui perturberaient l'estimation du nombre et du temps des différentes fixations. La valeur de $S_{Fixation}$ a été choisie afin d'être cohérente avec le paramètre $foveaSize$ utilisé pour générer les *heatmaps* à partir des différentes focalisations. Elle est donc fixée à 15% du plus grand côté de l'image observée.

En utilisant le même seuil $S_{Fixation}$ nous déterminons le nombre moyen de fixations par seconde $Freq_{Fixations}$, ainsi que la distance moyenne entre deux fixations $Dist_{Fixations}$.

3.3.4.2 Résultats et interprétation

« Temps de démarrage »

Le tableau 3.5 résume l'influence de la valeur initiale, du pas de temps utilisé pour la simulation (voir la partie implémentation en Annexe D), du nombre de sous-itérations du système proies / prédateurs à chaque itération du système attentionnel et enfin du facteur de *feedback* positif.

Le choix de la valeur initiale des différentes cartes est important : il doit être défini en fonction du paramétrage utilisé pour assurer la stabilité du système. En effet, plus la valeur initiale est proche de la valeur moyenne obtenue pour les différentes cartes de proies et prédateurs en fonctionnement stable du système, meilleur sera le temps de démarrage.

Le pas et le nombre de sous-itérations servent à trouver un juste équilibre entre :
– les ressources CPU nécessaires à la résolution du système : plus le pas est petit et le nombre d'itérations grand, plus les ressources nécessaires seront importantes ;
– la justesse des calculs effectués : un pas trop important peut engendrer des problèmes numériques lors de la résolution du système d'équations ;
– la dynamique du système : vitesse à laquelle le focus changera et réactivité aux changements entre deux images successives.

La valeur par défaut (pas de 1/3 et 3 itérations) permet d'obtenir un système relativement réactif (temps de démarrage de 25 trames), rapide à calculer, et juste. En réduisant le pas, on augmente la justesse, mais pour garder la même dynamique, il est nécessaire d'augmenter le nombre de sous-itérations. Les calculs deviennent alors trois fois plus lents. En réduisant le pas et le nombre d'itérations, le système est juste et rapide, mais trop peu dynamique (temps de démarrage moyen de 203 itérations sur la base « Bruce »).

Enfin, le facteur de *feedback* positif permet avec sa valeur par défaut (0.001) d'améliorer la dynamique (on passe d'un temps de démarrage de 30 à 25 itérations). Cependant, compte tenu de son influence quadratique, des valeurs plus grandes engendrent des problèmes calculatoires produisant l'effet inverse (passage de 30 à 97 itérations).

	StartTime	
	Bruce	LeMeur
Defautl+InitialValue=1	41,0	43,9
Defautl+InitialValue=4	29,4	32,0
Defautl+InitialValue=8	25,5	28,1
Default	25,5	30,1
Defautl+InitialValue=32	39,0	41,9
Defautl+InitialValue=64	47,9	50,9
Defautl+InitialValue=128	49,4	52,1
Defautl+InitialValue=256	66,4	54,6
Defautl+Step=0_1+Iterations=1	203,4	250,5
Defautl+Step=0_1+Iterations=10	26,2	29,3
Defautl+Step=1+Iterations=1	29,2	53,4
Defautl+QuadraticTerm=0	30,3	33,8
Defautl+QuadraticTerm=0_003	96,8	128,1

TABLE 3.5: Temps de démarrage moyen du système : influence des paramètres.

Statistiques liées aux « fixations »

La première constatation que l'on peut effectuer en consultant les statistiques présentées dans le tableau 3.6 est que la dynamique de notre système est fort différente de celle du système attentionnel humain. En effet, d'après [Le Meur 05a] et [Chauvin 03] le nombre de fixations et la durée des fixations moyens pour un observateur humain sont respectivement d'environ 3 et 300ms. Avec ses réglages par défaut notre système évolue 5 fois plus rapidement. La comparaison avec le modèle humain doit cependant s'arrêter là car notre système attentionnel est différent du système humain sur au moins deux points essentiels :

- comme nous l'avons déjà évoqué, notre système n'est relié à aucun organe moteur : il n'a pas à prendre en compte les limites physiques de ses effecteurs (en terme de vitesse de déplacement par exemple). Si celui-ci devait piloter les focalisations d'une caméra motorisée (vision active), sa dynamique serait certainement impactées par le centrage de l'objet d'attention courant dans le champ visuel et les limitations de déplacement de la caméra (vitesse, angles limites) ;
- lors des expérimentations humaines, le système attentionnel est (forcément) relié au système visuel. Celui-ci influe sur la durée des fixations en fonction du temps qui lui est nécessaire à acquérir l'information. Ce mécanisme n'est pas à l'œuvre dans le système attentionnel seul (non bouclé) étudié dans ce chapitre. Nous montrons au chapitre 4 que le rebouclage permet d'influencer fortement la dynamique du modèle.

D'autres paramètres rentrent en compte dans l'établisement de la dynamique globale de notre système. Premièrement, la fréquence de traitement a été arbitrairement fixée à 30 images par seconde afin de s'approcher de celle des caméras et flux vidéo. De plus, les paramètres régissant l'évolution du système proies / prédateurs (pas d'intégration et nombre d'itérations) ont été ajustés afin d'optimiser la réactivité du système au changement, pas sa fidélité au modèle humain.

D'autres paramètres, non liés à la réactivité du système, peuvent également influencer sa dynamique :

- Le biais central a tendance à concentrer les focalisations au centre de la scène. Les déplacements sont alors en moyenne plus courts et la dynamique du système ralentie.
- La diffusion peut, dans une moindre mesure, également ralentir la dynamique.
- L'hystérésis est comme prévu, un moyen efficace pour modifier la dynamique, sans perturber le fonctionnement du système compétitif (puisque c'est un filtrage appliqué lors de la détermination du maximum de la carte des prédateurs).
- Les valeurs initiales des différentes populations de proies et de prédateurs n'ont (logiquement) pas d'influence sur la dynamique.
- A l'opposé du biais central et de la diffusion, le bruit accélère la dynamique en rendant le système plus chaotique.
- Comme nous l'avons déjà évoqué plusieurs fois, le terme quadratique est à manier avec précaution. Une faible valeur (comme celle utilisée dans le paramétrage par défaut) peut légèrement accélérer le système, alors qu'une valeur plus forte fait rapidement ralentir et saturer le système.
- Etonnamment le filtre rétinien n'a que peu d'influence. Compte tenu du fait qu'il limite le champ visible, on aurait pu s'attendre à ce que les déplacements soient plus courts et moins fréquents, mais cet effet est très réduit.

Notons enfin que, de manière globale, un ralentissement de la dynamique du système tend à réduire également la distance moyenne entre les focalisations.

3.3.5 Bilan

Le tableau 3.7 résume l'influence des différents paramètres sur le fonctionnement du système. Nous ne rappellerons 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.

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.

	Déplacement moyen (pixels)		Nombre de fixations moyen par seconde		Temps de fixation moyen (ms)	
	Bruce	Le Meur	Bruce	Le Meur	Bruce	Le Meur
Default	59,51	57,36	14,80	13,76	67,54	72,66
Default+RetinalFilter	56,54	50,22	14,40	12,86	69,43	77,76
Default+CentralBias=0	74,17	69,28	16,55	15,38	60,44	65,04
Default+CentralBias=0_25	36,67	40,00	10,67	10,66	93,69	93,84
Default+CentralBias=0_5	21,71	24,53	6,67	7,36	149,86	135,89
Default+Diffusion=0	62,27	59,53	16,37	15,01	61,10	66,64
Default+Diffusion=0_5	52,05	52,20	12,31	11,93	81,21	83,83
Default+Diffusion=0_125	65,41	64,35	16,73	15,80	59,79	63,29
Default+Hysteris=0_1	54,33	53,25	13,58	12,84	73,64	77,87
Default+Hysteris=0_25	47,35	47,07	11,85	11,34	84,36	88,19
Default+Hysteris=0_5	39,61	39,91	9,98	9,63	100,19	103,84
Default+InitialValue=1	57,27	54,66	14,27	13,26	70,07	75,44
Default+InitialValue=4	58,34	54,92	14,52	13,31	68,88	75,11
Default+InitialValue=8	58,84	56,29	14,65	13,59	68,26	73,57
Default+InitialValue=32	57,88	56,03	14,38	13,48	69,54	74,18
Default+InitialValue=64	56,15	53,90	14,01	13,03	71,38	76,74
Default+InitialValue=128	56,00	53,32	13,94	12,89	71,75	77,58
Default+InitialValue=256	56,25	53,47	13,97	12,93	71,59	77,34
Default+Noise=0	39,02	41,58	9,14	9,38	109,38	106,64
Default+Noise=0_25	35,15	40,50	8,19	8,84	122,13	113,08
Default+Noise=0_75	82,58	81,67	19,93	19,84	50,18	50,42
Default+Noise=1	75,67	78,80	17,95	18,62	55,71	53,71
Default+QuadraticTerm=0	48,14	50,10	12,13	12,01	82,43	83,29
Default+QuadraticTerm=0_003	32,95	22,06	7,40	4,63	135,09	216,18
Default+Step=0_1+Iterations=1	4,71	5,62	1,19	1,31	843,29	764,22
Default+Step=0_1+Iterations=10	48,59	49,51	11,91	11,56	83,95	86,50
Default+Step=1+Iterations=1	65,80	55,43	16,51	13,47	60,57	74,24

TABLE 3.6: Influence des paramètres du système sur la dynamique des focalisations.

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	$\times$	$\times$	$\times$	$\uparrow$

TABLE 3.7: Résumé de l'influence des différents paramètres.

3.4 Conclusion

Nous avons décrit dans ce chapitre la partie purement *bottom-up* de notre modèle attentionnel. Celle-ci comporte un système visuel hiérarchique permettant de calculer de manière computationellement efficace , et sans déteriorer leur plausibilité, les cartes de singularité nécessaires au système attentionnel.

Ce second système, basé sur la compétition entre différentes colonies de proies et de prédateurs, permet de fusionner ces cartes et de générer un ensemble de focalisations attentionnelles. La comparaison des *heatmaps* obtenues à partir de ces focalisations avec une vérité terrain obtenue à partir d'expériences oculométriques permet de valider la plausibilité de notre modèle.

Comme le montrent les résultats de l'étude de ses différents paramètres, ce modèle possède déjà des capacités intéressantes d'adaptation à un comportement souhaité. En fonction du contexte, on peut ainsi ajuster sa fidélité, sa repoductibilité, sa dynamique ou sa stratégie d'exploration de l'espace.

Cependant, ces possibilités d'adaptation sont limitées : il n'est par exemple pas possible de favoriser certaines cartes de singularité au profit d'autres. De plus, elle ne sont pas dynamiques : le système est incapable de changer son comportement seul.

Pour dépasser ces limites, il est nécessaire d'ajouter de nouveaux mécanismes, basés sur la rétroaction, que nous présentons dans le prochain chapitre.

Points clés

Positionnement

- ❏ La vision adaptative nécessite un modèle attentionnel qui soit : adaptatable, extensible, rapide et dynamique.
- ❏ Pour atteindre ces objectifs, nous proposons un modèle hybride entre les appoches hierarchiques (centralisées) et compétitives (distribuées).

Contributions

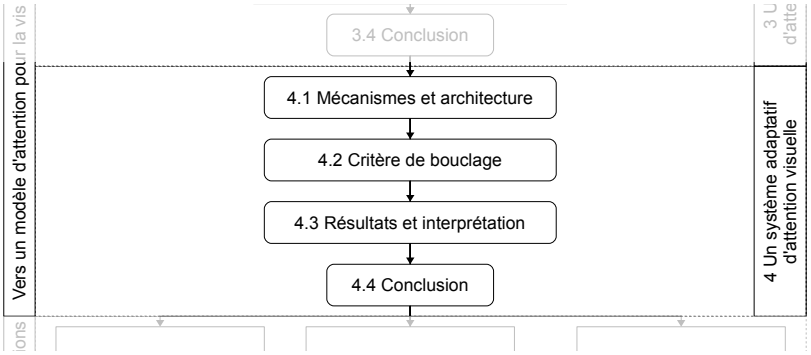
- ❏ Notre système visuel hiérarchique étend les travaux d'Itti et Frintrop afin de calculer en temps réel des cartes de caractéristiques plausibles.
- ❏ L'introduction d'un mécanisme de normalisation des cartes de caractéristiques basé sur les travaux de Mancas et la théorie de l'information permet une fusion rapide et efficace de ces cartes.
- ❏ Notre système d'attention compétitif permet le calcul d'un parcours visuel, sans carte de saillance ni mécanisme d'inhibition de retour. Celui-ci est paramétrable et permet d'adapter le modèle attentionnel à un contexte donné.

Évaluation

- ❏ Le système est robuste sur une large plage de valeurs de ses paramètres de natalité et mortalité. La transition vers ses comportements limites est progressive.
- ❏ Différents paramètres du système (bruit, hystérésis, préférence centrale, flou rétinien, diffusion, *feedback* positif) permettent d'ajuster différentes propriétés du système :
  - sa fidélité au modèle humain
  - sa reproductibilité
  - ses capacités d'exploration de l'espace
  - sa dynamique
- ❏ Le comportement du système est tout à fait plausible, comparativement au modèle humain.

Chapitre 4

Un système adaptatif d'attention visuelle



Dans ce chapitre nous décrivons la composante *top-down* permettant l'adaptation de notre système d'attention. Nous décrivons où et comment ajouter des points d'entrée dans le système proies / prédateurs afin de mieux contrôler son comportement. Nous donnons également un exemple de rebouclage du système permettant son adaptation automatique à une contrainte donnée (ici l'exploration de l'espace).

4.1 Mécanismes et architecture

La figure 4.1.1 présente le système décrit dans les prochaines sections. Il s'agit du système décrit précédemment auquel de nouvelles facultés d'adaptation ont été ajoutées.

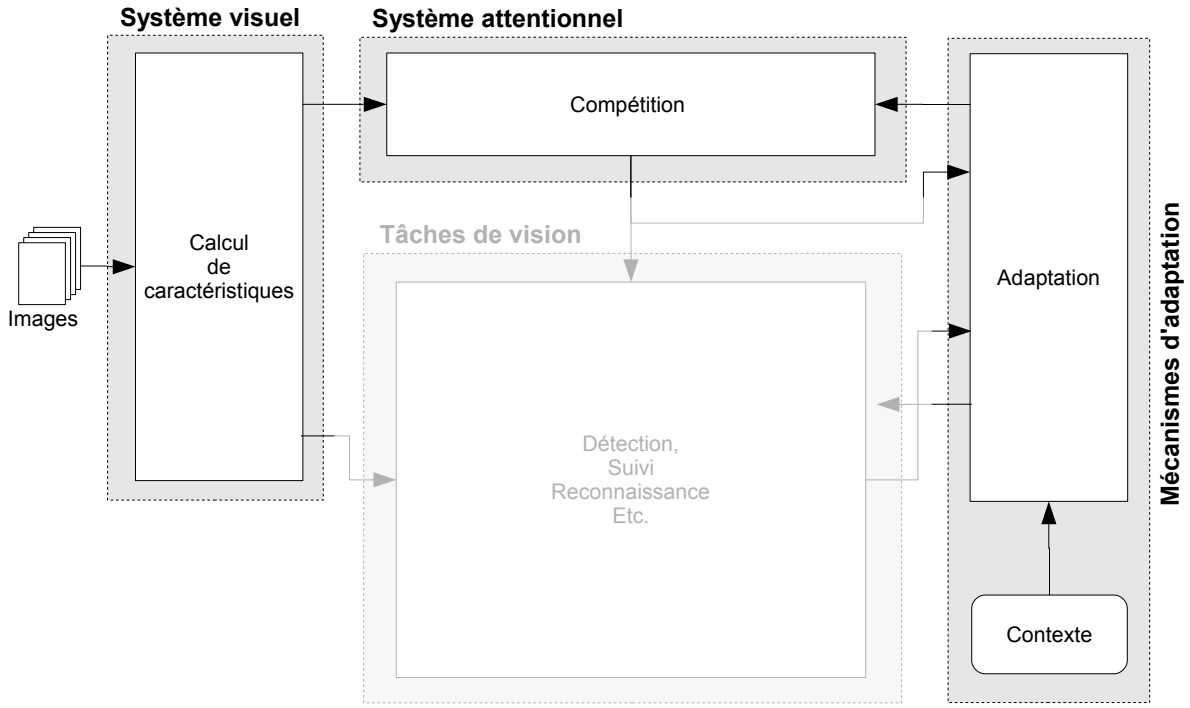


FIGURE 4.1.1: Parties du système présentées dans ce chapitre : nous traitons de la partie adaptative du système d'attention visuelle.

4.1.1 Schéma général

Comparativement au modèle décrit dans le chapitre précédent, la figure 4.1.1 fait apparaître plusieurs nouveautés.

L'adaptation s'effectue en fonction d'un contexte, qui représente les contraintes et objectifs que l'on souhaite appliquer au système d'attention. D'une certaine manière le contexte, bien que non présenté dans le chapitre précédent, était déjà présent lors de l'analyse des propriétés du système. Modifier le paramétrage de celui-ci pour répondre par exemple à un besoin d'exploration de l'espace, est une façon de prendre en compte un contexte particulier. Cependant, celui-ci prend ici une définition plus générale en permettant par exemple d'apporter de l'information *a priori* (par exemple : les objets recherchés sont colorés).

Concernant les mécanismes mis en œuvre, l'adaptation peut être une simple modification des paramètres du système, mais elle peut également utiliser les informations produites par ce même système afin de modifier son comportement. Le système devient alors bouclé (auto-adaptation).

Pour y parvenir, il faut tout d'abord définir ce qui peut être observé, et quels seront les points d'entrée permettant la modification du comportement attentionnel. C'est l'objectif de la sous-section suivante.

4.1.2 Mécanismes d'adaptation

Dans le chapitre 3 nous avons étudié comment l'ajustement des paramètres du système attentionnel permettait de modifier son comportement. Une autre façon d'influer sur le comportement attentionnel est d'injecter de l'information soit extérieure (information *top-down*), soit générée à partir des sorties du système et d'un objectif particulier (rétroaction / auto-adaptation). Cette sous-section présente l'intégration de ces deux mécanismes dans notre modèle d'attention.

4.1.2.1 Cartes *top-down*

Dans le cadre d'une recherche guidée (*guided-search*), 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 05b] 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 05a] *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.

D'autres raffinements sont encore possibles, puisque l'on peut également fournir un *a priori* sur l'intensité des caractéristiques attendues [Navalpakkam 06]. 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 :

$$\frac{d_{C^n_{x,y}}}{dt} = T^n_{x,y}\left(1 - \frac{C^n_{x,y}}{Max_{population}}\right)\left(hC^{*n}_{x,y} + hf\triangle_{C^{*n}_{x,y}}\right) - m_C C^n_{x,y} - sC^n_{x,y}I_{x,y} \qquad (4.1.1)$$

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 4.1.2).

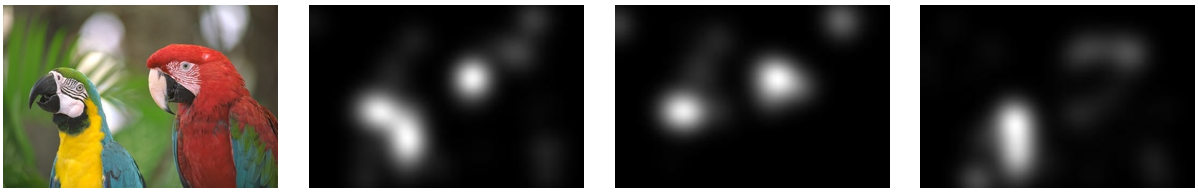


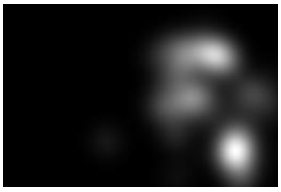
Image originale (a) (b) (c)

FIGURE 4.1.2: 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 ($W^i = 0.5$, $W^c = 1.0$, $W^o = 0.5$).

Sinon, l'accentuation de la saillance est localisée [Torralba 06]. On pourra par exemple attendre un objet coloré à la gauche de la scène (figure 4.1.3).

Image originale

Carte top-down de couleur T^c



$W^i = 0.5$, T^c et $W^o = 0.5$

FIGURE 4.1.3: Exemple de *heatmap* obtenue après modification du comportement attentionnel par pondération locale des caractéristiques. A l'aide de la carte *top-down* T^c, on favorise l'objet coloré situé à gauche de la scène.

4.1.2.2 Cartes de rétroaction

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 [Berthoz 09], comme un des dispositifs-clé permettant la compétition et la sélection, amenant à la simplexité. On aura alors :

$$\frac{d_{C^n_{x,y}}}{dt} = R_{x,y}T^n_{x,y}\left(1 - \frac{C^n_{x,y}}{Max_{population}}\right)\left(hC^{*n}_{x,y} + hf\triangle_{C^{*n}_{x,y}}\right) - m_C C^n_{x,y} - sC^n_{x,y}I_{x,y} \quad (4.1.2)$$

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

4.2 Un critère de bouclage : l'exploration de l'espace

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.

De manière similaire à la section 3.3.3, 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) ;

La suite de ce chapitre présente le calcul de la carte des zones visitées, ainsi que la modulation de celle-ci pour le calcul de la carte de rétroaction.

4.2.1 Une carte des zones visitées

Le calcul de la carte des zones visitées est basée 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 pas exemple de la résolution variable de la rétine). Ce principe est très proche de celui défini en section 3.3.3 pour calculer le masque de flou M_{flou}. On peut donc réutiliser l'équation 3.3.3 afin de construire la carte des zones visitées $M_{visites}$.

$$M_{visites}(x,y,t) = \max(M_{visites}(x,y,t-1), \frac{N_{Levels} - \min\left(\frac{dist(x,y,x_f,y_f)}{BlurSize}, N_{Levels}\right)}{N_{Levels}} \quad (4.2.1)$$

On aura donc $M_{visites}(x,y) \in [0,1]\, \forall x,y$.

Cette équation peut être étendue en prenant en compte le fait que notre système (tout comme l'homme) a une mémoire limitée : il oublie graduellement les informations acquises précédemment. Nous formalisons ce principe en introduisant un facteur d'oubli $F_{forget} \in [0,1]$qui servira à atténuer le rôle des focalisations plus anciennes, dans le calcul de $M_{visites}$:

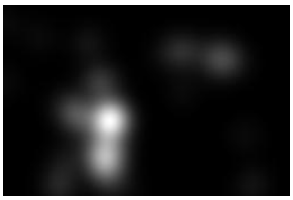
$$M_{visites}(x,y,t) = \max(F_{forget} \times M_{visites}(x,y,t-1), \frac{N_{Levels} - \min\left(\frac{dist(x,y,x_f,y_f)}{BlurSize}, N_{Levels}\right)}{N_{Levels}} \quad (4.2.2)$$

La figure 4.2.1 montre l'influence du facteur d'oubli sur la carte des zones visitées $M_{visites}$.

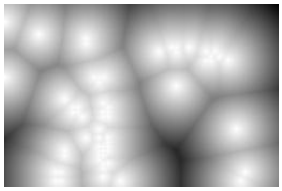
4.2.2 Calcul de la carte de rétroaction

A partir de la carte des zones visitées $M_{visites}$, nous pouvons construire la carte de rétroaction R. Nous devons alors pouvoir moduler la rétroaction en fonction de l'inten-

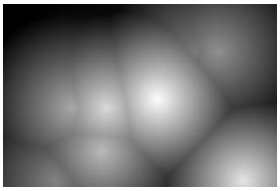
(a) Image originale



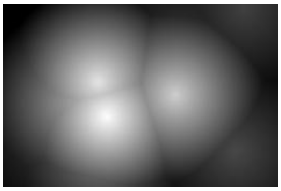
(b) *Heatmap.*



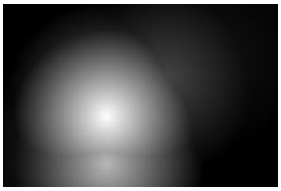
(c) Zones visitées, $F_{forget} = 1.0$.



(d) Zones visitées, $F_{forget} = 0.95$



(e) Zones visitées, $F_{forget} = 0.90$.



(f) Zones visitées, $F_{forget} = 0.85$.

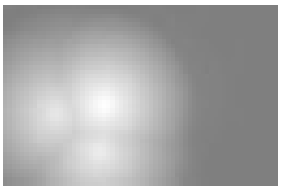
FIGURE 4.2.1: Influence du facteur d'oubli sur la carte des zones visitées après 100 itérations.

sité voulue, et du type d'influence (positive ou négative). Ceci est contrôlé par un seul paramètre $F_{feedback}$:

$$R(x,y) = \begin{cases} \frac{1+|F_{feedback}|\times M_{visites}(x,y)}{1+|F_{feedback}|} & si\ F_{feedback} \geq 0 \\ \frac{1+|F_{feedback}|\times(1-M_{visites}(x,y))}{1+|F_{feedback}|} & sinon \end{cases} \tag{4.2.3}$$

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

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). Les cartes de rétroaction R correspondant à ces deux cas de figure sont représentées figure 4.2.2.



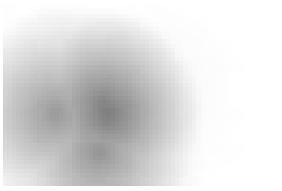


FIGURE 4.2.2: Cartes de rétroaction : à gauche $F_{feedback} = 1.0$, à droite $F_{feedback} = -1.0$.

4.3 Résultats et interprétation

Dans cette section, nous présentons différents éléments permettant d'apprécier l'influence du bouclage (*feedback*) sur le comportement de notre modèle attentionnel. L'influence des cartes *top-down* n'est pas étudiée, car par essence ces cartes dépendent d'un objectif fixé de manière extérieure au système attentionnel. En l'absence de connexion avec l'extérieur (un système de vision, par exemple), leur étude sort du cadre de ce chapitre.

Pour étudier l'influence du *feedback*, nous effectuons une première analyse qualitative, en observant les simulations effectuées sur des exemples d'images fixes et de vidéos. Puis, nous décrivons les résultats de l'analyse quantitative, utilisant les mêmes mesures qu'au chapitre précédent, afin de caractériser l'influence du bouclage sur la fidélité au modèle humain, la reproductibilité et l'exploration de l'espace.

Tous les résultats présentés dans cette section ont été réalisés en utilisant un facteur d'oubli $F_{forget} = 0.95$.

4.3.1 Analyse qualitative

Il est assez délicat de choisir une représentation adaptée à la description de la dynamique d'exploration du modèle en fonction de la valeur du paramètre de *feedback*. Dans cette sous-section, nous avons choisi d'afficher l'état de la *heatmap* à différents stades de l'évolution du système. On peut ainsi juger de l'évolution de la répartition des différentes focalisations au cours du temps.

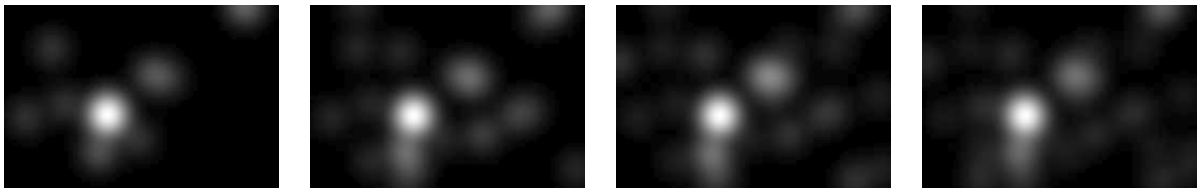
4.3.1.1 Images fixes

La figure 4.3.1 permet d'observer les *heatmaps* générées par le système pour différentes valeurs de $F_{feedback}$:

- en l'absence de *feedback*, le système visite en majorité le perroquet jaune, ceci dès le début de la simulation (t=25) ;
- avec un *feedback* négatif important ($F_{feedback} = -1$), le système visite très tôt des zones moins saillantes de l'image. L'exploration est également plus diversifiée tout au long de la simulation ;
- avec un *feedback* positif important ($F_{feedback} = 1$), le système visite uniquement le perroquet jaune et n'explore plus aucune autre zone.

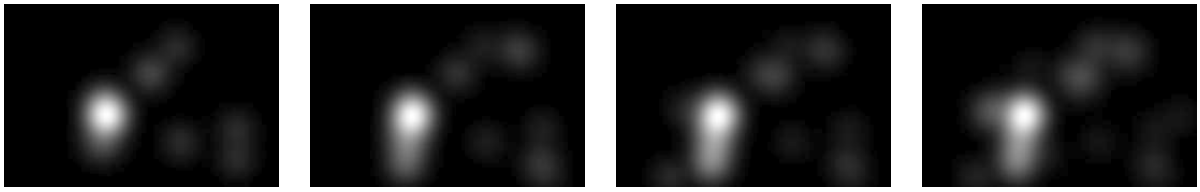
Ces observations confortent le comportement attendu. Un *feedback* négatif permet une exploration de la scène plus exhaustive, alors qu'un *feedback* positif renforce l'observation de l'objet le plus saillant de la scène.

(a) Image d'origine « Parrots ».



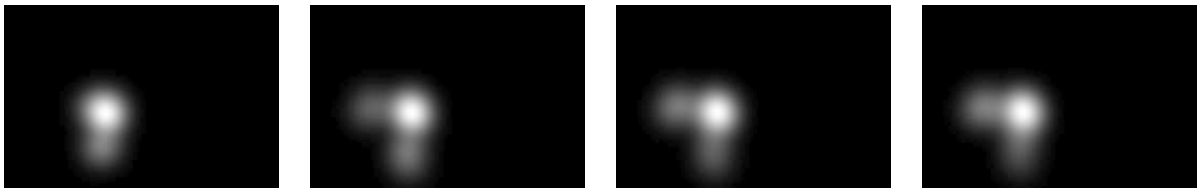
t=25 t=50 t=75 t=100

(b) Exploration : $F_{feedback} = -1$.



t=25 t=50 t=75 t=100

(c) Normal : $F_{feedback} = 0.0$.



t=25 t=50 t=75 t=100

(d) Suivi : $F_{feedback} = 1$.

FIGURE 4.3.1: Évolution des *heatmaps* générées par le système pour différentes valeurs du facteur de *feedback*. Afin de mieux faire ressortir les différences, le gamma des images a été ajusté à 1.5.

4.3.1.2 Vidéo

Le cas de la vidéo est un peu différent de celui des images fixes, l'interprétation des *heatmaps* devant alors prendre en compte la trajectoire des objets présents sur la scène. Dans le cas de la séquence d'exemple « Boule+Grille », la boule effectue une trajectoire en « v » partant du bord supérieur droit de la scène et arrivant au bord supérieur gauche.

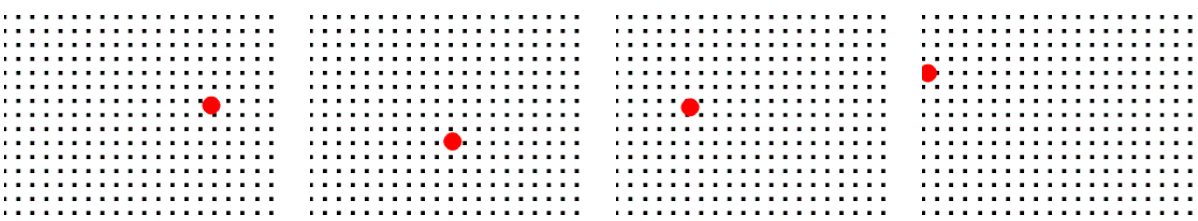
En observant les trajectoires obtenues sur la *heatmap*, pour différentes valeurs de $F_{feedback}$ on peut remarquer que :

- en l'absence de *feedback*, le système suit globalement la trajectoire de la boule. Il lui arrive cependant régulièrement de « décrocher » pendant quelques pas de simulation afin d'observer quelques points de la grille en mouvement. Les décrochages sont relativement nombreux mais très brefs ;
- avec un *feedback* négatif important ($F_{feedback} = -1$), le phénomène de décrochage ne s'amplifie pas, par contre ils sont plus longs. Globalement le système passe moins de temps à observer la boule rouge (on peut le constater en comparant l'intensité moyenne de la ligne traçant la trajectoire sur les *heatmaps* avec celle de la simulation sans *feedback*) ;
- avec un *feedback* positif modéré ($F_{feedback} = 0.5$), le système ne décroche plus, il suit la boule correctement. Le point hors trajectoire que l'on observe dès $t = 25$ correspond à la focalisation initiale du système, qui n'a pas suivi tout de suite la boule rouge ;
- avec un *feedback* positif fort ($F_{feedback} = 1$), le système a du mal à suivre la boule rouge. En effet, le *feedback* positif contraint trop fortement les focalisations vers la position occupée au pas de simulation précédent. Le système a alors trop d'inertie et fini par décrocher, restant à sa dernière position.

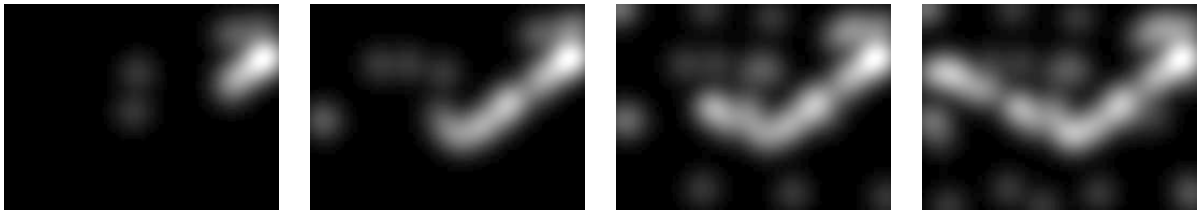
L'effet du *feedback* lors du traitement de vidéo par notre système est plus important que sur des images fixes. Il modifie la dynamiques des focalisations, influant ainsi sur la prise en compte des mouvements dans la scène. Il faut alors utiliser des valeurs plus faibles afin de ne pas avoir un effet inverse à celui voulu (perte du suivi pour un fort *feedback* positif).

4.3.2 Analyse quantitative

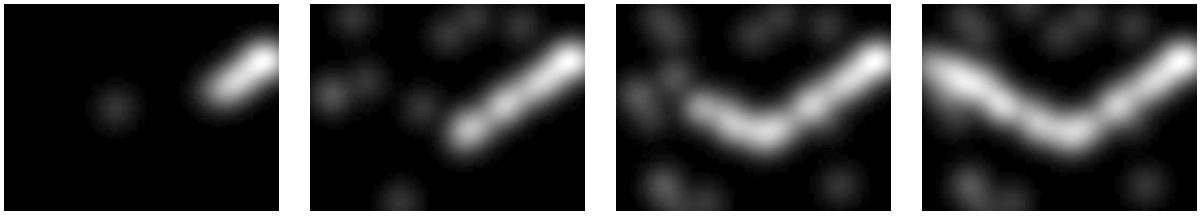
La quantification de l'influence du *feedback*, a été effectuée avec les mêmes mesures qu'au chapitre 3. La stabilité du système n'a cependant pas été étudiée puisque les paramètres concernés sont restés inchangés. L'influence de $F_{feedback}$ a été étudiée pour des valeurs comprises entre -1 (exploration) et 1 (focalisation), ceci en utilisant comme base le système attentionnel avec ses paramètres par défaut.



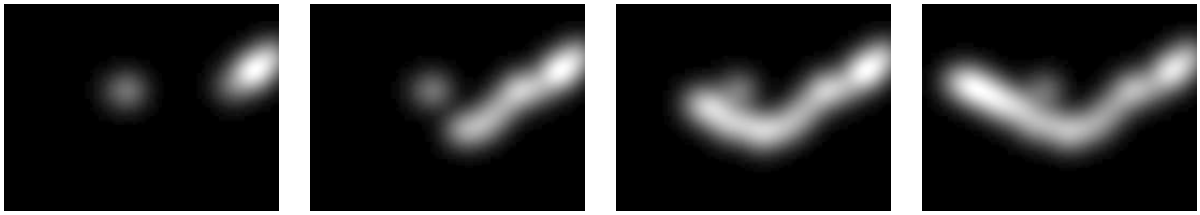
(a) Séquence d'origine « Boule+grille ». La vidéo complète est disponible en ligne : `http://www.youtube.com/watch?v=6sHcxPPs4UA`.



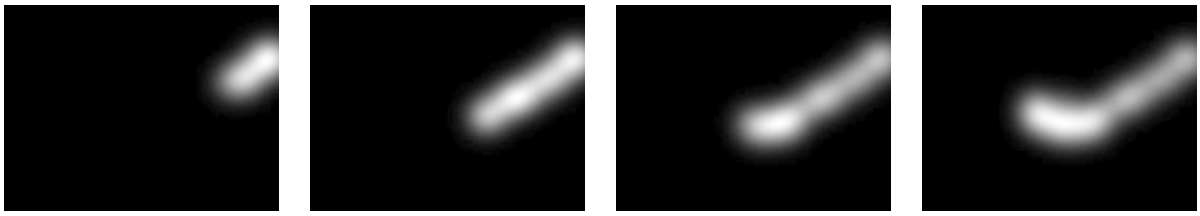
(b) Exploration : $F_{feedback} = -1$.



(c) Normal : $F_{feedback} = 0.0$.



(d) Suivi : $F_{feedback} = 0.5$.



(e) Suivi : $F_{feedback} = 1$.

FIGURE 4.3.2: Évolution des *heatmaps* générées par le système pour différentes valeurs du facteur de *feedback*. Séquence « Boule+Grille ». Les images affichées représentent l'état du tystème aux temps t=25, 50, 75 100.

	CrossCorrelation		KullbackLeiblerDivergence		NormalizedScanpathSaliency		Gain moyen	Gain moyen
	Bruce	LeMeur	Bruce	LeMeur	Bruce	LeMeur	CC +NSS	KLD
Default-feedback=-1,0	0,28	0,24	1,94	1,65	0,80	0,45	-19%	-1%
Default-feedback=-0,8	0,28	0,24	1,93	1,64	0,81	0,46	-18%	0%
Default-feedback=-0,6	0,30	0,24	1,91	1,64	0,84	0,46	-16%	1%
Default-feedback=-0,4	0,31	0,25	1,87	1,65	0,88	0,47	-13%	1%
Default-feedback=-0,2	0,34	0,26	1,81	1,67	0,94	0,50	-7%	3%
Default	0,35	0,30	1,80	1,76	0,95	0,56	0%	0%
Default-feedback=0,2	0,38	0,33	1,88	2,05	1,03	0,63	9%	-10%
Default-feedback=0,4	0,41	0,30	2,11	2,40	1,12	0,59	10%	-26%
Default-feedback=0,6	0,41	0,31	2,38	2,62	1,15	0,61	12%	-40%
Default-feedback=0,8	0,42	0,31	2,66	2,87	1,17	0,62	14%	-55%
Default-feedback=1,0	0,44	0,33	2,86	2,97	1,22	0,66	20%	-64%

TABLE 4.1: Influence du *feedback* sur la plausibilité du modèle, comparé à une vérité terrain humaine.

4.3.2.1 Fidélité au modèle humain

L'influence du *feedback* sur la fidélité du modèle est plus difficile à décrypter que précédemment. En effet, comme on peut le constater dans le tableau 4.1, les variations moyennes observées semblent contradictoires :

- pour la corrélation croisée et la *normalized scanpath salience*, l'utilisation d'un *feedback* positif semble améliorer la plausibilité de notre modèle ;
- pour la divergence de Kullback Leibler, elle semble au contraire la faire baisser.

Ces résultats sont cependant explicables (figure 4.3.3) .

La corrélation et la NSS sont sensibles aux zones de forte corrélation entre le modèle et la vérité terrain. Les zones où le modèle ou la vérité terrain ont de faibles valeurs influent peu sur la valeur finale de correspondance entre les deux cartes. La divergence de Kullback-Leibler mesure par contre la dissimilarité entre les deux distributions, c'est donc les zones de divergence entre les deux cartes qui influent le plus sur sa valeur.

L'utilisation d'un *feedback* négatif important, impose au système attentionnel de mieux explorer l'espace. Il insistera alors moins sur les zones les plus saillantes et visitera plus de zones non saillantes. La corrélation et le NSS vont légèrement baisser car notre modèle aura des valeurs moins importantes dans les zones les plus saillantes. Cette baisse n'est pas compensée par l'exploration plus exhaustive de l'espace. La divergence de Kullback-Leibler restera par contre stable, car ce qui est perdu en similarité dans les zones les plus saillantes sera en moyenne compensé par la très légère baisse de dissimilarité dans les zones moins saillantes.

L'utilisation d'un *feedback* positif important cantonne l'évolution des focalisations autour des zones les plus saillantes. La corrélation et la NSS augmentent légèrement car on obtient alors une meilleure correspondance dans les zones de forte saillance. Cependant, ce type d'exploration « étale » les zones de forte saillance définies par notre modèle. Ainsi la différence entre la vérité terrain et notre modèle s'en trouve paradoxalement augmentée.

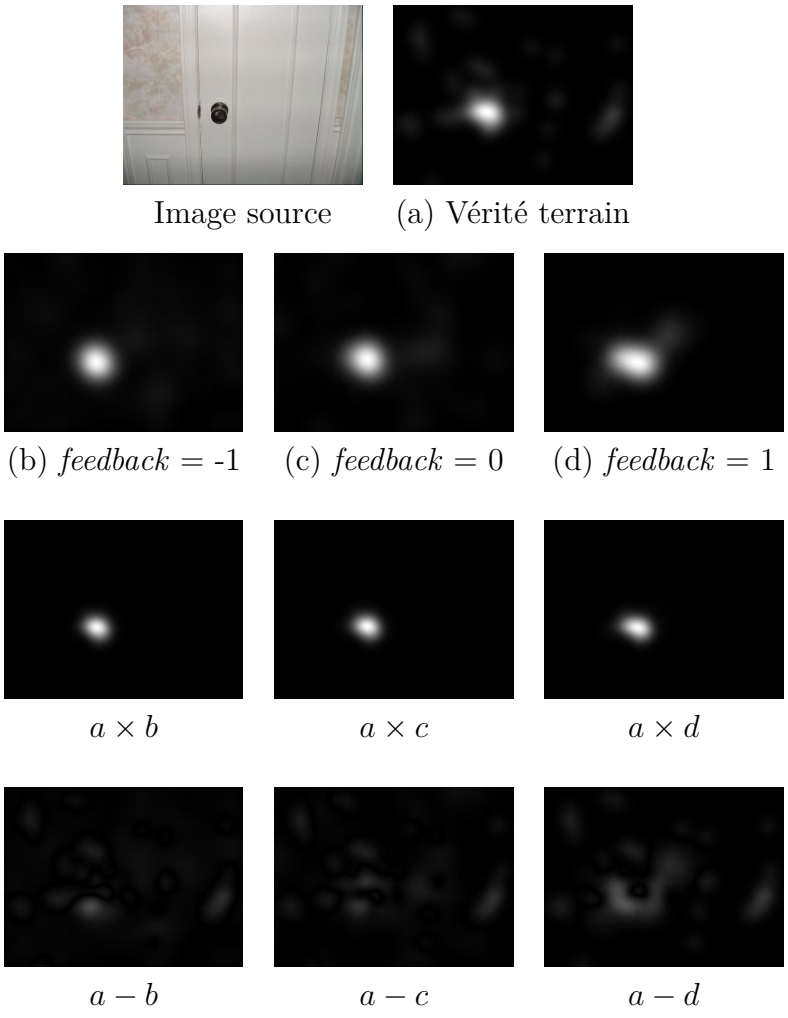


FIGURE 4.3.3: Illustration des différences entre corrélation / NSS et divergence de Kullback-Leibler. Sur cet exemple, lorsque l'on utilise un *feedback* positif important, la zone de correspondance entre le modèle et la vérité terrain augmente légèrement (ligne du milieu, à droite). Mais la différence entre les deux *heatmaps* ainsi générées augmente également (ligne du bas, à droite). On a ainsi une augmentation conjointe de la correlation, de la NSS et de la divergence de Kullback-Leibler.

Il est donc assez difficile de juger de l'influence du *feedback*, car celle-ci est double. On peut cependant conclure que le *feedback* permet d'améliorer légèrement la corrélation du modèle avec la vérité terrain dans les zones les plus saillantes, au prix d'une augmentation de la différence avec la vérité terrain dans les zones moins saillantes.

4.3.2.2 Reproductibilité

L'étude de la reproductibilité du système (tableau 4.2) fait apparaître les mêmes résultats contradictoires que lors de l'étude de sa fidélité.

Un *feedback* négatif ayant tendance à éparpiller les focalisations sur la *heatmap*, corrélation et NSS ont alors tendance à baisser : la variabilité dans les zones de forte saillance est plus grande. La divergence de Kullback-Leibler baisse également, car la plus grande exploration de l'espace à tendance à homogénéiser les zones peu saillantes visitées, et donc légèrement diminuer la différence moyenne entre les simulations.

L'effet contraire est observé lorsqu'un *feedback* positif est utilisé. La corrélation croît légèrement car les zones de forte saillance sont plus stables. Toutes les focalisations étant concentrées dans les zones de forte saillance, les dissimilarités (même si elles sont peu nombreuses) sont alors également plus importantes, faisant croître la NSS.

	Cross Correlation		Kullback LeiblerDivergence		Normalized Scanpath Saliency		Gain moyen	Gain moyen
	Bruce	LeMeur	Bruce	LeMeur	Bruce	LeMeur	CC +NSS	KLD
Default-feedback=-1,0	0,72	0,82	0,19	0,23	1,23	1,78	-20%	18%
Default-feedback=-0,8	0,74	0,83	0,19	0,22	1,27	1,80	-19%	19%
Default-feedback=-0,6	0,76	0,84	0,20	0,22	1,33	1,87	-16%	18%
Default-feedback=-0,4	0,79	0,86	0,20	0,22	1,43	1,96	-12%	17%
Default-feedback=-0,2	0,83	0,87	0,22	0,23	1,59	2,10	-7%	12%
Default	0,85	0,89	0,26	0,25	1,82	2,35	0%	0%
Default-feedback=0,2	0,83	0,87	0,40	0,35	2,01	2,55	3%	-46%
Default-feedback=0,4	0,83	0,87	0,53	0,44	2,37	2,83	11%	-89%
Default-feedback=0,6	0,84	0,87	0,57	0,49	2,70	2,93	17%	-108%
Default-feedback=0,8	0,83	0,85	0,71	0,61	2,88	3,04	20%	-160%
Default-feedback=1,0	0,82	0,85	0,86	0,69	2,94	3,01	20%	-204%

TABLE 4.2: Influence du *feedback* sur la reproductibilité du système.

4.3.2.3 Exploration de l'espace

Le mécanisme de bouclage présenté dans ce chapitre a été conçu pour influer sur la manière dont le système attentionnel explore l'espace. Les résultats obtenus avec les différentes mesures de quantité d'information que nous avons définies (ratio de compression JPEG et PNG, moyenne des valeurs absolues des différences) confirment l'influence attendue (tableau 4.3) :

– un fort *feedback* négatif force une exploration plus rapide mais pas forcément plus exhaustive de la scène (le paramétrage par défaut permet déjà de couvrir la quasi totalité de la scène à $t = 300$) ;

– l'utilisation d'un fort *feedback* positif permet de limiter fortement la zone couverte : même à $t = 300$, le taux de couverture de la scène reste inférieur à celui obtenu avec le paramétrage par défaut à $t = 50$. L'exploration de la scène est donc bien limitée et n'augmentera que peu pour $t > 300$.

4.3.2.4 Dynamique

L'influence du *feedback* sur le temps de démarrage du système attentionnel est résumée dans le tableau 4.4. Lorsque l'on utilise un fort *feedback* négatif (exploration), le système proies / prédateurs atteint plus rapidement son régime oscillant « stationnaire ». A l'opposé, le *feedback* positif semble ralentir le temps de démarrage : le fait de restreindre l'évolution du système proies / prédateurs à une zone limitée de la scène perturbe considérablement son évolution. Le système a alors du mal à atteindre un régime oscillant autour d'une valeur stable. Le système n'évolue pas plus lentement, mais il peine à se stabiliser (figure 4.3.4).

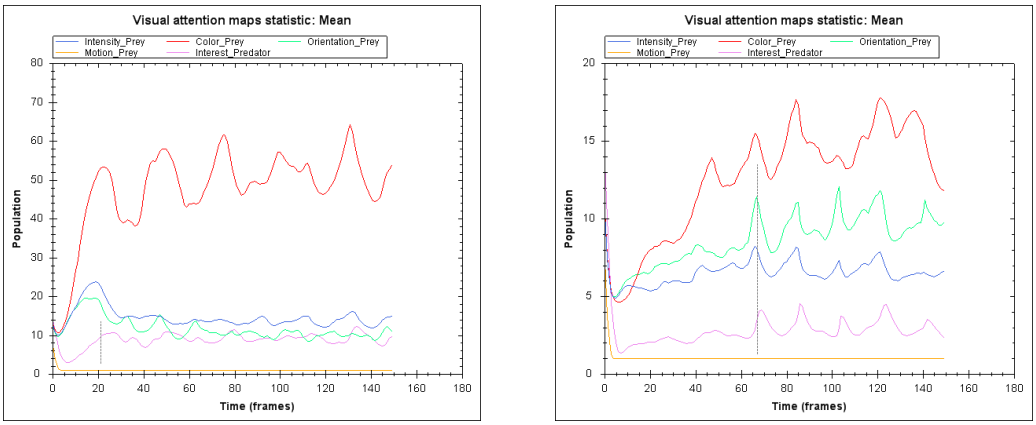


FIGURE 4.3.4: Détails du temps de démarrage du système sur l'image « Parrots ». A gauche, avec les paramètres par défaut un régime oscillant « stationnaire » est rapidement atteint (ligne pointillée grise). A droite, avec un *feedback* positif, le système a du mal à osciller autour d'une valeur stable. Le temps de démarrage mesuré est alors bien plus important.

L'influence du *feedback* sur la dynamique est beaucoup plus directe (tableau 4.5). Pour des valeurs négatives (exploration), les déplacements sont importants et plus fréquents. Pour des valeurs positives (focalisation) les déplacements se réduisent et sont moins fréquents : on obtient alors une dynamique se rapprochant de celle du modèle humain.

	Bruce			Le Meur		
	t=50	t=150	t=300	t=50	t=150	t=300
Default-feedback=-1,0	0,800	0,932	0,971	0,758	0,897	0,940
Default-feedback=-0,8	0,796	0,931	0,970	0,760	0,898	0,941
Default-feedback=-0,6	0,788	0,930	0,970	0,756	0,896	0,940
Default-feedback=-0,4	0,765	0,924	0,965	0,739	0,888	0,935
Default-feedback=-0,2	0,746	0,912	0,957	0,725	0,875	0,921
Default	0,732	0,881	0,931	0,696	0,840	0,891
Default-feedback=0,2	0,622	0,771	0,841	0,597	0,731	0,794
Default-feedback=0,4	0,564	0,702	0,783	0,537	0,665	0,751
Default-feedback=0,6	0,528	0,656	0,729	0,515	0,641	0,713
Default-feedback=0,8	0,514	0,624	0,680	0,485	0,599	0,655
Default-feedback=1,0	0,491	0,599	0,656	0,470	0,590	0,633

Ratio de compression JPEG.

	Bruce			Le Meur		
	t=50	t=150	t=300	t=50	t=150	t=300
Default-feedback=-1,0	0,948	0,990	0,997	0,932	0,976	0,987
Default-feedback=-0,8	0,947	0,989	0,997	0,934	0,977	0,987
Default-feedback=-0,6	0,933	0,981	0,989	0,926	0,974	0,986
Default-feedback=-0,4	0,927	0,988	0,996	0,918	0,973	0,985
Default-feedback=-0,2	0,915	0,984	0,994	0,914	0,969	0,982
Default	0,906	0,972	0,987	0,893	0,954	0,969
Default-feedback=0,2	0,800	0,903	0,940	0,819	0,897	0,929
Default-feedback=0,4	0,757	0,859	0,904	0,769	0,854	0,901
Default-feedback=0,6	0,729	0,824	0,867	0,749	0,836	0,876
Default-feedback=0,8	0,719	0,804	0,839	0,725	0,809	0,841
Default-feedback=1,0	0,700	0,790	0,827	0,706	0,795	0,826

Ratio de compression PNG.

	Bruce			Le Meur		
	t=50	t=150	t=300	t=50	t=150	t=300
Default-feedback=-1,0	0,021	0,009	0,005	0,028	0,014	0,009
Default-feedback=-0,8	0,021	0,009	0,005	0,028	0,014	0,008
Default-feedback=-0,6	0,027	0,016	0,012	0,028	0,013	0,008
Default-feedback=-0,4	0,024	0,010	0,005	0,030	0,015	0,009
Default-feedback=-0,2	0,026	0,011	0,006	0,032	0,016	0,011
Default	0,030	0,013	0,008	0,037	0,019	0,014
Default-feedback=0,2	0,050	0,026	0,017	0,055	0,033	0,025
Default-feedback=0,4	0,064	0,037	0,025	0,065	0,043	0,030
Default-feedback=0,6	0,072	0,046	0,033	0,072	0,048	0,036
Default-feedback=0,8	0,076	0,052	0,042	0,080	0,057	0,046
Default-feedback=1,0	0,081	0,058	0,046	0,084	0,059	0,051

Moyenne de la valeur absolue des différences.

TABLE 4.3: Influence du *feedback* sur l'exploration de l'espace.

	StartTime	
	Bruce	LeMeur
Default-feedback=-1,0	14,4	14,1
Default-feedback=-0,8	15,0	15,0
Default-feedback=-0,6	16,0	16,0
Default-feedback=-0,4	17,2	18,0
Default-feedback=-0,2	19,4	21,0
Default	25,5	30,1
Default-feedback=0,2	56,5	76,2
Default-feedback=0,4	74,0	77,2
Default-feedback=0,6	65,3	59,0
Default-feedback=0,8	55,8	42,7
Default-feedback=1,0	51,3	37,8

TABLE 4.4: Influence du *feedback* sur le « temps de démarrage » du système proies / prédateurs.

	Déplacement moyen (pixels)		Nombre de fixations		Temps de fixation moyen (ms)	
	Bruce	Le Meur	Bruce	Le Meur	Bruce	Le Meur
Default-feedback=-1,0	85,51	79,33	19,56	18,24	51,12	54,83
Default-feedback=-0,8	84,15	78,39	19,34	18,07	51,70	55,35
Default-feedback=-0,6	82,03	76,53	18,96	17,71	52,75	56,47
Default-feedback=-0,4	78,57	73,97	18,34	17,18	54,53	58,21
Default-feedback=-0,2	72,13	68,57	17,18	16,12	58,22	62,02
Default	59,51	57,36	14,80	13,76	67,54	72,66
Default-feedback=0,2	36,33	39,84	10,00	9,91	99,99	100,88
Default-feedback=0,4	22,52	27,96	6,19	6,83	161,46	146,33
Default-feedback=0,6	15,28	20,38	4,02	4,94	248,60	202,43
Default-feedback=0,8	11,42	15,12	2,81	3,73	356,39	267,92
Default-feedback=1,0	9,29	11,96	2,05	2,89	487,11	346,02

TABLE 4.5: Influence du *feedback* sur la dynamique des focalisations.

4.4 Conclusion

Les mécanismes proposés dans ce chapitre permettent de modifier, en cours de simulation, le comportement du système. Cela peut se faire en fonction des connaissances *a priori* sur la cible à rechercher (cartes *top-down*) ou en ajustant la stratégie de parcours de la scène (bouclage). Ce dernier mécanisme est le plus intéressant car il permet de jouer sur des propriétés (ajustement entre focalisation et exploration) qui, à notre connaissance, ne sont pas prises en compte par les modèles existants.

Les expérimentation menées montrent que le bouclage (*feedback*) n'a pas d'influence majeure sur la fidélité ou la reproductibilité du modèle. Par contre, il devient possible de moduler finement le degré de focalisation ou au contraire d'exploration du modèle. De même, la dynamique du système peut être adaptée de manière importante, permettant a celui-ci d'approcher le rythme de fixation moyen du modèle humain.

Les possibilités offertes par ces deux mécanismes sont nombreuses, mais pourraient encore être améliorées. En effet, le système compétitif travaille pour l'instant à partir des cartes de singularité ; il n'est pas possible de favoriser une caractéristique précise (couleur rouge, ou orientation de 45° par exemple). Pour cela, il faudrait soit :

- modifier notre architecture pour qu'elle travaille directement sur les cartes de caractéristiques ;
- multiplier les cartes de caractéristiques par des cartes de poids lors de leur normalisation (dans la partie hiérarchique du système attentionnel).

Compte tenu de la complexité architecturale engendrée par la première solution, la pondération des cartes de caractéristiques semble être la perspective la plus envisageable pour étendre les possibilités du modèle.

Une autre extension possible concerne le mécanisme de *feedback* présenté dans ce chapitre. Celui-ci est basé sur une mesure très simple de l'exploration de l'image, permettant au système de fonctionner de façon autonome. Des mesures plus complexes pourraient être utilisées, en interagissant notamment avec le système de vision hôte, afin de connaître ses besoins. Un tel type de mesure nécessiterait une mémoire de travail afin de stocker des informations de haut niveau sur les zones déjà visitées, ainsi qu'un estimateur de la pertinence des zones visitées en fonction par exemple de leur correspondance avec une base de modèles d'objets connus ou recherchés.

Comme on peut le constater, les perspectives d'extension du modèle sont nombreuses. Cependant, celui-ci est déjà suffisamment complet pour permettre son utilisation dans différentes applications, que nous présentons dans la prochaine partie (chapitres 5, 6 et 7).

Points clés

Positionnement

- ❏ Afin d'optimiser leur exploration de la scène, les systèmes de vision doivent pouvoir modifier le comportement du système attentionnel auquel ils sont connectés. Cette adaptation est généralement effectué en « injectant », *via* la voie *top-down* des modèles d'attention, de la connaissance *a priori* sur les éléments à détecter dans la scène.
- ❏ Il peut également être intéressant de modifier le comportement d'un modèle attentionnel de manière plus globale, en ajustant par exemple la stratégie d'exploration de la scène (en privilégiant la focalisation ou, au contraire, l'exploration).

Contributions

- ❏ Il est possible d'adapter le comportement du système à la recherche d'éléments aux caractéristiques particulières à l'aide de cartes *top-down*. Il est alors de la responsabilité du système de vision, connecté au système attentionnel, de mettre à jour ces cartes en fonction du contexte.
- ❏ Un autre mécanisme d'adaptation, celui-ci automatique, permet *via* un rebouclage des informations produites par le système attentionnel, de modifier le comportement de celui-ci afin d'explorer plus largement la scène ou au contraire se focaliser sur un objet saillant et le suivre.

Évaluation

- ❏ Le *feedback* a une influence moyenne sur la fidélité et la reproductibilité de notre modèle d'attention.
- ❏ Il permet par contre un ajustement précis de la stratégie d'exploration de l'image, tant d'un point de vue statique (taux d'exploration) que dynamique (manière d'explorer).
- ❏ Si on souhaite se rapprocher du modèle humain, l'utilisation d'un *feedback* positif est à envisager car elle permet d'améliorer la plausibilité du modèle tant d'un point de vue statique (comparaison des *heatmaps*) que dynamique (fréquence des fixations).

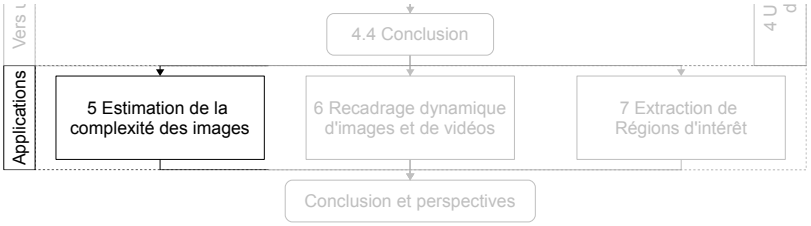
Applications

« Un expert c'est quelqu'un qui a fait toutes ses erreurs dans un champ réduit d'applications. »

(Niels Bohr)

Chapitre 5

Estimation de la complexité des images

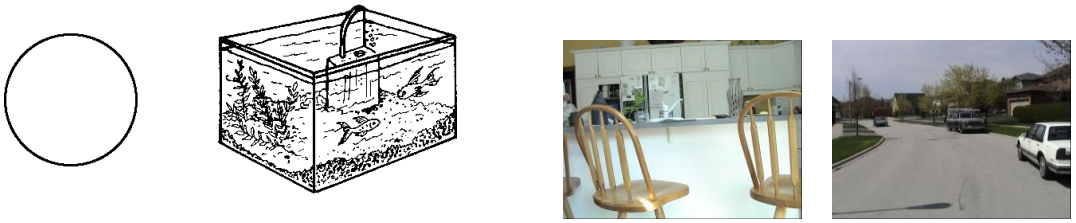


Dans ce chapitre nous proposons une nouvelle méthode de caractérisation de la complexité des images. Celle-ci est basée, non pas sur l'analyse des pixels de l'image, mais sur l'étude du parcours formé par les différentes focalisations de notre système attentionnel lorsque des images lui sont présentées. Cette méthode permet une mesure plus orientée sur la perception visuelle des images, et non sur leur contenu intrinsèque.

5.1 Introduction

L'étude de la complexité visuelle est un phénomène relativement récent. Les précurseurs dans ce domaine sont les psychologues Snodgrass et Vanderwart [Snodgrass 80] qui, au début des années 1980 ont établi un classement de la complexité d'un jeu d'images. Les recherches effectuées à cette époque visaient avant tout à étudier les processus cognitifs liés à l'analyse humaine des *stimuli* visuels. La complexité était alors déterminée expérimentalement par un certain nombre de sujets devant noter, sur une échelle donnée, la complexité des images qu'on leur présentait. Les images étudiées n'étaient pas des photographies, mais des dessins en noir et blanc représentant des objets familiers (figure 5.1.1a et 5.1.1b). Afin d'obtenir une mesure plus objective de la complexité, des mesures plus « mathématiques » ont ensuite été introduites (nombre de segments

de droites, de croisements de lignes, etc.). Celles-ci n'étaient cependant pas calculées informatiquement.



(a) (b) (c) (d)

FIGURE 5.1.1: La complexité des images n'est pas toujours facile à définir subjectivement. L'image (b) est clairement plus complexe que l'image (a). Mais (c) est-elle plus complexe que (d) ?

Plus récemment, avec le développement des techniques de traitement d'images, la communauté informatique s'est également intéressée à la notion de complexité visuelle. Les applications potentielles sont en effet variées. Tout d'abord, connaitre la complexité d'une image permet de mesurer qualitativement les performances relatives d'un algorithme de vision ou de traitement d'images en fonction de la complexité des images qui lui sont présentées [Peters 90]. Celle-ci peut aussi être une caractéristique intéressante pour un système de recherche d'images par le contenu [Perkiö 09] : on pourra alors rechercher des images possédant les mêmes attributs de formes, couleur, etc. et de même complexité que l'image requête. La complexité peut également servir à estimer la capacité de watermarking d'une image [Yaghmaee 10] : plus celle-ci sera complexe, plus il sera facile d'y insérer de l'information sans en dégrader la qualité. La liste des applications possibles est bien évidemment plus longue, mais ces quelques exemples permettent de souligner la diversité des possibilités.

L'estimation de la complexité d'une image est donc intéressante, tant d'un point de vue de l'étude psychologique des phénomènes de perception, d'analyse et de mémorisation de l'information visuelle, que d'une exploitation plus informatique. Cependant, son calcul se heurte à un problème majeur : la définition même de la complexité. En effet, celle-ci est relativement floue. Le « Petit Robert » nous donne par exemple la définition suivante :

> « Complexe : qui se compose d'éléments différents, combinés d'une manière qui n'est pas immédiatement saisissable. »

Les mesures de complexité découlant d'une telle définition sont alors nombreuses. Lloyd [Lloyd 01], en propose une liste non exhaustive, classée en trois catégories : les mesures de difficulté de description, les mesures de difficulté de création, et les mesures de difficulté d'organisation. Dans le domaine de la complexité des images, les propositions sont moins nombreuses, mais tout aussi variées. Les méthodes utilisées peuvent par exemple

être fractales [Lam 02], floues [Petrou 06] ou, plus classiquement liées à la théorie de l'information [Rigau 05]. Concernant ce dernier point, des études récentes montrent que les méthodes liées à la complexité de Kolmogorov (mesure du taux de compression GIF ou JPEG des images) permettent d'obtenir une estimation automatique assez fidèle de l'évaluation de la complexité d'images faites par des humains [Mulhern 08, Forsythe 09].

Les méthodes évoquées plus haut évaluent la complexité des images, indépendamment de leur perception par notre système visuel et attentionnel. Nous proposons de construire une mesure basée non plus sur la complexité « directe » de leur contenu, mais sur leur complexité de perception, au travers du filtre attentionnel. Ainsi, nous mesurons la complexité des trajectoires de focalisation de notre système attentionnel afin de déterminer si une image est complexe à parcourir. Cette mesure n'a pas pour objectif de concurrencer ou remplacer les mesures traditionnelles, mais plutôt de fournir un complément autour de la complexité attentionnelle des images.

5.2 Méthodes

Notre mesure de la complexité de la trajectoire du focus d'attention est basée sur les principes de la théorie de l'information, déjà utilisés en section 3.3.3 pour évaluer la quantité d'information explorée dans une image. Comme précédemment nous utilisons la complexité de Kolmogorov. Cependant, celle-ci n'étant pas directement calculable, nous en effectuons une estimation *via* deux méthodes. L'une utilisant la compression [Mulhern 08] (principe de la *minimum description length*), l'autre l'entropie spectrale [Toh 05].

5.2.1 Taux de compression des coordonnées de la trajectoire

Cette méthode exploite le principe du *minimum description length* [Rissanen 78] afin d'estimer la complexité de la trajectoire du focus d'attention. D'après ce principe, plus la trajectoire sera complexe, plus elle sera difficile à compresser efficacement (puisqu'*a priori* non prévisible). Nous mesurons, pour une liste de coordonnées (x_i, y_i) avec $i \in [0, \ldots, nbPoints]$ le ratio entre la taille $T_{RawCoord} = 2 \times 4 \times nbPoints$ des coordonnées des focalisations lorsqu'elles sont stockées sous forme d'entiers en mémoire, et la taille de ces mêmes données, compressées avec la méthode de compression *deflat*e.

Cette méthode de compression est définie par la RFC 1951 [1]. C'est une combinaison de l'algorithme LZ77 [Ziv 77] et d'un codage de Huffman [Huffman 52]. LZ77 permet d'éliminer les séries d'octets dupliquées en les remplaçant par une référence vers la première occurrence de celle-ci. Le codage de Huffman quant à lui propose une représentation plus compacte des données, basée sur leur fréquence d'apparition. Ensemble, ces deux algorithmes permettent de réduire efficacement la taille des données sur lesquelles ils sont

1. `http://tools.ietf.org/html/rfc1951`

appliqués. Le taux de compression obtenu sera alors fonction de la complexité de celles-ci (présence de motifs / valeurs répétées, évolutions sur une plage limitée de valeurs, etc.).

En comparant les taux de compression obtenus pour différentes images avec un même temps d'observation, on va aussi classer celles-ci selon leur complexité.

5.2.2 Entropie spectrale de la longueur des saccades

L'autre méthode que nous proposons, également basée sur la théorie de l'information, travaille uniquement à partir d'une série de données s_k avec$k \in [0, \ldots, nbPoints - 2]$, formée par les distances entre les différentes focalisations successives.

On aura alors $\forall k \in [0, \ldots, nbPoints - 2]$:

$$s_k = \sqrt{(x_k - x_{k+1})^2 + (y_k - y_{k-1})^2}$$

Bien que moins riche spatialement, cette représentation permet une analyse plus fine de la dynamique. La comparaison avec la première méthode permettra d'évaluer si cette perte d'information spatiale à une influence importante sur l'évaluation de la complexité.

Nous calculons la transformée de Fourier $S = \mathcal{F}(s)$ de la série de données ainsi générée puis transformons celle-ci en une distribution statistique en divisant chaque coefficient par la somme des coefficients :

$$p(S_k) = \frac{S_k}{\sum_k S_k}$$

L'entropie spectrale est l'entropie E de cette distribution, définie par :

$$E = \sum_k p(S_k) \log(p(S_k))$$

On peut alors comparer l'entropie obtenue pour différentes images, avec un même temps d'observation.

5.3 Résultats

Nous avons mesuré l'entropie spectrale et le taux de compression des coordonnées des trajectoires des focalisations obtenues grâce à notre modèle attentionnel pour les 147 images des bases « Bruce » et « Le Meur ». Le paramétrage utilisé était celui par défaut, sans *feedback*.

Les figures 5.3.1 et 5.3.2 permettent de visualiser une partie des résultats obtenus. Les 7 images les plus simples et plus complexes sont représentées, pour des durées de simulation de 50, 150 et 300 itérations. On peut constater que le classement varie beaucoup entre $t = 50$ et $t = 150$. Au-delà, il semble être plus stable (les changements entre

$t = 150$ et $t = 300$ sont bien moins importants). Ceci est particulièrement vrai pour la mesure basée sur la compression *deflate* : les classements observés pour $t = 150$ et $t = 300$ sont très proches. La mesure basée sur l'entropie spectrale est un peu moins stable, puisqu'on constate des changements plus importants entre $t = 150$ et $t = 300$.

Le classement obtenu est cohérent avec le critère de complexité mesuré. Les images comportant un objet très saillant sont considérées comme simples. La trajectoire des focalisations nécessaires pour parcourir ces images n'est pas très complexe, puisque le système reste principalement focalisé sur l'objet le plus saillant. A l'opposé, les images ne comportant pas d'objet particulièrement saillant sont considérées comme complexes car le système attentionnel génère des focalisations dans des endroits variés. Les deux mesures donnent des résultats globalement proches, cependant, la perte d'information engendrée par l'utilisation des seules distances entre focalisations lors du calcul de l'entropie spectrale rend cette mesure moins stable.

La mesure de complexité obtenue répond à l'objectif fixé : déterminer la complexité du parcours oculaire lors de l'observation d'une image. Bien entendu, la complexité d'une image est une notion bien plus riche que cette simple mesure. Celle-ci permet cependant de contribuer à son estimation.

De plus, cette mesure peut avoir des applications directes : il peut être, par exemple, être intéressant de connaitre la complexité du parcours d'une image afin d'adapter la rééducation de patients souffrant de troubles oculaires.

5.4 Conclusion

Les deux méthodes proposées dans ce chapitre permettent une estimation itérative de la complexité d'une image, à partir de la trajectoire des focalisations générées par notre système attentionnel. Une première étude qualitative permet de valider la cohérence du classement obtenu. Une validation quantitative reste encore à effectuer, mais celle-ci est délicate car notre méthode est difficilement comparable avec un classement humain, basé sur le contenu de l'image et non sur le parcours oculaire.

Notre méthode étant également applicable aux trajectoires oculaires, il serait intéressant d'évaluer la corrélation entre le classement de complexité obtenu par notre modèle et le classement issu d'une observation humaine. Cela permettrait de valider, en partie, le comportement dynamique de notre modèle.

Enfin, puisque les deux méthodes de calcul de la complexité proposée sont complémentaires (l'une conserve les informations spatiales mais ne traite que partiellement la dynamique, l'autre supprime les informations saptiales, mais analyse correctement la dynamique) il serait intéressant de les combiner afin d'obtenir une méthode plus précise spatialement et dynamiquement..

FIGURE 5.3.1: Classement des 7 images les plus simples et les plus complexes en fonction du temps de simulation, pour la méthode « *Deflate compression ratio* ».

FIGURE 5.3.2: Classement des 7 images les plus simples et les plus complexes en fonction du temps de simulation, pour la méthode « *Saccade length fourier entropy* ».

Points clés

Positionnement

- ❏ La complexité des images est un domaine peu exploré.
- ❏ Elle est généralement estimée à partir d'une analyse de l'organisation de ses pixels.
- ❏ Nous proposons une mesure de complexité complémentaire, calculée à partir des focalisations attentionnelles nécessaires au parcours de l'image.

Contributions

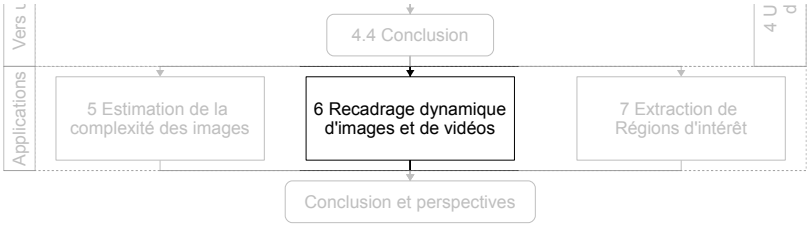
- ❏ La complexité des focalisations attentionnelles effectuées lors de l'observation d'une image peut être utilisée pour déterminer sa complexité :
  - en estimant le taux de compression de la trajectoire du focus d'attention ;
  - en calculant l'entropie spectrale de la longueur des « saccades » de la trajectoire.

Evaluation

- ❏ Les deux mesures permettent d'obtenir un classement cohérent et relativement stable au bout de 150 itérations.
- ❏ La mesure basée sur l'entropie spectrale de la longueur des « saccades » de la trajectoire, est moins stable et précise, du fait de la perte des informations spatiales lors du calcul de l'entropie.

Chapitre 6

Recadrage dynamique d'images et de vidéos



Dans ce chapitre, nous proposons une méthode de recadrage d'images et de vidéos adaptée à notre modèle d'attention. Celle-ci a la particularité d'être dynamique : le recadrage évolue en fonction du temps et peut être influencé (en temps réel) par des changements de paramétrage du modèle attentionnel. Nous verrons entre autres, que des modifications du paramètre de *feedback* permettent d'ajuster finement la stratégie de recadrage, et donc de découverte d'une image ou d'une vidéo.

6.1 Introduction

Il existe actuellement de nombreux périphériques multimédia mobiles, permettant de visualiser des images et vidéos. Ces périphériques sont généralement de petite taille (afin de pouvoir tenir dans une poche) et par conséquent, leur écran est également de taille réduite. L'affichage des photos et vidéos sur ce type d'appareil peut alors être problématique, la faible résolution de l'écran rendant le sujet principal de la photo / vidéo difficile à discerner. Pour faire face à ce problème, Le Meur [Le Meur 06] a proposé d'utiliser les informations fournies par une *heatmap* (obtenues *via* un oculomètre) ou une carte de saillance (obtenue grâce à son modèle d'attention) afin de recadrer les images

en n'en gardant que les éléments les plus saillants. Son approche est efficace, mais souffre de quelques limitations :

- dans sa version originale, elle est limitée aux images. Une extension à la vidéo, permettant de déterminer un rectangle de recadrage en fonction des N points les plus saillants de chaque trame, est proposée dans [Chamaret 09]. Les points sont extraits par un mécanisme de *Winner Takes All* permettant d'obtenir un ensemble de fixations à partir d'une carte de saillance. Cette méthode n'est cependant applicable ni aux focalisations issues de notre modèle, ni à des données obtenues *via* un oculomètre, car ceux ci ne génèrent pas assez de fixations par trame traitée ;
- le recadrage des images est totalement statique. Il pourrait être intéressant de faire découvrir l'image à l'utilisateur de manière dynamique, voire interactive.

Dans ce chapitre, nous proposons d'étendre les travaux originaux de Le Meur et présentons une méthode exploitant la dynamique de notre système attentionnel afin d'appliquer un recadrage variable dans le temps, sur des images et vidéos.

6.2 Méthode

6.2.1 Calcul du recadrage

Dans [Le Meur 06], l'auteur utilise la notion de couverture (*coverage*) définie par Wooding [Wooding 02] afin de définir une boite englobante des éléments les plus saillants de l'image. D'après cette méthode, l'estimation du taux de couverture d'une *heatmap* s'effectue en seuillant celle-ci à un niveau $S_{coverage}$ devant être défini empiriquement. Cette méthode est efficace, mais pose un problème lorsque l'on souhaite générer un recadrage dynamique (mis à jour au fur et à mesure de la construction de la *heatmap*). Dans ce cas, cette méthode s'avère très sensible à l'ajout de nouvelles focalisations et génère une animation contenant des transitions parfois trop brusques (figure 6.2.1). Afin de limiter ce problème, nous proposons une méthode basée sur le centroïde et la variance de la *heatmap*. Cette méthode permet de s'affranchir de seuil en prenant en compte tous les pixels de la *heatmap*, modulant leur influence en fonction de leur intensité. Pour une *heatmap* H, le centroïde (c_x, c_y) et les différentes variances v_g, v_d, v_h, v_b (représentant respectivement la variance des pixels situés à gauche, à droite, au-dessus et en-dessous du centroïde) sont calculés comme suit :

$$c_x = \frac{\sum_x \sum_y xH(x,y)}{\sum_x \sum_y H(x,y)} \qquad c_y = \frac{\sum_x \sum_y yH(x,y)}{\sum_x \sum_y H(x,y)}$$

et

$$
\begin{aligned}
v_g &= \sqrt{\frac{\sum\sum \delta^-(x,c_x)^2 H(x,y)}{\sum\sum \delta^-(x,c_x)^2}} \\
v_d &= \sqrt{\frac{\sum\sum \delta^+(x)^2 H(x,y)}{\sum\sum \delta^+(x)^2}} \\
v_h &= \sqrt{\frac{\sum\sum \delta^-(y,c_y) H(x,y)}{\sum\sum \delta^-(y,c_y)^2}} \\
v_b &= \sqrt{\frac{\sum\sum \delta^+(y,c_y) H(x,y)}{\sum\sum \delta^+(y,c_y)^2}}
\end{aligned}
\quad \text{avec} \quad
\begin{aligned}
\delta^-(i,j) &= \begin{cases} i-j & si\ i < j \\ 0 & sinon \end{cases} \\
\delta^+(i,j) &= \begin{cases} i-j & si\ i \geq j \\ 0 & sinon \end{cases}
\end{aligned}
$$

On peut alors construire le rectangle R englobant les éléments les plus saillants de la scène :

$$
\begin{aligned}
x_R &= c_x - \alpha v_g \\
y_R &= c_y - \alpha v_h \\
W_R &= \alpha(v_d - v_g) \\
H_R &= \alpha(v_b - v_h)
\end{aligned}
$$

avec (x_R, y_R) les coordonnées du coin haut gauche du rectangle R, W_R et H_R sa largeur et sa hauteur et α un paramètre permettant d'ajuster la quantité d'éléments saillants à garder.

Dans la suite de ce chapitre, nous avons utilisé une valeur de $\alpha = 2$. Pour une *heatmap* contenant une seule focalisation (représentée par une gaussienne), cela correspond à garder 95% de la saillance. Ceci n'est bien sûr valable que dans le cas d'une seule focalisation. Ensuite, la distribution n'est plus gaussienne et la part de saillance représentée chute.

Notre système est ainsi moins sujet aux agrandissements brutaux du rectangle de recadrage lors de l'ajout de nouvelles focalisations, mais ce phénomène reste encore parfois problématique. Pour le résoudre nous proposons d'ajouter de l'inertie lors de la mise à jour du rectangle de recadrage. On aura alors une procédure de mise à jour itérative :

$$
\begin{aligned}
x_R(t) &= inertie \times x_R(t-1) + (1 - inertie) \times (c_x - \alpha v_g) \\
y_R(t) &= inertie \times y_R(t-1) + (1 - inertie) \times (c_y - \alpha v_h) \\
W_R(t) &= inertie \times W_R(t-1) + (1 - inertie) \times \alpha(v_d - v_g) \\
H_R(t) &= inertie \times H_R(t-1) + (1 - inertie) \times \alpha(v_b - v_h)
\end{aligned}
$$

De cette façon, on pourra filtrer les variations importantes du rectangle de recadrage, au prix toutefois d'une moins grande réactivité.

Le système attentionnel permettant de générer incrémentallement une *heatmap* au fur et à mesure que la simulation se déroule, il devient possible d'utiliser le mécanisme décrit ci-dessus afin de créer une animation de découverte de l'image. Le recadrage est

dynamique, mais ne permet pas une réelle découverte séquentielle des différents éléments de la scène (figure 6.2.2). Nous présentons dans la sous-section suivante, différentes façons d'utiliser les propriétés dynamiques et d'adaptation de notre système attentionnel afin d'améliorer ce parcours et d'étendre le principe à la vidéo.

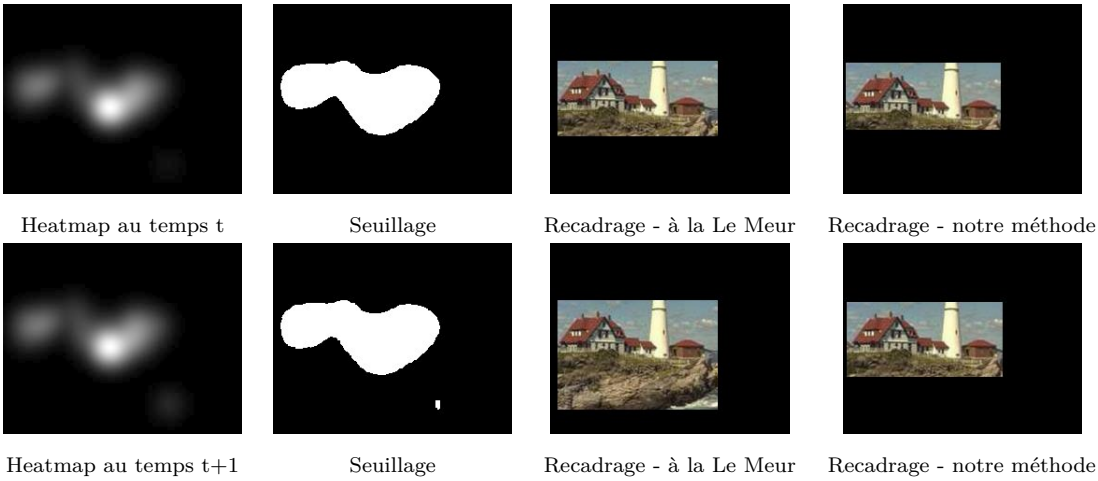


FIGURE 6.2.1: Sensibilité des méthodes de recadrage à l'ajout de nouvelles focalisations. En évitant l'usage d'un seuil, notre méthode est moins sensible aux nouvelles focalisations : le recadrage dynamique est plus fluide.

FIGURE 6.2.2: Exemple de recadrage dynamique simple. L'image est découverte progressivement, mais trop rapidement (un recadrage stable est obtenu dès t= 50, ensuite le système n'évolue quasiment plus). L'algorithme ne permet pas un parcours séquentiel des différents éléments saillants de l'image

6.2.2 Recadrage avancé

Afin d'obtenir un recadrage plus conforme à nos attentes, ainsi que pour permettre un meilleur contrôle sur le déroulement du recadrage dynamique, nous avons utilisé deux mécanismes simples.

Ajout d'un facteur d'oubli lors de la génération de la heatmap

Jusqu'à présent, la *heatmap* générée à partir des différentes focalisations de notre système attentionnel était générée à partir de l'équation 3.2.3 (présentée au chapitre 3) et que nous rappelons ici :

$$HM_s(x,y) = \left(\sum_{i=1}^{N}(\delta_{x,y}^{x_i,y_i})\right) * g_{\sigma_x,\sigma_y}(x,y) \tag{6.2.1}$$

En réalité, la *heatmap* est construite itérativement à chaque pas de la simulation. On a alors :

$$HM_s(x,y,t) = HM_s(x,y,t-1) + \left(\delta_{x,y}^{x_t,y_t} * g_{\sigma_x,\sigma_y}(x,y)\right) \tag{6.2.2}$$

avec (x_t, y_t)les coordonnées de la focalisation générée au pas de temps t.

Le recadrage sera plus dynamique s'il dépend d'une estimation de la saillance calculée principalement à partir des focalisations les plus récentes. Ainsi, nous souhaitons que les focalisations plus anciennes soient oubliées progressivement lors de l'évolution du système. Pour cela, il suffit de multiplier $HM_s(x,y,t-1)$ par un facteur d'oubli $forgetting \in [0,1]$:

$$HM_s(x,y,t) = forgetting \times HM_s(x,y,t-1) + \left(\delta_{x,y}^{x_t,y_t} * g_{\sigma_x,\sigma_y}(x,y)\right) \tag{6.2.3}$$

Avec une valeur de $forgetting = 1$, la *heatmap* est calculée comme auparavant : elle représente la répartition de la saillance sur l'image pour toute la durée de simulation. Avec une valeur plus faible (typiquement $forgetting = 0.95$), la *heatmap* représente la répartition des focalisations pour une période de temps beaucoup plus courte. On obtient alors une estimation de la saillance plus locale dans le temps.

Dans le cas du recadrage d'images et vidéos, l'utilisation d'une telle *heatmap* permet de rendre le système beaucoup plus dynamique : on ne présente à l'utilisateur que la partie de l'image ayant attiré l'attention de notre système récemment. Ainsi, le recadrage évolue en permanence.

Utilisation du facteur de *feedback*

L'objectif du recadrage dynamique proposé dans ce chapitre est de faire découvrir progressivement l'image à l'utilisateur, en étant guidé par la saillance des différents éléments de la scène. Ainsi, le recadrage sera de meilleure qualité si on peut contrôler le degré de focalisation du système attentionnel à chaque instant. C'est le rôle du facteur de *feedback* introduit au chapitre 4. Nous proposons d'utiliser ce mécanisme de *feedback* de deux manières :

– pour les images fixes, on fera varier progressivement le *feedback* de 1 (focalisation importante sur les éléments les plus saillants) à -1 (exploration de toute la scène).

Ainsi, on devrait balayer l'image en faisant d'abord découvrir les quelques éléments les plus saillants, puis en élargissant le champ de recadrage afin d'afficher l'ensemble de la scène.
- pour les vidéos, le facteur de *feedback* pourra être choisi en fonction du type de recadrage voulu (plan large ou serré) ainsi que du type de vidéo (film, sport, actualité, etc.).

6.3 Résultats

Les résultats présentés dans cette section permettent d'estimer qualitativement les résultats obtenus en utilisant les mécanismes de recadrage avancés présentés dans la section précédente. Les exemples proposés concernent des images statiques ainsi que différentes séquences vidéo.

6.3.1 Images

La figure 6.3.1, donne un aperçu du recadrage dynamique effectué sur différentes images en faisant varier progressivement le *feedback* de 1 à -1. Les paramètres du système attentionnels étaient ceux par défaut, les valeurs de $forgetting$ et $inertie$ respectivement de 0.95 et 0.9.

Dans la première moitié des séquences, les images sont parcourues en plan serré et les éléments les plus saillants visités. Dans la seconde moitié, le cadrage s'élargit, laissant apparaitre la quasi totalité des images. Une vidéo permettant de mieux apprécier l'évolution de la découverte dynamique des images est disponible sur : `http://www.youtube.com/watch?v=Tl9-MpJrAKg`.

6.3.2 Vidéos

Nous avons également appliqué notre algorithme de recadrage dynamique à des séquences vidéo de différents types :
- Une vidéo de sport : « TOP 20 Tennis Master Points » (figure 6.3.2). Comme son nom l'indique cette séquence est une compilation de différents extraits de matchs de tennis.
- Une publicité : « Levis - Mr Oizo » (figure 6.3.3).
- Une bande annonce du film « Hancock » (figure 6.3.4).

Chacune des vidéos a été traitée avec une valeur de *feedback* différente (dont la justification est donnée dans les prochains paragraphes) . Les autres paramètres sont communs : le système attentionnel était configuré avec ses valeurs par défaut, l'inertie était fixée à 0.5.

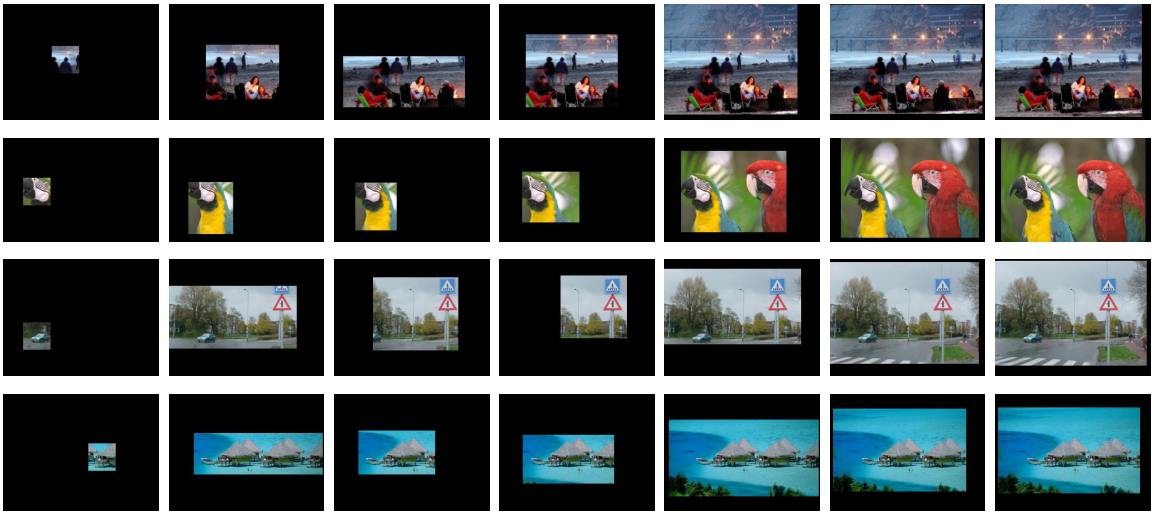


FIGURE 6.3.1: Différents exemples de découverte dynamique d'image avec variation progressive du *feedback* de 1 (très focalisé) à -1 (forte exploration). Les trames représentées correspondent respectivement aux couples (nombre d'itérations / *feedback)* suivants : (0 / -1), (50 / -0.66), (100 / -0.33), (150 / 0), (200 / 0.33), (250 / 0.66) et (300 / 1) (de gauche à droite).

La vidéo de tennis a été traitée de manière particulière. L'objectif étant de suivre les joueurs, nous avons utilisé deux valeurs de *feedback* positives : une première à 0.5 (focalisation forte) afin de suivre au plus près les déplacements du / des joueur(s), une seconde à 0.25 permettant (en théorie) un suivi plus large des joueurs. Les vidéos complètes correspondant aux résultats obtenus sont consultables sur youtube aux adresses suivantes : `http://www.youtube.com/watch?v=gpQSYY-GXlO` et `http://www.youtube.com/watch?v=MZQlWOHIwkg`.

Les deux autres vidéos ont été traitées avec le même jeu de valeurs : aucun *feedback*, afin de voir l'impact du recadrage avec un système attentionnel non contraint, et un *feedback* de 0.25 afin d'observer comment se comporte le système lorsque l'attention est un peu plus focalisée. Les vidéos complètes des résultats sont consultables aux adresses suivantes : `http://www.youtube.com/watch?v=-UHMRneYLDE` et `http://www.youtube.com/watch?v=CtHdWJUGqSM` (« Levis - Mr Oizo »), `http://www.youtube.com/watch?v=0QgCLHtLlpw` et `http://www.youtube.com/watch?v=0CCjmZAW584` (« Hancok »).

On peut constater que globalement le système propose un recadrage cohérent : les éléments principaux restent dans le cadre, et la vidéo recadrée est toujours compréhensible. Le pourcentage moyen de surface de la vidéo originale affichée dans sa version recadrée est directement proportionnel au facteur de *feedback* :

– 38% pour un *feedback* de 0.5 ;

– 48% pour un *feedback* de 0.25 ;
– 70% pour un *feedback* de 0.

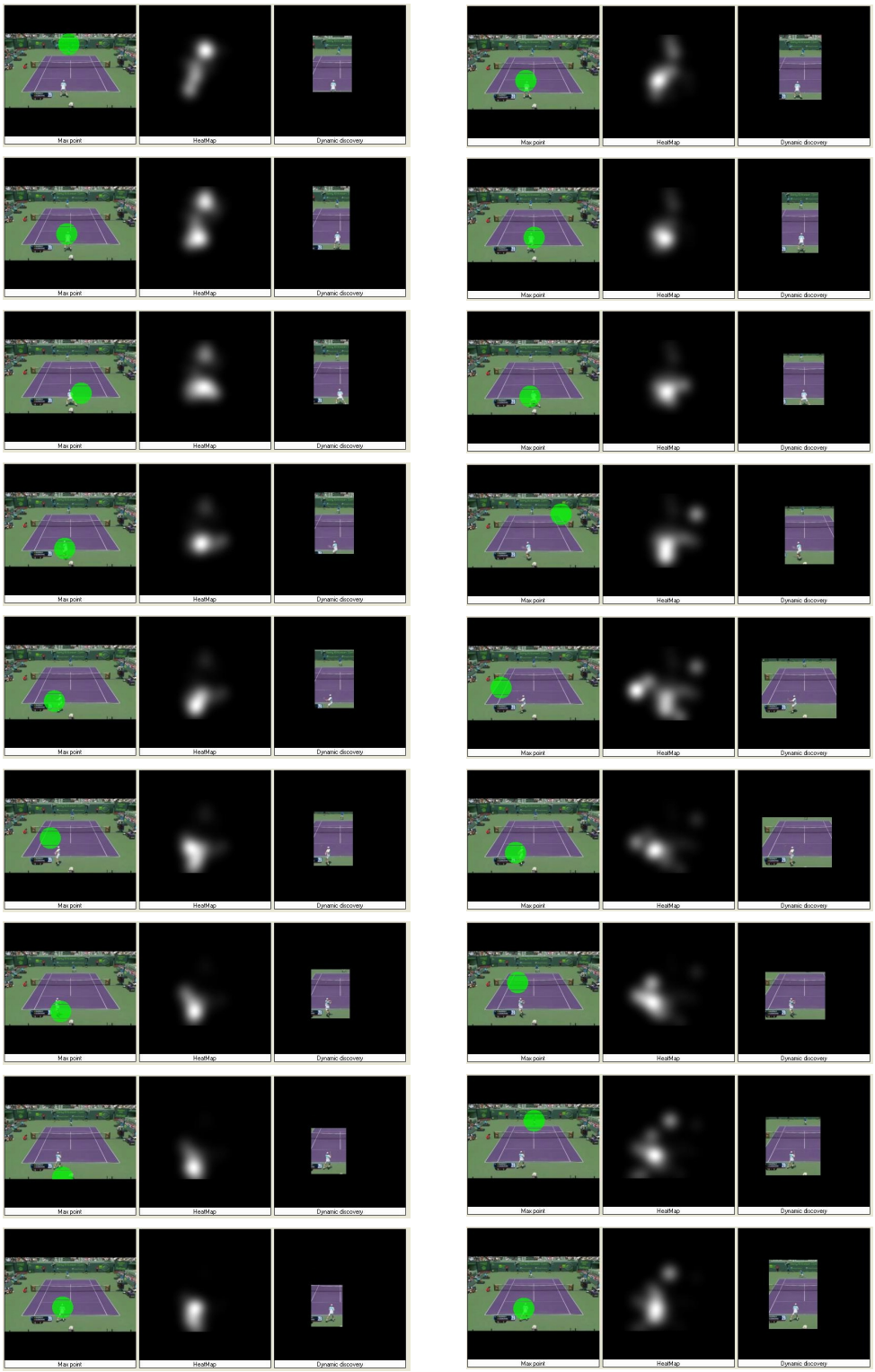
6.4 Conclusion

Dans ce chapitre nous avons proposé une approche dynamique du recadrage d'image et de vidéo. Celle-ci apporte de nouvelles possibilités par rapport aux travaux originaux de Le Meur [Le Meur 06].

Dans le cas de l'image, notre approche élargit le champ d'application du recadrage en ne le cantonnant plus qu'à la seule optimisation d'affichage pour périphériques mobiles. Le fait que le recadrage puisse être dynamique et interactif ouvre la voie à une autre façon de découvrir les images. Une application potentielle pourrait par exemple être une extension / amélioration des logiciels existants de lecture de bandes-dessinées sur ordinateur et téléphones mobiles [1]. Actuellement ces logiciels de lecture proposent un parcours animé de chaque vignette de la bande-dessinée. Ces parcours sont construits « à la main » par des opérateurs humains. Notre algorithme de recadrage dynamique pourrait permettre d'automatiser tout ou partie de ce fastidieux travail.

Concernant les vidéos, le recadrage dynamique basé sur la saillance permet de réduire la taille des données à afficher tout en en préservant le sens général. L'application initiale d'Olivier Le Meur (le recadrage à destination de périphériques mobiles) est alors également possible pour les flux vidéos. On pourra également exploiter ce procédé pour des applications classiques de compression de données (compression plus forte dans les zones non recadrées) ou de streaming adaptatif (ajustement du *feedback* en fonction de la bande passante disponible afin de dégrader plus ou moins la vidéo dans les zones non recadrées).

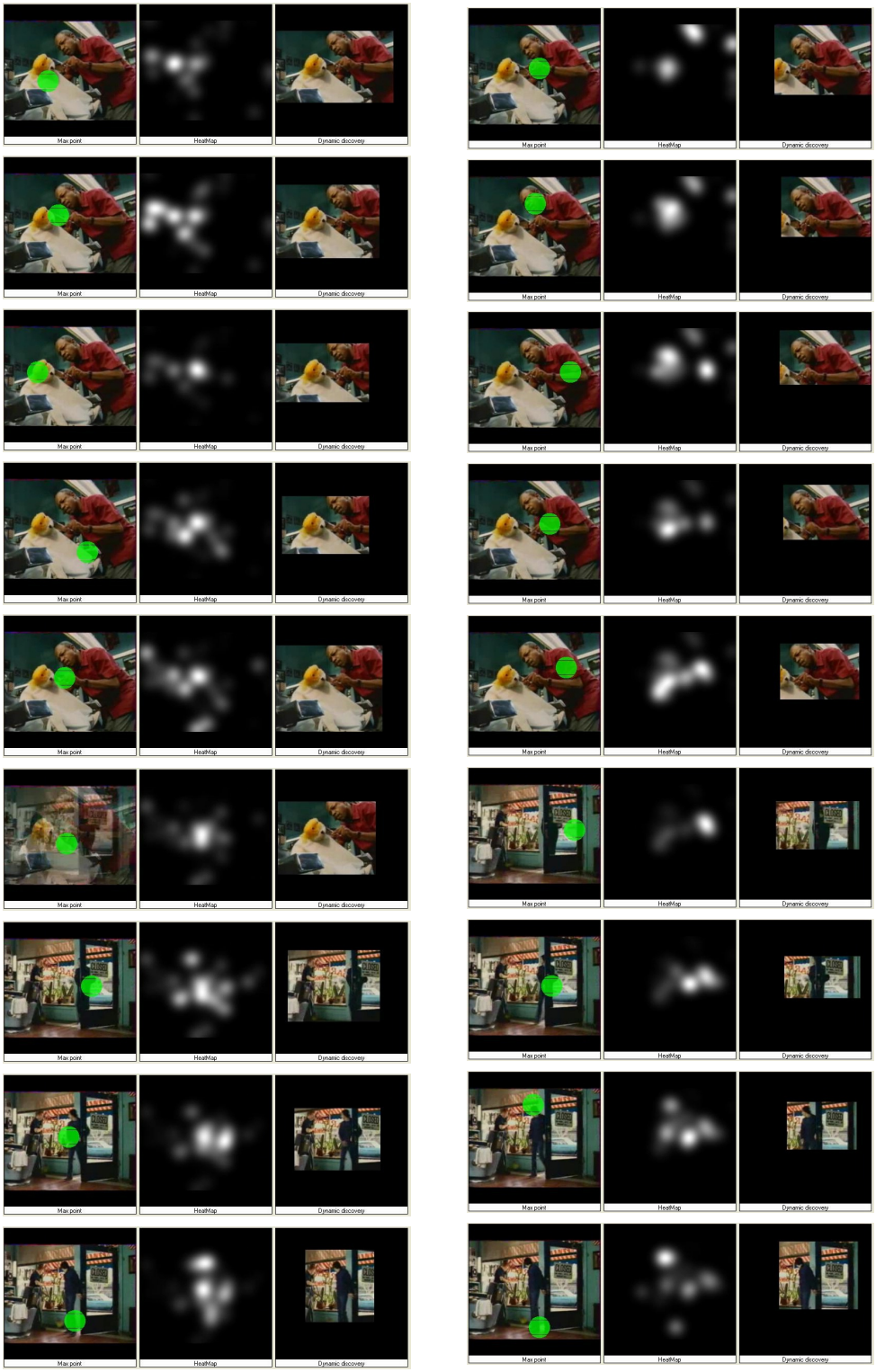
1. Ce type de service est actuellement proposé par des sociétés comme Avé Comics (`http://www.avecomics.com`).



(a) *Feedback* = 0.5, le taux moyen de surface affiché est de 38%.

(b) *Feedback* = 0.25, le taux moyen de surface affiché est de 46%.

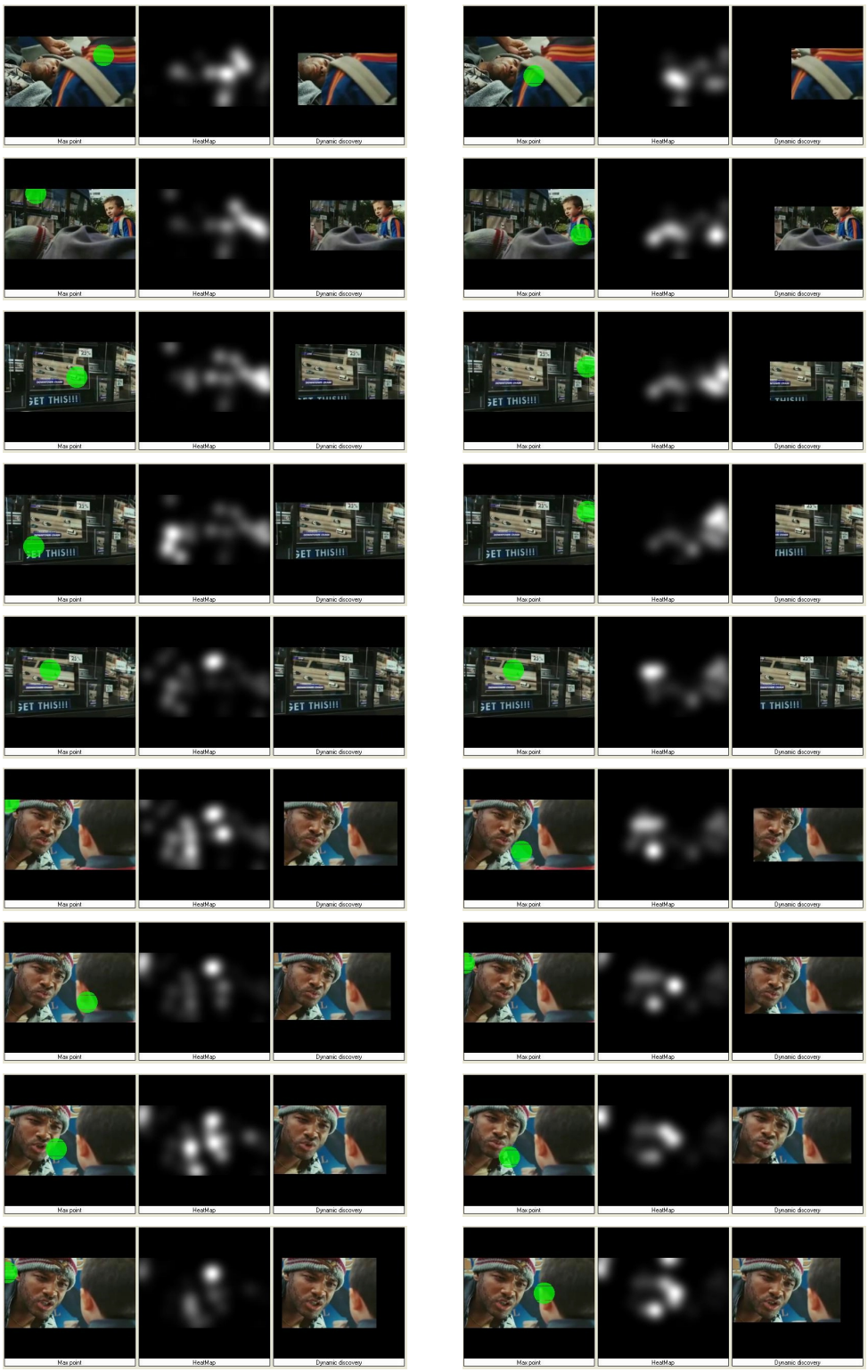
FIGURE 6.3.2: Quelques trames de la vidéo « TOP 20 Tennis Master Points ».



(a) *Feedback* = 0, le taux moyen de surface affiché est de 69%.

(b) *Feedback* = 0.25, le taux moyen de surface affiché est de 51%.

FIGURE 6.3.3: Quelques trames de la publicité « Levis - Mr Oizo ».



(a) *Feedback* = 0, le taux moyen de surface affiché est de 70%.

(b) *Feedback* = 0.25, le taux moyen de surface affiché est de 49%.

FIGURE 6.3.4: Quelques trames de la bande-annonce du film « Hancok ».

Points clés

Positionnement

- ❏ Pour faciliter leur consultation sur des périphériques mobiles posédant un écran de petite taille, il peut être intéressant de recadrer des images ou vidéo, en fonction de leur saillance. Le Meur propose une solution à ce problème, mais celle-ci ne permet pas d'exploiter les spécificités de notre modèle.
- ❏ Nous étendons cette proposition par un recadrage dynamique, ajustable en temps réel, pouvant être appliqué sur des images fixes et des vidéos.

Contributions

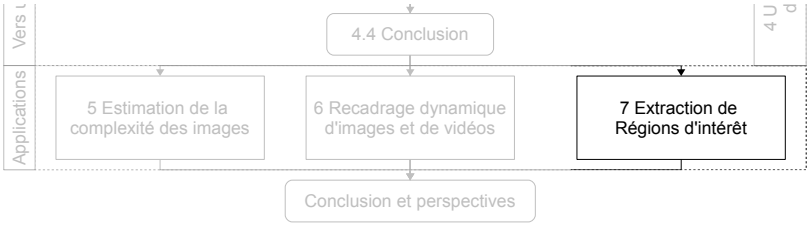
- ❏ Le calcul de la « boite de recadrage » d'une image, basé sur la variance de la *heatmap* permet de s'affranchir du seuil de saillance utilisé par Le Meur. Le recadrage dynamique peut alors s'effectuer avec des transitions plus douces.
- ❏ L'utilisation d'un facteur d'oubli lors de la construction itérative d'une *heatmap* permet d'obtenir une représentation temporellement limitée de la saillance de l'image.
- ❏ Le recadrage dynamique peut être ajusté en temps réel, notamment grâce au paramètre de *feedback* du modèle attentionnel.

Évaluation

- ❏ Le recadrage permet une découverte dynamique des différents éléments saillants d'une image.
- ❏ Globalement, le recadrage préserve le sens des vidéos. Les éléments essentiels sont affichés.
- ❏ Le paramètre de *feedback* du modèle attentionnel agit directement sur le pourcentage de surface d'image affiché.

Chapitre 7

Extraction de régions d'intérêt



Dans ce chapitre nous proposons de compléter notre modèle d'attention afin de le rendre plus compatible avec la plupart des systèmes de vision, qui travaillent généralement à partir de régions ou d'objets. En se basant sur les différentes focalisations produites, nous générons un ensemble de régions d'intérêt. Celles-ci évoluent au cours du temps, en fonction des nouvelles focalisations générées par notre modèle.

7.1 Introduction

Notre algorithme attentionnel, comme de nombreux autres modèles, est basé sur une représentation spatiale de l'attention. Ainsi, il propose un nouveau point de focalisation à chaque pas de temps de la simulation. Cependant, la plupart des algorithmes de vision ou de traitement d'images, ne travaillent pas sur des points d'intérêt, mais sur des régions d'intérêt (le plus souvent rectangulaires). Nous savons également que le système attentionnel humain ne fonctionne pas que de manière spatiale et qu'une partie au moins des processus attentionnels est basée sur des objets [Sun 03, Sun 08] ou proto-objets [Walther 06b, Orabona 08]. Ainsi, il parait nécessaire de faire le lien entre attention visuelle (fixations oculaires) et régions d'intérêt [Privitera 00].

Pour combler le vide entre les fixations générées par notre système attentionnel et les régions d'intérêt traitées par la plupart des systèmes de vision, nous proposons un nouveau mécanisme. Celui-ci permet de générer un ensemble de régions d'intérêt, classé par ordre de saillance, en fonction de la *heatmap* construite à partir des focalisations générées par notre système attentionnel. Le mécanisme proposé est également applicable aux *heatmaps* générées à partir de fixations oculaires humaines.

7.2 Méthode

La méthode de segmentation d'images et vidéos en régions d'intérêt proposée est basée sur les maximums locaux de la *heatmap* générée par notre modèle attentionnel (figure 7.2.1b). Cette *heatmap* étant construite itérativement à chaque pas de simulation du système attentionnel, les régions segmentées évoluent donc au même rythme. On dispose alors d'une méthode de segmentation *anytime*, permettant d'avoir très tôt un estimatif des régions les plus intéressantes, puis d'affiner (ou mettre à jour en fonction du contexte et des directives reçues par le système attentionnel) son estimation initiale. Comme précisé plus haut, la segmentation est basée sur les maximums locaux de la *heatmap*. En effet, ces maxima résument l'ensemble des zones que le système attentionnel a visité. De plus, leur intensité est directement proportionnelle au nombre de focalisations effectuées dans la zone et donc (dans le cas de notre système attentionnel) de la saillance de la zone.

Notre algorithme de segmentation est découpé en trois étapes :
- détermination des maximums locaux de la *heatmap* ;
- regroupement (*clustering*) des points de focalisation autour des différents maximums locaux ;
- construction du rectangle englobant de chaque région d'intérêt à partir des maximums locaux et de leurs focalisations associées.

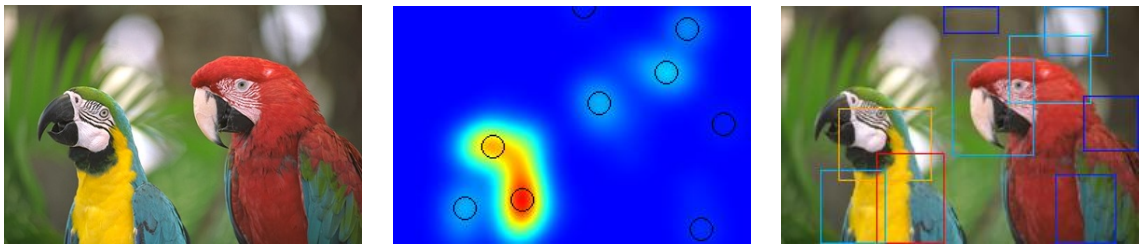
Ces étapes sont décrites dans les prochaines sous-sections.

7.2.1 Détermination des maximums locaux de la *heatmap*

Il existe différentes méthodes de recherche de maximums locaux dans une image : changement de signe de la dérivée, passage par zéro de la dérivée seconde (avec ou sans filtrage préalable), etc. Cependant, dans notre cas, les données à traiter ont été générées à partir d'une somme de gaussiennes, elles ne sont pas bruitées et ne nécessitent donc pas l'utilisation d'une méthode très robuste. Ainsi, nous utilisons une méthode simple, basée sur la comparaison de chaque pixel avec son voisinage 8×8 (algorithme 7.1). Afin d'éviter la prolifération de maximums trop proches, on peut filtrer les résultats obtenus en regroupant tous les maximums dont la distance est inférieure à un seuil donné (la moitié de la taille de la gaussienne utilisée lors de la création de la *heatmap* par exemple).

Algorithm 7.1 Détermination des maximums locaux de la *heatmap*.

// ENTREES
// H : la heatmap dont on cherche à calculer les maximums locaux
// SORTIES
// $Maximums$: liste de maximums
pour tous les pixels de H **faire**
  si $H[x,y] \geq H[x-1,y-1]$
  et $H[x,y] \geq H[x,y-1]$
  et $H[x,y] \geq H[x+1,y-1]$
  et $H[x,y] \geq H[x-1,y]$
  et $H[x,y] \geq H[x+1,y]$
  et $H[x,y] \geq H[x-1,y+1]$
  et $H[x,y] \geq H[x,y+1]$
  et $H[x,y] \geq H[x+1,y+1]$
  et $H[x,y] > \frac{1}{8}(H[x-1,y-1] + H[x,y-1] + H[x+1,y-1] + H[x-1,y] + H[x+1,y] + H[x-1,y+1] + H[x,y+1] + H[x+1,y+1])$ **alors**
    ajouter (x,y) A $Maximums$
  fin si
fin pour
renvoyer $Maximums$



(a) Image source

(b) *Heatmap* et maximums locaux. Les cercles noirs sont centrés sur les maximums locaux.

(c) Résultat de la segmentation. La couleur des rectangles indique la saillance des différentes régions (rouge : très saillant, bleu : très peu saillant).

FIGURE 7.2.1: Exemple de segmentation obtenue pour la *heatmap* de l'image « Parrots » à t=150.

7.2.2 Regroupement des points de focalisation

Le regroupement des points de focalisation peut également être effectué avec une multitude d'algorithmes de *clustering* guidé. Cependant, puisque nous connaissons une bonne estimation des centres des *clusters* que nous cherchons à obtenir, nous avons opté pour l'algorithme K-mean (figure 7.2.2). Les maximums locaux obtenus à l'étape précédente permettent de déterminer le nombre de *clusters* voulu, ainsi que l'estimation initiale des centroïdes des clusters. On exécute alors l'algorithme un nombre de fois suffisant pour stabiliser les classes. Théoriquement, nous devrions utiliser un critère d'arrêt estimant la variabilité du *clustering* d'une itération à l'autre, mais étant donné que notre estimation initiale des centres des *clusters* est particulièrement adaptée aux données, l'algorithme converge rapidement. On peut alors utiliser un nombre d'itérations fixe (ici 25).

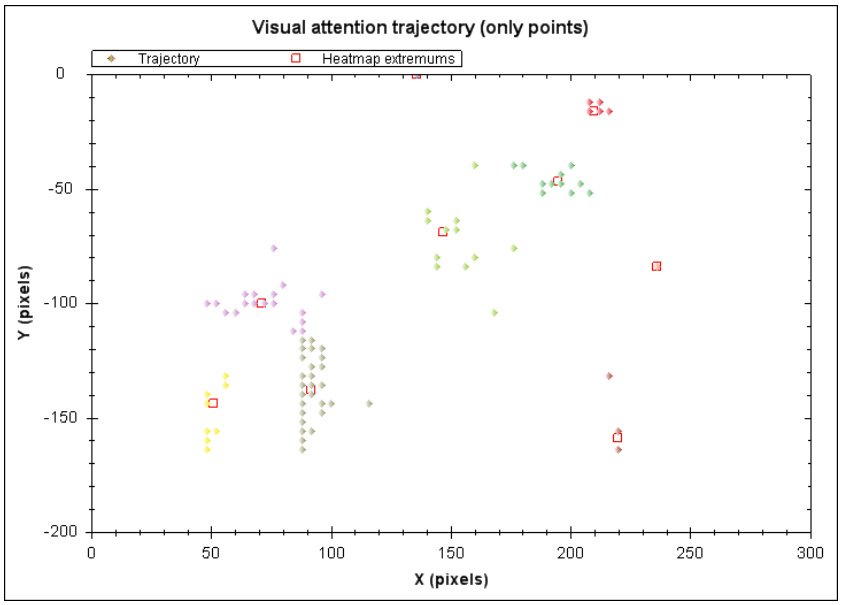


FIGURE 7.2.2: Exemple de *clustering* obtenu pour l'image « Parrots », à t=150. Les focalisations (losanges) sont regroupées en différentes classes (auxquelles une couleur est associée) autour des maximums locaux de la *heatmap* (carrés rouges).

7.2.3 Construction des rectangles englobant des régions

Nous utilisons la même méthode que dans la section 6.2.1. Nous ne travaillons cependant plus à partir de tous les pixels de la *heatmap*, mais des focalisations associées à chaque maximum local de celle-ci.

Les rectangles ainsi obtenus doivent voir leur taille ajustée en fonction des dimensions de la gaussienne utilisée pour générer les *heatmaps* (qui dépendent du paramètre

$foveaSize$, défini en section 3.2.1.3). En effet, si cet ajustement n'est pas effectué, un cluster ne comportant qu'un seul point génèrera un rectangle de taille nulle. On aura ainsi :

$$\begin{aligned} x_R &= c_x - (\alpha v_g + L \times 0.5 \times foveaSize) \\ y_R &= c_y - (\alpha v_h + L \times 0.5 \times foveaSize) \\ W_R &= \alpha(v_d - v_g) + L \times foveaSize \\ H_R &= \alpha(v_b - v_h) + L \times foveaSize \end{aligned}$$

avec $L = \max(W, H)$ où W et H sont respectivement la largeur et la hauteur de l'image d'origine.

Pour chaque région R_i ainsi segmentée, on calcule un indice de saillance relative SR_i, représentant la part de saillance de chacune des régions dans la saillance totale de l'image. Celui-ci est associée à un maxima M_i de la *heatmap* H (avec $i \in [0, \ldots, nbMaximums - 1]$), de la manière suivante :

$$SR_i = \frac{H[M_i]}{\sum_k H[M_k]}$$

Notons qu'ici, on approxime la saillance totale de l'image par la somme des maximums de la *heatmap*.

7.2.4 Segmentation de vidéos

Le cas particulier des vidéos est abordé de manière similaire au chapitre précédent. Au lieu d'utiliser une *heatmap* classique, nous utilisons une *heatmap* dynamique (typiquement, avec un facteur d'oubli $forgetting = 0.95$). Ainsi les régions générées ne correspondent qu'à des zones ayant une saillance temporellement limitée.

Quelques modifications doivent cependant être appliquées à l'algorithme de segmentation afin de prendre en compte l'oubli progressif des focalisations les plus anciennes. Ainsi, lors de la mise à jour des centroïdes des clusters du K-mean, il est nécessaire de pondérer l'influence de chaque focalisation (x_i, y_i) par un facteur $AgeWeight_i$ proportionnel à son âge :

$$AgeWeight_i = \begin{cases} 1.0 & si\ forgetting = 1.0 \\ forgetting^{N-1-i} & sinon \end{cases}$$

avec N le nombre de focalisations depuis le début de la simulation attentionnelle et $i \in [0, \ldots, N-1]$ (la dernière focalisation effectuée sera donc(x_{n-1}, y_{n-1})).

$AgeWeight_i$ baissant très rapidement lorsque i diminue, on peut optimiser le calcul

des clusters en ne travaillant que sur les K dernières focalisations, correspondant à $AgeWeight_i \geq Seuil_{age}$. On aura alors :

$$\begin{aligned} forgetting^{N-1-K} &\geq Seuil_{age} \\ \log(forgetting^{N-1-K}) &\geq \log(Seuil_{age}) \\ (N-1-K)\log(forgetting) &\geq \log(Seuil_{age}) \\ K &\geq (N-1) - \frac{\log(Seuil_{age})}{\log(forgetting)} \end{aligned}$$

Si on veut ignorer les focalisations dont le poids sera inférieur à 1%, on ne traitera donc que les $i \in [N-1-90, \ldots, N-1]$, soit les 91 dernières focalisations.

7.3 Résultats

Dans cette sous-section, nous proposons une estimation qualitative de la segmentation obtenue sur différentes images et vidéos.

7.3.1 Images

Notre modèle d'attention étant un modèle dynamique, le focus d'attention, la *heatmap* et la segmentation d'image associée évoluent à chaque pas de la simulation. La figure 7.3.1 permet de comparer les segmentations obtenues après des temps de 50, 100, 150 et 300 itérations. Le modèle a été utilisé avec ses paramètres par défaut, sans *feedback* ni facteur d'oubli.

On peut effectuer plusieurs constatations :

- la plupart des éléments les plus saillants de l'image sont correctement segmentés dès t=50, mais la saillance relative de certains éléments peut être sous-estimée en début de simulation. Cela est corrigé ensuite.
  Ce phénomène correspond à la définition d'un algorithme *anytime* : les premiers résultats sont une approximation grossière de la solution escomptée, celle-ci s'affine avec le temps ;
- il y a une variabilité dans le classement des régions (la vignette en $3^{\text{ème}}$ position à $t = 100$ et $t = 150$ n'est plus que $5^{\text{ème}}$ et $6^{\text{ème}}$ à $t = 300$ et $t = 600$), mais celle-ci reste limitée et ne touche que les régions moyennement saillantes. Un élément saillant le sera généralement tout au long de la simulation. S'il devient moins saillant ce ne sera que temporairement afin de laisser le temps au système de mettre en avant d'autres éléments potentiellement importants ;
- certains éléments de l'image peuvent être présents dans plusieurs régions (les rectangles se chevauchent). C'est un effet de bord dû au *clustering* des focalisations autour des différents maximums. Celui-ci est cependant limité puisque les extremums ont été filtrés pour être suffisamment distants les uns des autres ;

– les régions et leur classement ne se stabilisent jamais : le système est en constante évolution.

Notre modèle dynamique d'attention permet donc, pour le système de vision qui l'intègre, de bénéficier d'une estimation des zones les plus saillantes de l'image pouvant s'affiner au cours du temps. On pourra également ajuster le paramétrage du système attentionnel en cours de simulation, en fonction des besoins du système de vision hôte. L'aspect dynamique de notre système sera alors encore mieux exploité.

La figure 7.3.2 permet de juger de la qualité de la segmentation pour différentes images d'exemple (après un temps de simulation de 150 itérations). La qualité des résultats est à apprécier en fonction de la saillance relative de chacune des régions extraites. Pour les zones de moyenne et forte saillance ($SR > 10\%$) la qualité des vignettes obtenues est bonne : les vignettes représentent des zones indiscutablement saillantes. Pour des valeurs de saillance relative plus faible, la qualité est difficile à apprécier car les régions extraites ne sont liées qu'à quelques focalisations, pas forcement très significatives. Ainsi, en cas d'exploitation des résultats de la segmentation par un algorithme de vision il faudrait filtrer les résultats en fonction de la somme de leur saillance relative (par exemple, les N vignettes dont la somme représente 80% de la saillance) et non simplement par leur ordre (ici, les 7 vignettes les plus saillantes).

7.3.2 Vidéos

L'évaluation, même qualitative, des résultats de la segmentation sur des vidéos est plus problématique que dans le cas de l'image. Dans un cadre général, il n'est pas évident de déterminer subjectivement quelles régions sont spatio-temporellement saillantes. Dans cette section, nous avons donc choisi d'appliquer notre algorithme de segmentation sur des vidéos d'un type particulier : le trafic routier. Pour ce type de vidéos, la définition des éléments saillants (spatiallement et spatio-temporellement) est plus aisée. On y retrouve par exemple :
- les panneaux routiers ;
- le clignotant des autres voitures ;
- les feux de signalisation (rouge, orange, vert) ;
- les mouvements des piétons et véhicules traversant transversalement la chaussée.

Ainsi, nous présentons en figure 7.3.3, quelques trames caractéristiques d'une vidéo bien plus longue, représentant un trajet routier en périphérie de Paris. La vidéo complète est consultable en ligne à l'adresse suivante : `http://www.youtube.com/watch?v=Yotxq9_vjPg`. Notre algorithme a été appliqué sur cette vidéo avec ses paramètres par défaut, plus un *feedback* de 0.25 (pour éviter que le système ne bascule trop rapidement d'un élément à l'autre) et un facteur d'oubli de 0.95 (pour que la *heatmap* générée soit

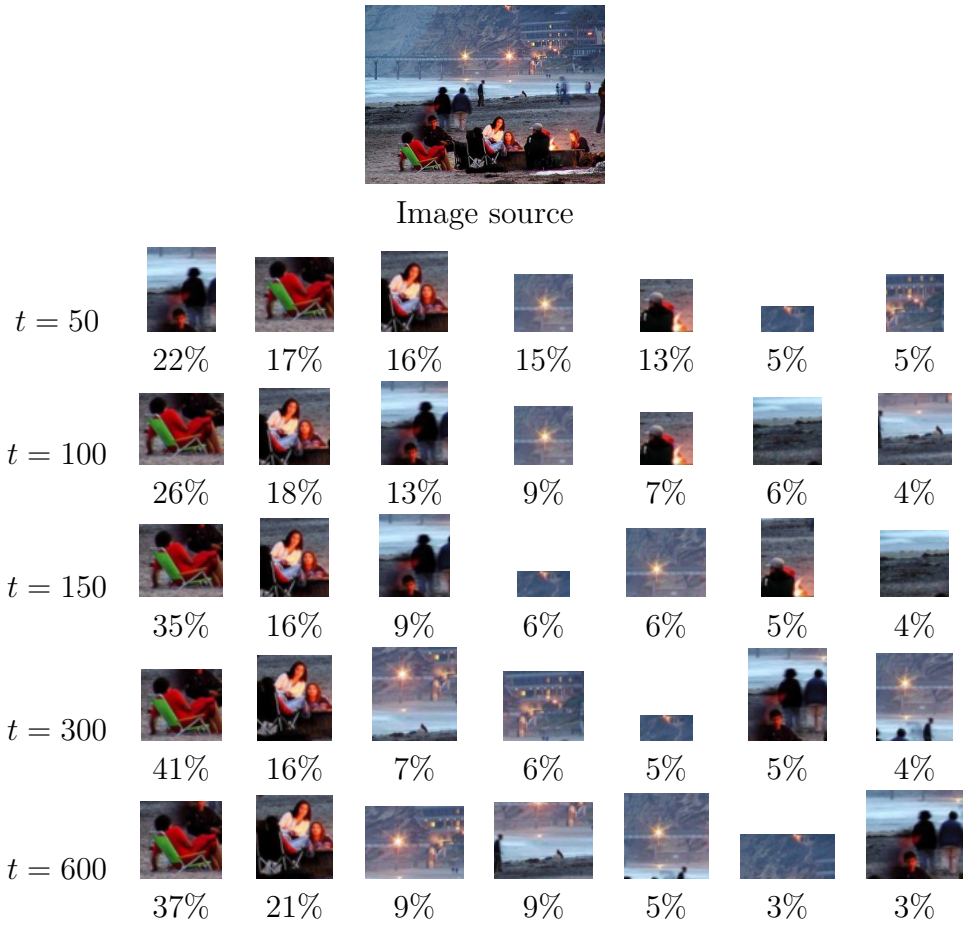


FIGURE 7.3.1: Évolution des 7 vignettes les plus saillantes en fonction du temps de simulation.

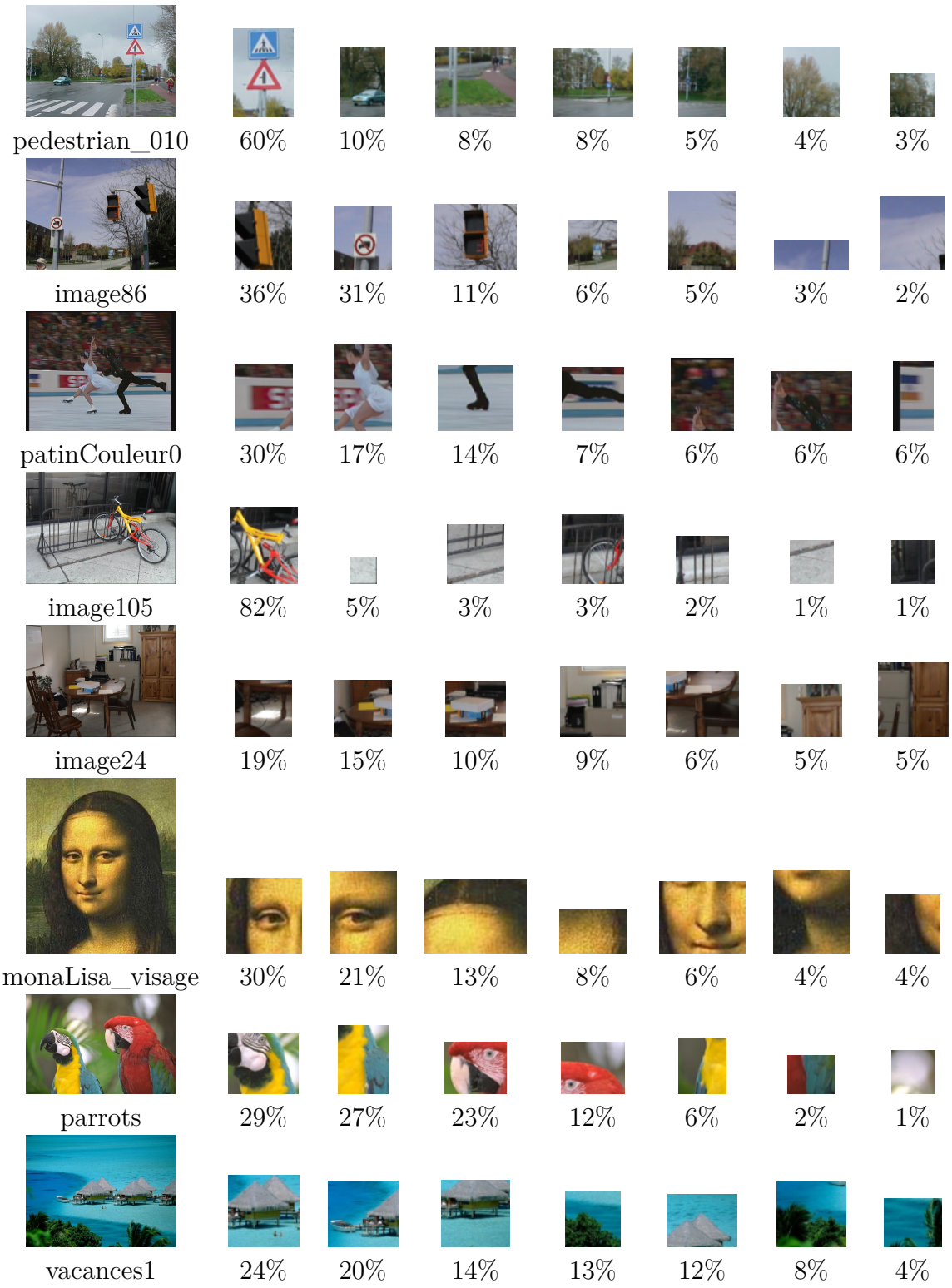


FIGURE 7.3.2: Exemple des 7 vignettes les plus saillantes pour différentes images, à t=150. Colonne de gauche : images sources. Colonnes de droite : les vignettes et leur saillance relative dans l'image.

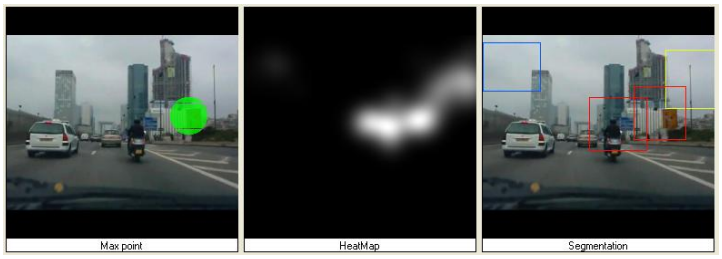
temporellement limitée).

Notre système attentionnel segmente à chaque nouvelle trame un sous-ensemble des éléments les plus importants de la scène. Ainsi, les panneaux routiers, clignotants et feux stop des véhicules, etc. sont d'une manière générale bien segmentés. Pour une bonne interprétation des résultats, il faut cependant garder à l'esprit que le système attentionnel fonctionne ici seul (non connecté à un système de vision) et sans aucun *a priori* lié au type d'élément saillant dans le contexte considéré. Ainsi, un panneau segmenté à un instant t, ne sera pas forcément suivi et segmenté tout au long de son déplacement dans la scène. Il appartiendrait au système de vision connecté au système attentionnel, d'effectuer les tâches de suivi et d'analyse de panneau. Le rôle de notre système attentionnel « se cantonne », pour chaque trame de la vidéo, à fournir une liste des régions potentiellement intéressantes ainsi qu'une évaluation de leur saillance relative. Ces valeurs de saillance relative varient en fonction des nouveaux éléments apparaissant dans la scène et de l'évolution propre du système d'attention.

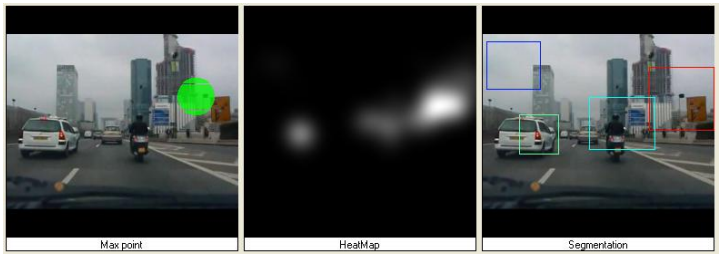
7.4 Conclusion

Nous avons proposé dans ce chapitre de transformer une série de points de focalisation en un ensemble de régions, classées selon leur saillance relative. Compte tenu des propriétés dynamiques de notre modèle d'attention, ces régions évoluent au cours du temps, permettant ainsi de proposer à un système de vision différentes possibilités d'analyse. La qualité des régions générées semble bonne, même si, en l'absence de connexion à un système de vision, une évaluation quantitative des performances est difficile. Cette connexion sort du cadre de cette thèse, elle s'inscrit très logiquement dans les perspectives de nos travaux consistant, entre autre, à valider en conditions réelles la capacité de notre système.

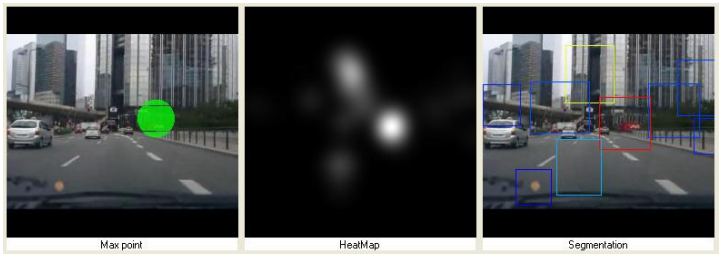
Il serait également intéressant de tester d'autres méthodes de segmentation des points de focalisation ou de la *heatmap* dynamique. Dans ce dernier cas, l'approche basée sur un partitionnement flou, proposée par Ma [Ma 03], pourrait être une voie intéressante.



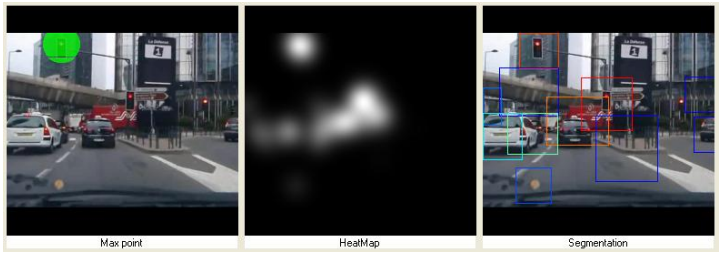
(a) Trame 311. Deux régions sont particulièrement saillantes : le panneau jaune à droite, le scooter et le panneau bleu au centre.



(b) Trame 325. La voiture de gauche a mis son clignottant en marche. Celui-ci est segmenté dans une nouvelle région.



(c) Trame 1459. Un poid-lourd rouge arrive transversalement sur le côté droit de la route.



(d) Trame 1649. La voiture arrive au croisement. Le poid-lourd est toujours segmenté, ainsi que les deux feux rouges du centre et de gauche et le feu stop de la voiture blanche à gauche.

FIGURE 7.3.3: Une sélection de trames issue des résultats de notre algorithme de segmentation attentionnelle sur une vidéo de trafic routier.

Points clés

Positionnement

- ❏ Les systèmes de vision traitent généralement des régions ou des objets.
- ❏ Il est nécessaire de mettre en place un mécanisme permettant à notre modèle d'attention visuelle spatiale de générer des régions d'intérêt.

Contributions

- ❏ Différentes régions sont générées à partir d'un *clustering* des points de focalisations autour des maximums locaux de la *heatmap*. Le K-mean est un moyen simple et efficace d'effectuer ce *clustering*.
- ❏ Lors de la segmentation de vidéos, l'algorithme doit être adapté afin de prendre en compte l'oubli progressif des focalisation les plus anciennes

Évaluation

- ❏ Dans le cas des images, notre méthode permet une segmentation itérative en différentes régions d'intérêt. La saillance relative de ces régions évolue au cours du temps, proposant ainsi différentes possibilités de segmentation. La variabilité du système reste cependant limitée et a tendance à baisser au fur et à mesure de l'évolution du système.
- ❏ Dans le cas de vidéos, le système met à jour la liste des régions saillantes et leur saillance relative à chaque nouvelle trame. Les valeurs de saillance relative évoluent en fonction des nouveaux éléments apparaissant dans la scène et de l'évolution propre du système d'attention.

Conclusion et perspectives

Rappel des objectifs

Le but de cette thèse était de proposer un modèle d'attention visuelle efficient disposant d'un ensemble de propriétés, le rendant particulièrement approprié à une application en vision adaptative. Notre cahier des charges définissait trois propriétés principales :

- fiabilité : le système devait délivrer le comportement attendu ;
- vitesse : les calculs devaient être rapides, afin de permettre au système attentionnel de jouer son rôle de « pré-traitement » dans un système de vision plus complet ;
- flexibilité : le modèle proposé devait fournir un maximum de possibilités d'adaptation, afin de pouvoir modifier son comportement en fonction du contexte changeant du système de vision adaptatif. Cette flexibilité devait également permettre de garantir que le modèle d'attention pourrait être utilisé pour une large gamme d'applications (dépassant même celle de la vision adaptative).

Bilan des travaux effectués

Pour respecter ces contraintes, nous avons mené notre étude en deux étapes. La première étape consistait en une analyse des théories et modèles d'attention existants. Nous avons tout d'abord étudié leurs applications afin d'établir un ensemble de critères (dérivés également de notre cahier des charges) permettant de caractériser les propriétés nécessaires à chaque type de tâche (segmentation, vision, CBIR, etc.). A partir de ces critères, nous avons défini une taxonomie des différents algorithmes et estimé leur adéquation avec notre application cible.

Cette analyse des différents modèles nous a permis de conclure qu'un algorithme basé sur une approche mixte hiérarchique / compétitive, permettrait d'obtenir une grande partie des avantages de ces deux familles de modèles (efficacité computationnelle et facilité d'extension pour la première, gestion dynamique de la compétition entre différentes sources d'information pour la seconde). Cette approche, déjà explorée en partie par les modèles connexionnistes centralisés n'est vraiment intéressante que si l'on s'affanchit de la carte de saillance. En appliquant la partie compétitive non plus sur cette dernière, mais sur les cartes de singularité ou de caractéristiques, on exploite au mieux le principe de compétition. L'attention est alors une propriété émergente des interactions entre les

différentes cartes.

La seconde étape de notre étude consistait à mettre en œuvre ces principes et étudier les propriétés du modèle ainsi obtenu. On peut résumer la partie *bottom-up* de son architecture comme suit :

- la partie hiérarchique est dérivée de l'architecture du modèle de Itti [Itti 98]. Celle-ci permet de calculer des cartes de caractéristiques et de singularité. Elle reprend et étend les améliorations apportées par Frintrop [Frintrop 05a] afin de rendre le modèle plus plausible sur certains points (calcul des filtres centre-périphérie, séparation des cartes on/off et off/on, etc) tout en améliorant les temps de calcul des différentes cartes. Nous proposons également un nouveau mécanisme de normalisation des cartes, basé sur la théorie de l'information, permettant de réhausser les éléments rares, sans utiliser de seuil ;
- la partie compétitive est basée sur un système dynamique proies / prédateur. Elle fournit un mécanisme efficace de compétition entre les cartes, permettant également de gérer l'évolution de la dynamique du focus d'attention.

La combinaison de ces deux sous-systèmes permet d'obtenir une gamme de comportements variés, qui peuvent être contrôlés par les paramètres du modèle. On peut ainsi ajuster simplement sa reproductibilité, sa dynamique, ou encore sa stratégie d'exploration de la scène. Cette grande flexibilité du modèle ne se fait pas au détriment de sa plausibilité, puisque les expérimentations menées montrent que celui-ci, lorsqu'il est comparé au système visuel humain, obtient des résultats comparables à des modèles reconnus (Itti, Le Meur, Bruce).

Ce système *bottom-up* n'était cependant pas aussi flexible qu'aurait pu l'être un modèle prenant en compte une influence *top-town*. Pour combler ce manque, nous avons étudié les mécanismes permettant de prendre en compte le contexte dans le système dynamique proies / prédateurs utilisé. Cette adaptation peut être réalisée de deux façons :

- assez classiquement, en utilisant des cartes *top-down* permettant de biaiser le comportement du système en fonction d'*a priori* externes ;

- en utilisant un mécanisme de rebouclage (*feedback*) permettant, par exemple, d'utiliser une carte représentant le « degré de couverture » de la scène, par les focalisations générées par le modèle pour contrôler sa stratégie d'exploration.

Ce dernier mécanisme permet également d'ajuster de manière intéressante la dynamique du système, permettant ainsi de la rapprocher du comportement attentionnel humain.

Concernant les applications liées à notre modèle adaptatif d'attention, nous avons exploré différérentes pistes, permettant :

- d'estimer la complexité des images, en utilisant non pas directement leur contenu mais la trajectoire des focalisations attentionnnelles nécessaire à leur parcours. La mesure ainsi obtenue est un estimateur de la difficulté d'aquisition de l'informa-

tion dans l'image. Celle-ci se place comme un complément aux (rares) mesures de complexité d'images classiques ;
– de recadrer dynamiquement des images et vidéos, proposant ainsi une nouvelle manière de découvrir le contenu d'une image ou d'une vidéo, en fonction de la saillance de ses différents éléments ;
– d'extraire des régions d'intérêt, ouvrant ainsi la voie vers la transformation de notre modèle d'attention spatiale, en un modèle d'attention proto-objet.

Apports, limites et perspectives

Le modèle proposé dans cette thèse répond au cahier des charges simplexe que nous nous étions fixé :
– il est fiable. Les focalisations générées correspondent au comportement attendu : comparé au système visuel humain, notre modèle est aussi plausible que d'autres modèles reconnus. De plus, la caractérisation de l'influence de ses différents paramètres permet de s'assurer que le comportement désiré pourra être obtenu ;
– il est rapide. Les mécanismes de simplification des filtres de la partie hiérarchique du modèle, l'introduction des colonnes multi-résolution, ainsi que le nombre limité de cartes utilisées par le système proies / prédateurs permettent une utilisation temps réel de notre modèle sur un ordinateur standard ;
– il est flexible. Les nombreux paramètres de notre modèle, dont nous avons caractérisé l'influence, ainsi que ses possibilités de rebouclage permettent d'adapter son comportement à des contextes variés et / ou changeants.

Notre approche est cependant limitée sur un certain nombre de points, que nous décrivons dans la suite de ce paragraphe. Ces améliorations potentielles sont autant de perspectives intéressantes pour la suite de nos travaux.

Nous avons vu au chapitre 3 que le modèle proies / prédateurs travaillait à partir des cartes de singularité produites par le système visuel hiérarchique. Cette solution est intéressante pour sa simplicité, mais limite la compétition dynamique à ces seules cartes. Une approche plus générale consisterait à utiliser les cartes de caractéristiques (voire même les différents niveaux des pyramides multi-résolutions) comme entrée du système dynamique. Il serait alors nécessaire de déterminer une hiérarchie à différents niveaux dans les proies et les prédateurs (figure 7.4.1). Cependant, la difficulité d'une telle approche serait liée à la gestion du temps de propagation de l'information dans cette hierarchie afin de garder le système suffisamment réactif.

Comme nous l'avons vu en sous-section 2.3.2.1, il semblerait que dans le système visuel humain l'attention dynamique soit prioritaire sur l'attention statique. Cependant, notre modèle ne traite pas la priorité entre la voie statique de traitement de l'information visuelle (intensité, couleur, orientation) et son pendant dynamique (mouvement).

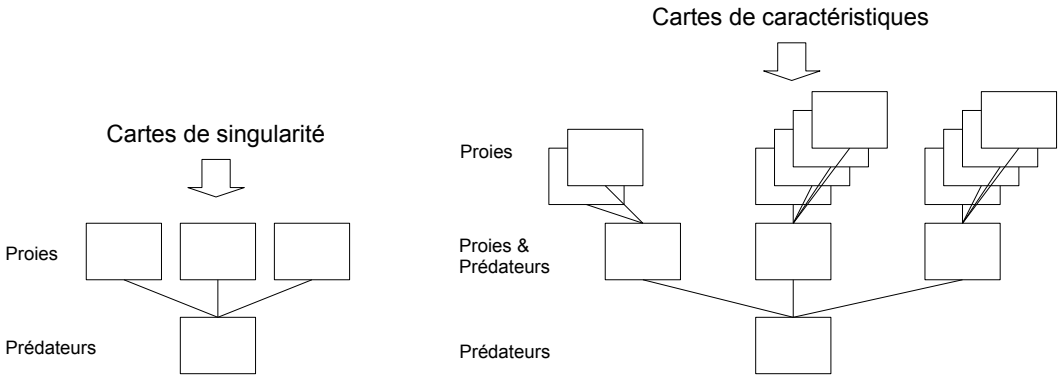


FIGURE 7.4.1: Prise en compte des cartes de caractéristiques dans le modèle proies / prédateurs. A gauche : la hiérarchie utilisée actuellement par notre modèle ; les proies sont associées aux cartes de singularié. A droite : une hiérarchie basée sur les cartes de caractéristiques ; on ajoute un étage au système proies / prédateurs, les cartes de singularité sont remplacées par les proies du niveau central (également prédatrices des proies du niveau supérieur).

La question de la manière d'incorporer cette propriété dans notre modèle est particulièrement intéressante, mais non triviale : il faudrait en particulier étudier l'impact de l'inhibition de la partie statique du modèle sur le comportement du système proies / prédateurs.

Enfin, une perspective plus générale concerne la réalisation d'un système de vision adaptatif exploitant notre modèle. On pourrait ainsi bénéficier pleinement de ses capacités d'adaptation et mesurer ses performances en conditions d'utilisation réelles. Cela permettrait également de tester d'autres critères, liés au système de vision, pour l'application du mécanisme de *feedback* présenté au chapitre 4.

La figure 7.4.2 donne un aperçu de ce que pourrait être ce système de vision hôte. Le schéma reprend celui présenté en section 1.3 lors de la définition de notre approche, et le complète par quelques indications sur la manière de mettre en œuvre la partie « système de suivi / reconnaissance ». La mise en place d'un tel système représente à elle seule le travail d'une nouvelle thèse, mais on peut déjà imaginer les différents modules à interconnecter. L'étude du module de reconnaissance, en particulier, soulève de nombreuse questions intéressantes :

– comment représenter l'information visuelle de manière compatible avec notre modèle d'attention [1] ?

1. On pourra par exemple s'inspirer des idées de [Rensink 00]

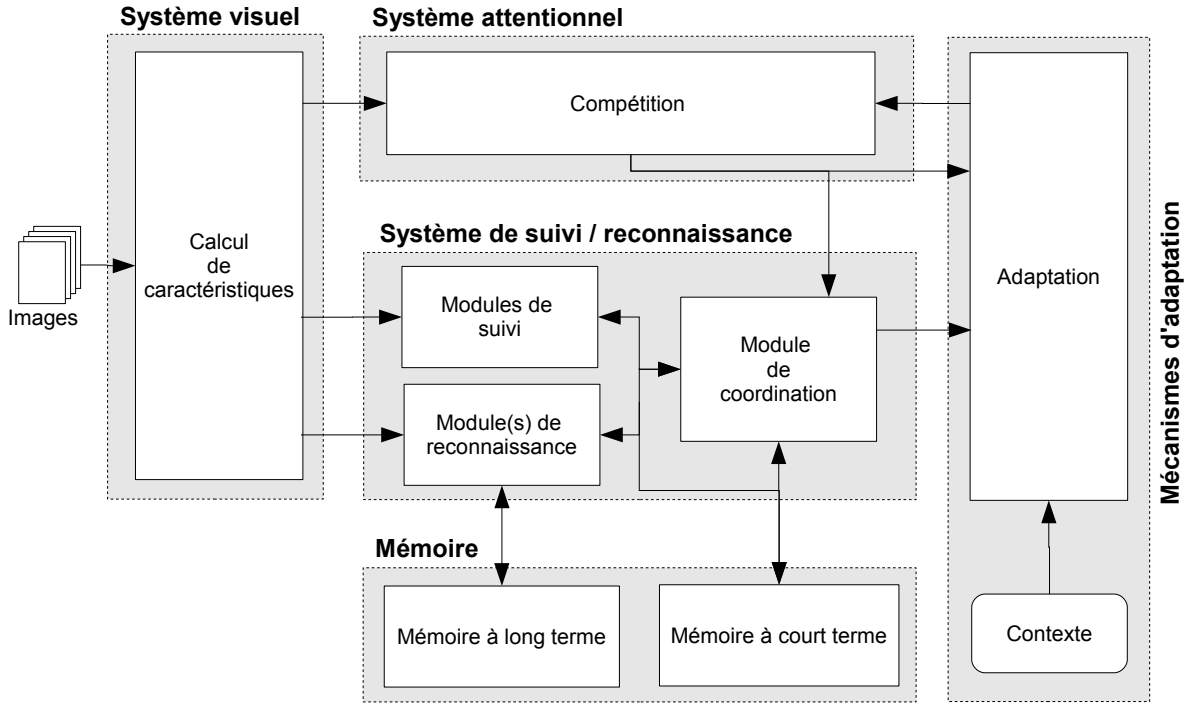


FIGURE 7.4.2: Exemple d'une architecture de vision attentionnelle adaptative.

– peut-on réutiliser efficacement les caractéristiques générées par le système visuel [2] ?
– comment relier attention, reconnaissance et mémorisation ?

Pour répondre à ces questions, on pourra s'inspirer, comme nous l'avons fait pour ces travaux de thèse, de solutions mises en place par la nature pour résoudre des problèmes similaires. Nous plaidons ici pour une approche transdisciplinaire, permettant d'exploiter des modèles issus par exemple de la biologie, l'économie ou la sociologie, qui répondrait à un champ de questions trop vaste pour une résolution purement informatique.

2. Une partie de la réponse à cette question est fournie par [Siagian 07].

Annexes

Annexe A

Etudes sur l'attention auditive

Les premières études de l'attention ne concernaient pas la modalité visuelle (plus difficile à étudier expérimentalement) mais auditive. Le principe expérimental utilisé était celui de l'écoute dichotique. Ce procédé consiste à faire écouter aux sujets, à l'aide d'un casque stéréo, des messages différents dans l'oreille droite et gauche. Ces études ont prouvé qu'il était presque impossible de mémoriser simultanément deux sources d'information auditives. Ainsi, lorsque l'on demande à un sujet de restituer le message A qu'il entend dans son oreille gauche (attentive), il lui est très difficile de restituer des éléments du message B entendu dans son oreille droite (non attentive). Seuls certains mots comme le nom du sujet, ou des informations de (relativement) bas niveau peuvent être restituées : volume, ton de la voix, etc.

L'attention sélective précoce

L'analyse de ces premières études par Broadbent [Broadbent 58] lui permit de créer un des tout premiers modèles d'attention. Ce modèle utilise l'hypothèse de l'attention comme processus de filtrage (2.1), permettant de ne pas surcharger notre cerveau aux capacités limitées. C'est un modèle de sélection précoce de l'information. Celui-ci considère que, après une phase d'extraction de ses propriétés physiques (parallèle et bas niveau), l'information auditive est filtrée afin que seuls les types de *stimuli* attendus puissent être analysés dans une phase de traitement de plus haut niveau (figure A.1).

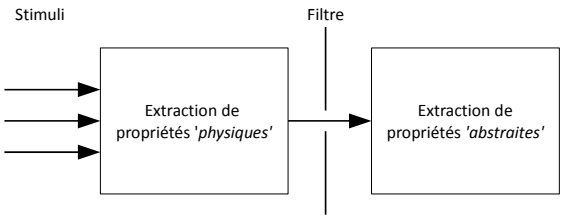


FIGURE A.1: Le modèle de l'attention sélective précoce [Broadbent 58].

Ce modèle expliquait très bien que les sujets en écoute dichotique ne puissent percevoir que les propriétés physiques simples des *stimuli* présentés dans leur oreille non attentive. Il échouait par contre à expliquer pourquoi les sujets pouvaient reconnaître leur nom dans cette même oreille non attentive.

L'attention sélective tardive

Pour combler les lacunes du modèle de Broadbent, un modèle concurrent, à l'approche radicalement opposée, fut proposé par Deutsch [Deutsch 63]. Ce modèle ne considère plus l'attention comme un filtre précoce permettant de limiter la quantité d'information à traiter de manière approfondie. Il considère au contraire que toutes les propriétés des *stimuli* perçus sont extraites en parallèle et que l'attention n'est que la conséquence des limitations de stockage de notre mémoire de travail (figure A.2). C'est un modèle d'attention sélective tardive.

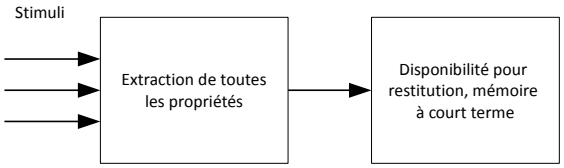


FIGURE A.2: Le modèle de l'attention sélective tardive [Deutsch 63].

L'atténuation sélective

A la frontière entre ces deux modèles extrêmes, Treisman [Treisman 60, Treisman 69] propose un modèle plus mesuré : l'atténuation sélective. Celui-ci reprend en grande partie l'esprit de la sélection précoce de Broadbent, mais Treisman considère que plutôt que de supprimer totalement les *stimuli* non-attendus, l'attention ne fait que les atténuer. Ainsi toute l'information parvient à l'étape de traitement de plus haut niveau, mais les *stimuli* atténués ont généralement un niveau trop faible pour être traités. Par contre, lorsque ceux-ci représentent une information importante dans le contexte ou pour le sujet (son nom par exemple), ils peuvent tout de même être traités et atteindre la conscience.

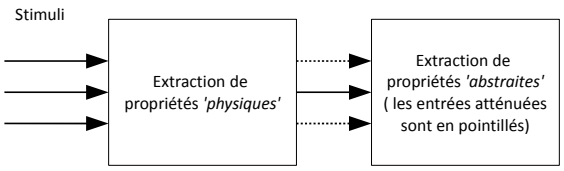


FIGURE A.3: Le modèle de l'atténuation sélective [Treisman 60, Treisman 69].

L'étude de l'attention auditive sera peu à peu délaissée au profit de l'attention visuelle. Les principaux modèles théoriques développés dans les années 70 à 90 concerneront quasi uniquement cette modalité.

Annexe B

Le système visuel humain

Le système visuel humain (figure B.1) est composé des éléments suivants (dans l'ordre de cheminement de l'information visuelle) :

- l'œil, lui même composé d'une partie optique (cornée, iris, cristallin) et d'un capteur (la rétine). Son fonctionnement, et en particulier celui de la rétine, est décrit dans le prochain paragraphe ;
- le nerf optique, comprenant 1.5 millions de fibres (axones) par œil. Compte tenu de la quantité d'information à transmettre, c'est l'axone le plus rapide du corps humain (60 mètres par seconde) ;
- le chiasma optique, qui est l'endroit où se croisent les nerfs optiques provenant des deux yeux. La conséquence de ce croisement est que la partie droite du cortex visuel traite la partie gauche du champ visuel et réciproquement. Notons que la ségrégation entre les deux champs n'est pas totale : une petite région au centre du champ visuel est transmise conjointement aux deux hémisphères ;
- le tractus optique, qui effectue la liaison entre le chiasma optique et le corps genouillé latéral ;
- le corps genouillé latéral, servant de relais sensoriel. Il est situé dans le thalamus et reçoit 90% des axones provenant du nerf optique (ce cheminement représente le système géniculostrié). Les 10% restant sont reliés au colliculus supérieur (pour former le système tectopulvinarien) ;
- les radiations optiques, relient le corps genouillé latéral au cortex visuel primaire (V1). A partir de cette jonction avec V1, le cheminement de l'information visuelle se complexifie : les interconnexions et réentrances sont plus nombreuses ;
- le cortex visuel primaire (V1) puis secondaire (V2, V3, V3A, V4, V5), décompose l'image en caractéristiques de plus en plus complexes. Son fonctionnement est décrit dans la sous-section B.

Les deux principaux lieux où l'information visuelle est « transformée » sont la rétine et le cortex visuel. Les deux prochains paragraphes abordent la façon dont l'information

y est modifiée, dans un but de simplification et d'extraction de caractéristiques de plus haut niveau.

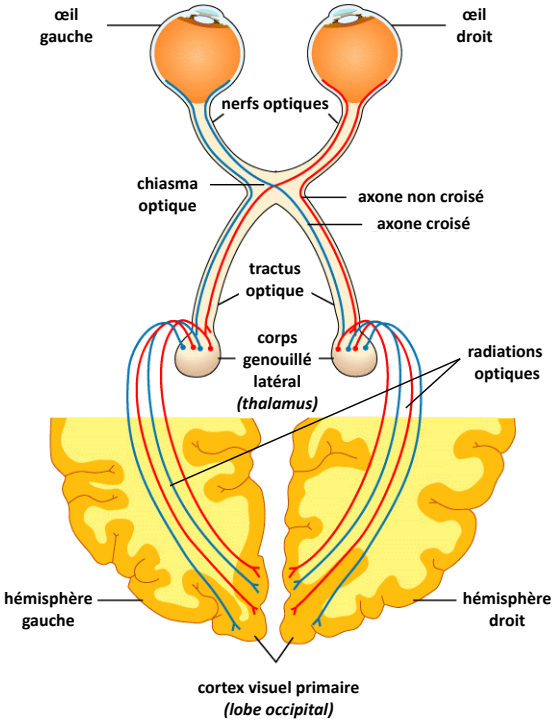


FIGURE B.1: Les voies visuelles : de la rétine au cortex visuel primaire (V1).

L'œil et la rétine

L'œil (figure B.2) est un dispositif optique assez classique. Les rayons lumineux traversent d'abord la cornée qui agit comme une première lentille, permettant aux rayons lumineux de se concentrer afin de traverser la pupille, dont le diamètre dépend de l'ouverture de l'iris. L'iris permet de réguler la quantité de lumière qui traversera le cristallin pour être projetée, de manière inversée, sur la rétine. Au centre de la rétine, dans le prolongement de l'axe optique, on trouve la fovéa. C'est dans cette zone, où l'on retrouve la plus grande quantité de photorécepteurs, que notre acuité visuelle est la plus grande.

La rétine représente la partie « capteur » de l'œil. Elle est composée de deux principaux types de cellules :

– les photorécepteurs : cônes et bâtonnets. Leur répartition n'est pas uniforme, plus on s'éloigne de la fovéa et moins les récepteurs sont nombreux. Les cônes, principalement utilisés pour la vision diurne, sont concentrés au niveau de la fovéa alors que les bâtonnets, principalement utilisés pour la vision nocturne, sont absents de

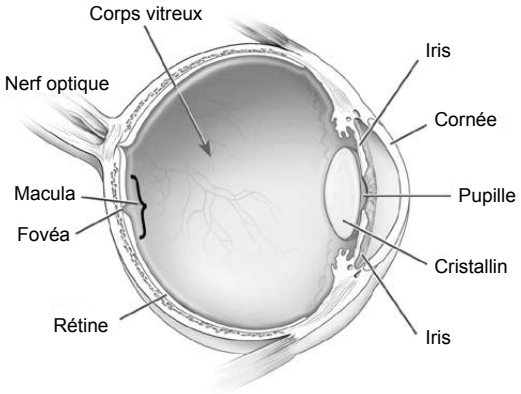


FIGURE B.2: Anatomie de l'oeil.

cette zone. Les principales caractéristiques de ces deux types des photorécepteurs sont résumées dans le tableau B.1 ;

Type	Localisation	Résolution spatiale	Réponse	Sensibilité	Couleur	Nombre
Cônes	dans la fovéa	élevée	rapide	basse	oui (R,V, B)	6 millions 5%
Bâtonnets	sauf la fovéa	basse	lent	élevée	non	125 milions 95%

TABLE B.1: Principales caractéristiques des cellules photoréceptrices de la rétine.

– les cellules transmission et transformation de l'information. Le signal provenant des cônes et bâtonnets est transformé par les cellules horizontales, amacrines et bipolaires avant de traverser différents types de cellules ganglionnaires. Le tableau B.2 présente une synthèse de leurs propriétés. Notons que les cellules koniocellulaires ont été assez peu étudiées jusqu'à présent ; on connaît donc assez mal leur rôle.

Les photorécepteurs ne sont pas situés directement à la surface de la rétine. La lumière doit d'abord traverser les cellules horizontales, amacrines, etc. avant d'atteindre les cônes et bâtonnets. Les cellules ganglionnaires ainsi que les fibres du nerf optique se trouvent sur la couche le plus extérieure de la rétine. Ainsi, la lumière ne peut pas traverser la forte densité de fibres présentes à l'endroit où le nerf optique est connecté à la rétine : c'est la tâche aveugle. Dans cette zone, l'œil ne capture aucune information visuelle, notre cerveau y « interpole » les données manquantes afin que notre perception semble complète et continue.

La répartition et les caractéristiques des cônes et bâtonnets ont des conséquences importantes sur le fonctionnement de notre vision :

Type	Champ récepteur	Type	Réponse	Caractéristiques	Source	Nombre
M magnocellulaire	grand	centre - périphérie	rapide	profondeur, mouvement	batonnet	5%
P parvocellulaire	petit	centre - périphérie	lent	couleur, forme et détails	cones (R,V)	95%
K koniocellulaire	très grand	centre	moyen	couleur	cones (B)	5%

TABLE B.2: Propriétés des cellules ganglionnaires.

- en condition diurne, notre vision n'est précise, et en couleur que sur quelques degrés de notre champ de vision ;
- notre vision périphérique, achromatique, est principalement utilisée lors de conditions de lumière faible ;
- « l'image » que nous percevons est de résolution variable : précise au centre, approximative en périphérie ;
- de par leur sensibilité importante, les bâtonnets permettent une grande aptitude à réagir aux mouvements en vision périphérique. Cependant, leur temps d'intégration étant assez considérable (environ 100 millisecondes), cette capture du mouvement n'est pas très précise (ni spatialement, ni temporellement).

Ces propriétés rendent le mouvement des yeux incontournable pour percevoir correctement l'ensemble de notre environnement. Il est également probable qu'elles conditionnement en partie le fonctionnement de notre système d'attention visuelle.

Les différentes interconnexions entre les cellules horizontales, amacrines, bipolaires et ganglionnaires, et les photorécepteurs, forment des champs récepteurs. La plupart de ceux-ci sont de type centre-périphérie (*center-surround*), et peuvent être :

- on center - off surround (figure B.3, à gauche), ayant une réponse maximale en présence d'une tache blanche entournée de noir ;
- off center - on surround (figure B.3, à droite), ayant une réponse maximale en présence d'une tache noire entourée de blanc.

Le rôle de ces champs récepteurs est de transformer le signal issu des photorécepteurs, codant l'intensité de la lumière, en un signal codant le contraste. On peut ainsi comparer le traitement effectué par la rétine à un algorithme de détection de bords (généralement modélisé informatiquement par une différence de gaussiennes). Ce type de représentation de l'information correspond à un type particulier de codage différentiel : le codage prédictif. C'est une méthode utilisée en théorie de l'information pour réduire la quantité de données à transférer [Sayood 00]. En se basant sur l'hypothèse que le signal à transmettre est majoritairement uniforme et / ou continu, la différence entre la donnée

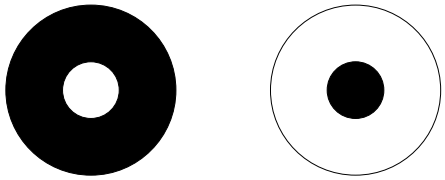


FIGURE B.3: Champs récepteurs centre-périphérie. A gauche : on center - off surround. A droite : off center - on surround.

à coder et la moyenne des valeurs de son environnement proche, sera généralement un nombre plus petit que la donnée elle-même. Elle nécessitera alors moins de quantité d'information (en informatique, de bits) pour être représentée.

Ce type de codage est nécessaire puisqu'il y a un rapport de 100 contre 1 entre le nombre de photorécepteurs dans la rétine (130 millions) et le nombre de fibres du nerf optique (1.2 millions).

Le cortex visuel

La figure B.4 donne un aperçu de la principale voie de traitement de l'information visuelle (90% des données), passant par le système géniculostrié. Les 10% restants passent par le colliculus supérieur et le pluvinar avant de rejoindre les aires V2 et V3 du cortex visuel. C'est le système tectopulvinarien, principalement utilisé pour le pilotage du mouvement des yeux.

Le système géniculostrié étant la source principale de notre vision consciente, c'est lui que nous allons décrire dans la suite de ce paragraphe.

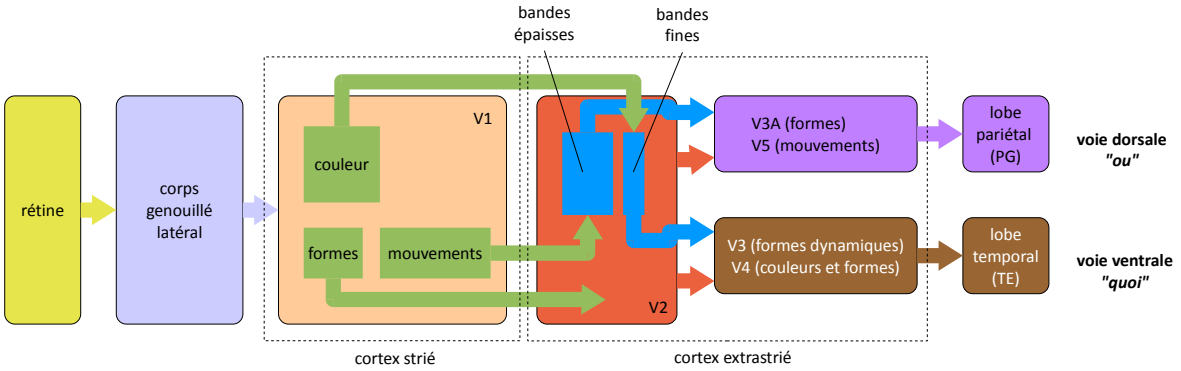


FIGURE B.4: Cheminement de l'information visuelle à travers les différentes aires du cortex visuel (ici on n'a représenté que le système géniculostrié).

Après avoir traversé le corps genouillé latéral où l'information visuelle est traitée selon

deux voies séparées (magnocellulaire (M) et parvocellulaire (P)), les signaux sont traités par le cortex visuel primaire V1 (cortex strié). Dans V1, les neurones sont organisés de manière rétinotopique : il y a une étroite correspondance entre la localisation d'une zone dans V1 et sa position dans le champ visuel. Le principal rôle de V1 est d'effectuer un filtrage spatio-temporel des données qu'il reçoit. On peut comparer cette fonction avec l'application de filtres de Gabor sur une image numérique. L'application de ce type de filtres permet d'extraire des informations de fréquence (spatiale), orientation, mouvement, direction et vitesse (fréquence temporelle). L'autre rôle de V1 est de traiter l'information couleur ; les cellules de cette zone effectuent un calcul de ratio entre les 3 couleurs de base (rouge, vert et bleu) en utilisant le mécanisme centre-périphérie décrit plus haut (avec par exemple un centre sensible au rouge et une périphérie sensible au vert). A la sortie de V1, les signaux concernant la forme, la couleur et le mouvement sont séparés.

Les cellules de V2 réagissent également à des propriétés simples telles que l'orientation, la fréquence spatiale ou la couleur. Cependant, des traitements d'un plus haut niveau semblent également être réalisés, notamment concernant la continuation des contours et d'autres propriétés de groupement perceptuel [Grossberg 97] tels qu'ont pu les définir les partisans de la théorie de la forme (cf. annexe C). Il semblerait également que la réponse des cellules de V2 puisse être légèrement modulée par les mécanismes attentionnels [Luck 97].

Après V2, une hypothèse largement admise considère que le traitement de l'information visuelle est effectué selon deux voies :

- la voie dorsale, dont le rôle serait de nous aider à diriger nos actions et à localiser les objets dans l'espace. Elle est également connue sous le nom de voie pariétale ou encore voie « où » ou « comment ». Son fonctionnement est lié aux aires V2a, V5 et au lobe pariétal ;
- la voie ventrale, dont le rôle serait la reconnaissance des objets et la représentation des formes. Elle est également connue sous le nom de voie temporale ou encore voie « quoi ». Son fonctionnement est lié aux aires V3, V4 et au lobe temporal.

Cette hypothèse est cependant controversée. Il semblerait qu'elle résulte d'une trop grande simplification car ces deux voies sont en réalité fortement reliées entre elles, leur indépendance fonctionnelle est alors difficilement démontrable.

L'aire V3 est sous divisée en plusieurs parties : l'aire V3a est considérée comme faisant partie de la voie dorsale alors que V3 serait liée à la voie ventrale. Son rôle n'est pas encore établi avec certitude, mais il semblerait que V3 joue un rôle dans la perception du mouvement et en particulier dans l'estimation du mouvement global de la scène et d'objets.

Comme V1 et V2, V4 réagit à l'orientation, la fréquence spatiale et la couleur, mais les

caractéristiques manipulées seraient d'un plus haut niveau de complexité. De manière encore plus importante que V2, la réponse des cellules de V4 peut être modulée par l'attention.

V5 aussi connue sous le nom d'aire MT est liée à la perception du mouvement (à un plus haut niveau que V3) et au contrôle des mouvements oculaires.

Plus loin dans la voie ventrale, lorsque les signaux atteignent le lobe temporal, les cellules réagissent à des caractéristiques de bien plus haut niveau (visages, objets complexes). A ce stade, les champs récepteurs des cellules sont tellement larges que toute information spatiale est perdue. Se pose alors le problème de la correspondance (*binding*) entre l'objet perçu (traité par la voie ventrale) et sa localisation dans le champ visuel (traitée par la voie dorsale). Certaines théories attentionnelles [Treisman 80]suggèrent que l'attention a un rôle clé dans la résolution de ce problème.

Annexe C

La théorie de la forme

La théorie de la forme (*Gestalttheorie*), également appelée psychologie de la forme, est un courant de pensée d'origine allemande qui s'est développé à partir de la fin du XIXème, et surtout dans la première moitié du XXème siècle. Ses principaux contributeurs étaient Max Wertheimer, Wolfgang Köhler, Kurt Koffka et Kurt Lewin. En France, c'est surtout Paul Guillaume [Guillaume 37] qui a participé à sa diffusion.

FIGURE C.1: Exemple classique illustrant notre perception globale des formes. Nous percevons correctement le dalmatien, bien qu'il ne soit constitué que d'un ensemble de tâches.

D'après cette théorie, nous percevons les objets comme un tout (une forme), et non comme la somme d'un ensemble de parties (figure C.1). La perception consiste alors en une « segmentation » fond / forme (objets).

La structuration des formes amenant à leur perception est réalisée selon un ensemble de règles (figure C.2) :

– continuité : les éléments d'une forme tendent à être perçus comme un tout lorsqu'il sont dans le prolongement les uns des autres ;

- proximité : les élements proches les uns des autres ont tendance à être regroupés ensemble en une forme unique ;
- similitude : nous regroupons ensemble les éléments similaires ;
- destin commun : des éléments se déplaçant avec la même trajectoire sont perçus comme faisant partie de la même forme ;
- fermeture : notre esprit peut regrouper des élements ensemble, s'ils contribuent à construire une forme fermée ;
- symétrie : des éléments formant une forme symétrique sont perçus comme un tout, indépendamment de leur distance.

L'ensemble de ces règles contribue à la perception de « bonnes formes », qui correspondent, selon le contexte, à des formes : géométriques, attendues, symétriques, familières, etc.

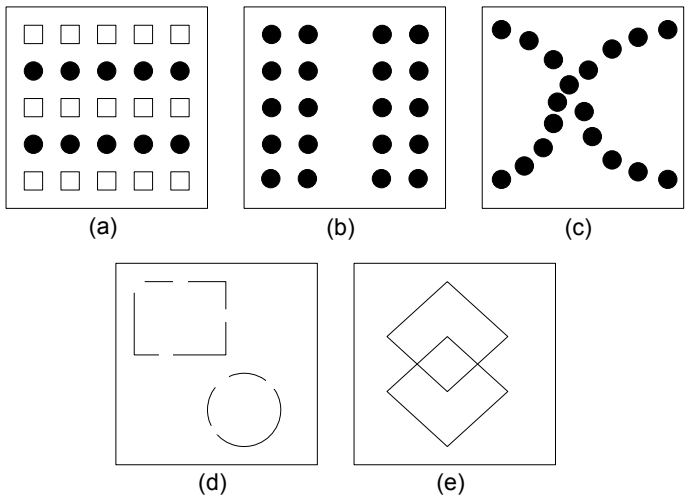


FIGURE C.2: Illustration des différentes lois de la Gestalt. (a) similarité : nous percevons des lignes (ou barres) de ronds noirs et carrés blancs ; (b) proximité : nous percevons deux blocs verticaux ; (c) continuité : la figure semble représenter deux courbes qui se croisent ; (d) fermeture : les différents segments et arcs de cercle sont perçus comme un rectangle et un cercle ; (e) symétrie : nous percevons deux losanges superposés plutôt que trois polygones.

Malheureusement cette théorie est principalement descriptive, et peine à mettre à jour les mécanismes permettant la mise en oeuvre de ces règles. De plus, la mise en application des différentes règles est problématique, car celles-ci entrent régulièrement en conflit. Des études théoriques et expériementales continuent cependant à être menées afin d'affiner ces règles, établir une priorité entre elles (pour gérer les conflits) et essayer de révéler les processus cognitifs sous-jacents [Palmer 99, Rosenthal 99, Sasaki 07]. Certains

de ces travaux théoriques effectuent un lien entre attention et groupement perceptuel [Roelfsema 06] : celui-ci serait réalisé en partie grâce à la modulation attentionnelle de certaines caractéristiques.

D'un point de vue computationnel, différents travaux ont tenté de mettre en application les lois de la Geslstat dans des applications variées : analyse de mouvement[Sabatini 03], traitement d'image [Wörgötter 04], vision par ordinateur [Desolneux 04], ou encore analyse d'images [Desolneux 08]. D'un point de vue plus général, [Petitot 08] propose une formulation neurogéométrique de certaines lois Gestaltistes, calculable à très bas niveau.

Enfin, Zou [Zou 05] utilise un système attentionnel afin de guider un algorithme de groupement perceptuel ; le modèle ainsi obtenu se rapproche alors d'un modèle d'attention objet.

Annexe D

Implémentation

Le modèle d'attention que nous avons présenté dans cette thèse est un modèle computationnel, il est donc associé à une implémentation informatique. Celle-ci permet d'observer son comportement et de vérifier ses propriétés. Ceci est effectué grâce à un démonstrateur logiciel, proposant une interface graphique permettant de soumettre des images ou vidéos à notre modèle attentionnel, modifier ses paramètres et bien sûr, observer les différentes cartes générées. Celui-ci sera décrit en section D.1.

D'autres outils, permettant de générer les données puis d'effectuer les mesures nécessaires à l'étude des propriétés de notre modèle, ont également été créés. Cependant, ceux-ci réutilisent les librairies créées pour le démonstrateur. L'architecture de ces librairies, ainsi que différents détails d'implémentation importants seront décrits en section D.2.

Enfin, le cahier des charges défini au chapitre 1, insistait sur la nécessaire rapidité du système. Pour pouvoir être utilisé conjointement avec un système de vision, notre système attentionnel se devrait d'être le plus proche possible du temps réel. Nous décrivons les performances liées à la vitesse de traitement de notre implémentation en section D.3.

D.1 Démonstrateur

L'objectif de ce démonstrateur était double :
- permettre la validation du modèle en cours de développement. Les possibilités de modification interactive des paramètres, ainsi que de visualisation de leurs conséquences, permettent l'étude empirique du comportement du modèle. Bien que peu valable scientifiquement, celle-ci permet de vérifier rapidement si les idées intégrées dans le modèle vont dans la bonne direction.
- bénéficier d'un outil de présentation de nos travaux, permettant de communiquer facilement nos résultats à un large public.

Le démonstrateur implémente la totalité des contributions présentées dans les parties 2.4 et 4.4 de ce rapport. Ceci inclus le modèle attentionnel, ainsi que les applications associées. Il n'est cependant pas possible de classer directement un jeu d'images en fonction de leurs complexité à partir du démonstrateur. Ce classement nécessite l'utilisation d'un outil en ligne de commande, plus adapté à la tâche.

Fonctionnalités

La figure D.1.1, donne un aperçu de l'interface graphique du démonstrateur. Celle-ci est assez riche car la quasi totalité des paramètres et fonctionnalités du modèle y sont contrôlables.

Le système peut traiter différents type de données :
- une image ;
- une séquence d'images ;
- un fichier vidéo ;
- un flux vidéo issu d'une webcam.

A partir de ces données d'entrée on peut calculer soit :
- un ensemble de cartes de caractéristiques et de singularité. Seule la partie hiérarchique de notre modèle est utilisée (pas de système proies / prédateurs). Les cartes générées peuvent alors être utilisées dans un autre modèle d'attention (pour comparer par exemple la qualité des cartes).
- l'évolution du système proies / prédateurs, à partir de cartes de singularité fournies par un modèle d'attention extérieure. On peut alors observer l'apport de notre système proies / prédateurs sur les cartes d'un modèle de l'état de l'art (c'est ce qui a été réalisé section 3.2.2 avec les cartes de singularité issues du modèle de Laurent Itti).
- l'évolution complète du système, à partir d'une image ou une vidéo. C'est le mode de fonctionnement le plus courant, utilisant l'ensemble de notre modèle.

Quel que soit le mode de fonctionnement choisi, on peut modifier les différents paramètres du modèle concernant la partie *bottom-up* (génération des cartes de caractéristiques et de singularité, modèle proies / prédateurs, calcul des focalisations, génération de la *heat-map*), *top-down* (cartes *top-down*, *feedback*, facteur d'oubli), ou les applications (inertie du recadrage dynamique).

Il est également possible de choisir quelles cartes seront affichées (caractéristiques, singularité, etc.). L'activation des différentes applications de notre modèle d'attention (recadrage, segmentation) est également effectuée par ce biais.

Enfin, il est possible d'afficher différents graphiques, permettant de représenter :
- diverses informations liées aux focalisations générées par le modèle : trajectoire des différentes focalisations générées, *clustering* des différentes focalisations autour des maximums locaux de la *heatmap*, distance entre les différentes focalisations en fonction du temps, entropie spectrale de la longueur des saccades en fonction du temps, taux de compression *deflate* de la trajectoire des focalisations en fonction du temps, et plans de phase de la trajectoire ;
- l'évolution en fonction du temps de différentes grandeurs calculées sur les cartes des proies et des prédateurs : minimum, maximum, moyenne, entropie, et plans de phase des 4 mesures précitées ;
- l'évolution des mesures liées à l'exploration de l'espace de notre modèle : moyenne des valeurs absolues des différences, taux de compression PNG, taux de compression JPEG.

Certaines des mesures utilisées aux chapitres 3 et 4 ne sont pas représentées graphiquement. Elle peuvent cependant être générées par des outils de mesure en ligne de commande.

D.2 Architecture et développement

Le modèle d'attention, son démonstrateur, et l'ensemble des outils permettant de mesurer son comportement, ont été développés en C#, dans une architecture modulaire, permettant une réutilisation aisée de ses différents composants. Nous ne décrirons ici que l'architecture du système attentionnel, les autres outils (démonstrateurs, etc.) n'étant que des utilisateurs de celle-ci.

Le modèle est organisé en 3 principaux modules (figure D.2.1) :
- *VisualSystem*, représente l'ensemble du modèle, composé d'une partie hérarchique calculant les cartes de caractéristiques et singularité (*ConspicuityMaps*) et d'un système dynamique, calculant l'évolution du système proies / prédateurs (*AttentionalSystem*).
- *ConspicuityMaps,* calcule les différentes cartes *via* les mécanismes de pyramides ou colonnes center-surround (*CenterSurroundPyr* et *CenterSurroundCol*).
- *AttentionnalSystem* calcule l'évolution du système dynamique grâce à *DataMapsSolver*.

VisualSystem fait également appel à *BasicTopDownSystem* pour gérer l'adaptation *via* les cartes *top-down* ou les mécanismes de rebouclage (*feedback*). Lors d'une utilisation en mode purement *bottom-up*, *BasicTopDownSystem* est remplacé par *NullTopDownSystem*.

L'ensemble des manipulations concernant les images (chargement, traitement, affichage, etc.) est réalisé grâce à la librairie *SharpVision*, que nous avons développée tout

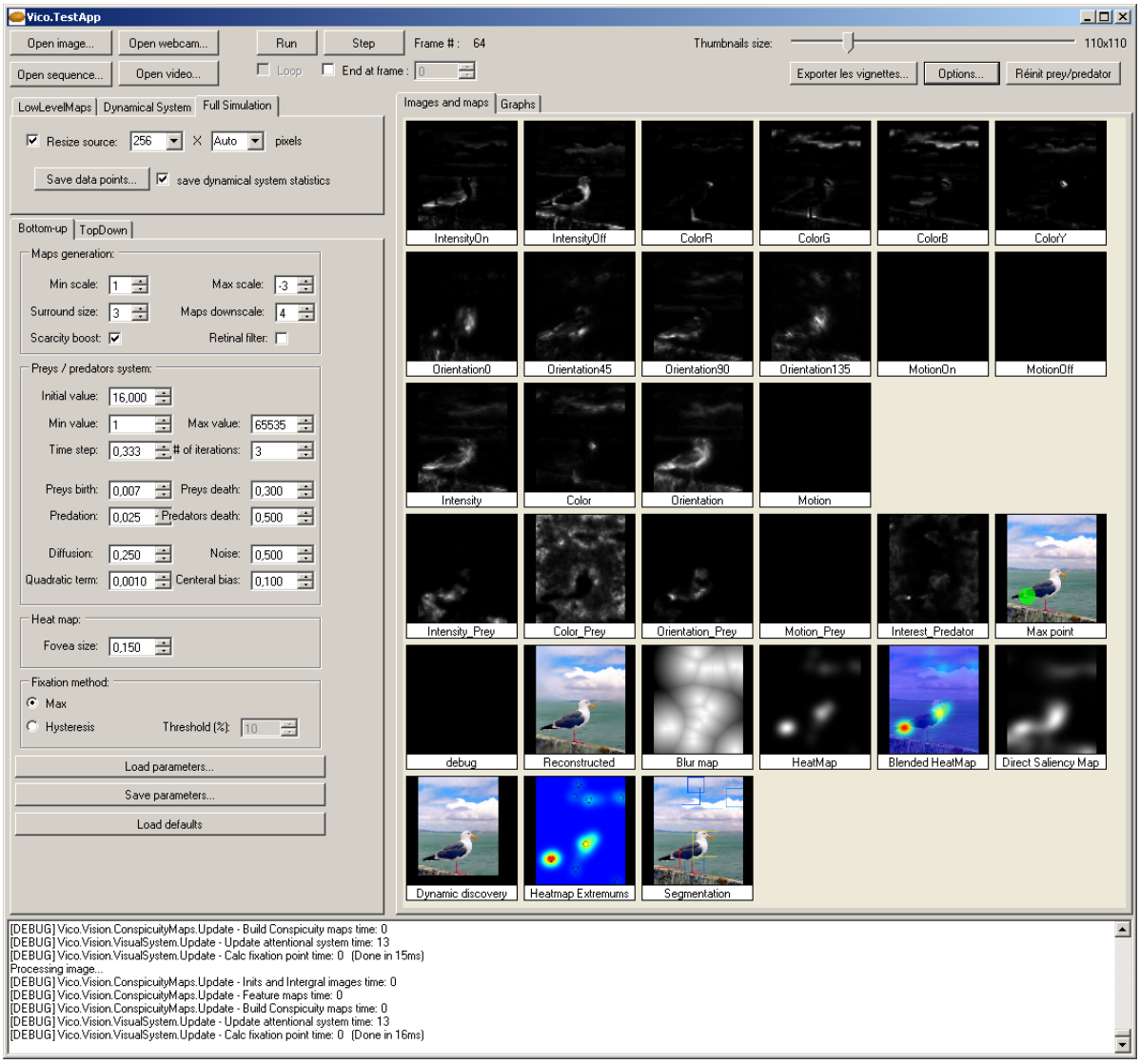


FIGURE D.1.1: Copie d'écran du démonstrateur de notre modèle d'attention. A gauche, les différents modes de fonctionnement et paramètres. A droite, la zone d'affichage des différentes cartes générées par le système. En bas, les différents temps de calcul.

au long de cette thèse. Celle-ci fournit un ensemble de fonctionnalités (tableau D.1), pour la plupart parallélisées, permettant une mise en œuvre efficace (facilité de programmation et performance d'exécution) de la majorité des opérations basiques de traitement d'images.

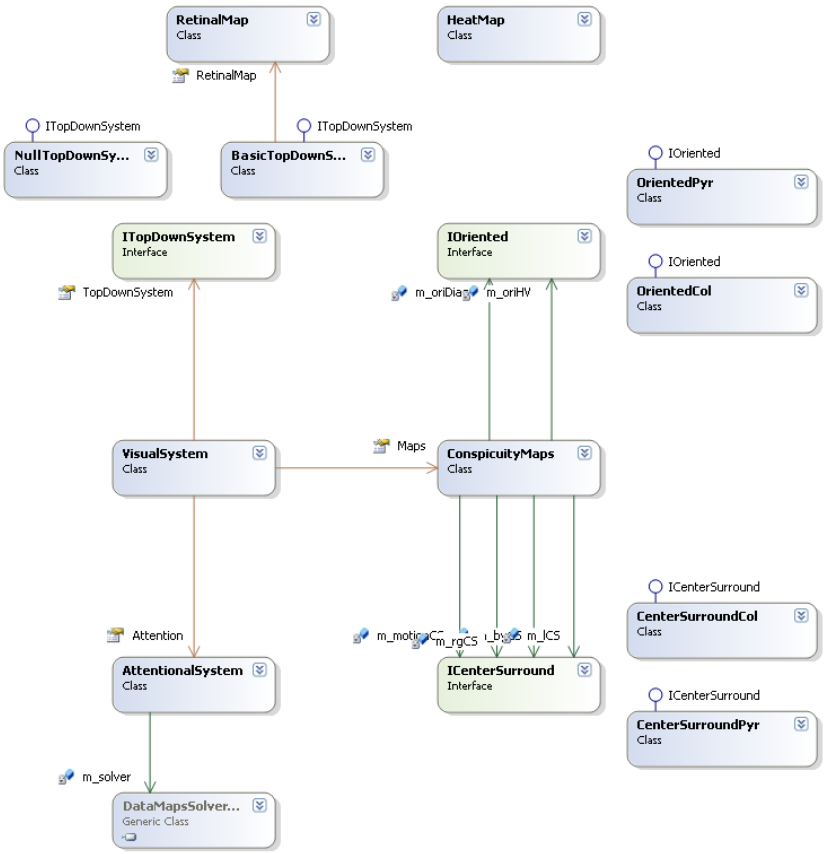


FIGURE D.2.1: Organisation des différentes classes du modèle attentionnel.

D.3 Performances

Dans cette section, nous étudions les performances de notre système proies / prédateurs selon différents angles. En sous-section D.3.1, nous justifions l'intérêt d'utiliser les images intégrales pour accélérer le calcul des cartes de caractéristiques et singularité. En sous-section D.3.2, nous évoquons la vitesse de traitement de notre modèle pour différentes configurations matérielles et pour différentes tailles d'image. Enfin, en sous-section D.3.3, nous soulignons le rôle de la parallélisation des calculs dans la performance computationnelle de notre modèle.

Fonctionnalité	Détails
Acquisition	flux brut ou MJPEG issus de webcam / caméra IP séquences d'images fichiers vidéo
Image	chargement, sauvegarde, affichage opérations arithmétiques (+,-,x,/) opérations mathématiques (min, sqrt, sin, etc.) gestion des bords (convolution, etc.) interpolation (plus proche voisin, bilinéaire, bicubique)
Image intégrale	somme et moyenne de zones rectangulaires
Pyramide multi-résolution	gaussienne, laplacienne, moyenne somme trans-échelles
Transformée de Fourier	directe et inverse magnitude, densité spectrale
Transformée en ondelettes	décomposition standard et non standard par *liftting* Harr, Daubechies4, CDF9/7
Dessin	lignes, rectangles, cercles
Filtres	flou (moyenne, gaussien) convolution (séparable, non séparable) noyaux de convolution (sobel, laplacien, gabor, gaussien, etc.) changement d'espace couleur (Lab, LMS, YUV, YCrCb, etc.) transformations géométriques (translation, rotation, échelle) normalisation (min/max, moyenne/écart-type, etc.) symétrie radiale (Loy & Zelinski) redimentionnement (sur / sous échantillonnage) seuillage (binaire, masque, adaptatif, etc.)
Histogrammes	1D, 2D, 3D équalisation, normalisation, rétro-projection
Statistiques	min, max, moyenne, écart-type, médian, entropie, etc. min et max locaux

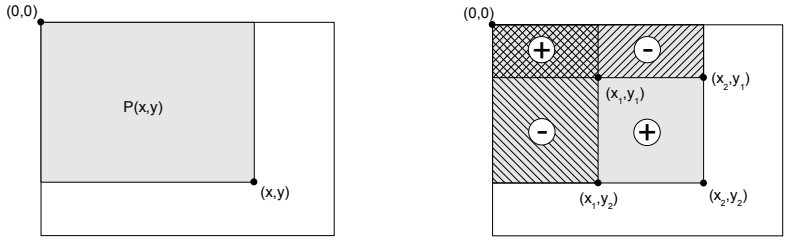
TABLE D.1: Principales fonctionnalités de la librairie SharpVision.

D.3.1 Intérêt des images intégrales

La partie hiérarchique de notre modèle d'attention fait un usage massif des filtres centre-périphérie. La vitesse de calcul de ces filtres est donc un élément clé de la performance computationnelle du système.

Ces filtres sont généralement calculés par une différence de gaussienne. Cette opération est relativement rapide, car réalisable par une convolution séparable. En acceptant une perte de précision des calculs, il est cependant possible d'aller plus vite en remplaçant les gaussiennes par des filtres boite (moyenneur), calculés au moyen d'images intégrales.

Principe de fonctionnement



(a) Somme $P(x, y)$ sur le rectangle $(0, 0)$ (x, y).

(b) Somme $S(x_1, x_2, y_1, y_2)$ sur le rectangle (x_1, y_1) (x_2, y_2).

FIGURE D.3.1: Calcul de la somme d'une zone rectangulaire (x_1, y_1) à partir d'une image intégrale P.

Le principe des images intégrales a été introduit par Crow [Crow 84] puis généralisé par Heckbert [Heckbert 86] au milieu des années 1980. Il exploite le fait que la convolution d'une fonction f avec une fonction g est équivalente à la convolution de l'intégrale de f par la dérivée de g :

$$f \star g = (\int f(x)dx) \star (\frac{dg}{dx})$$

La dérivée d'un filtre boite se résume à deux impulsions : une positive au début du filtre et une négative à la fin de celui-ci. En une dimension, on peut alors calculer la somme d'un segment d'un signal en soustrayant deux termes. En étendant ce principe à deux dimensions, le calcul de la somme des éléments contenus dans une zone rectangulaire quelconque est réalisable en additionnant et soustrayant quatre termes. On pourra ainsi, pour une image I, calculer la somme S des pixels de la zone rectangulaire délimitée par les points (x_1, y_1) et (x_2, y_2)(figure D.3.1) de la manière suivante :

$$\begin{aligned} S(x_1, x_2, y_1, y_2) &= \sum_{x_1 \leq x \leq x_2} \sum_{y_1 \leq y \leq y_2} I(x, y) \\ S(x_1, x_2, y_1, y_2) &= P(x_2, y_2) - P(x_1, y_2) - P(x_2, y_1) + P(x_1, y_1) \end{aligned}$$

avec

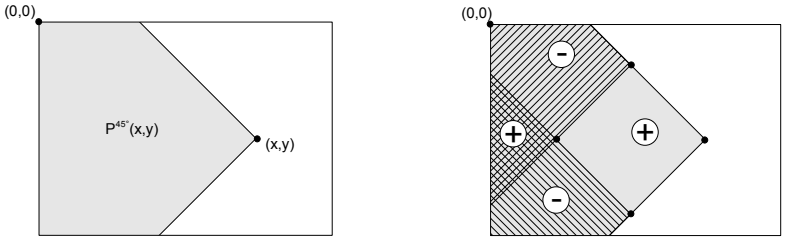
$$P(x,y) = \sum_{i=0}^{x} \sum_{j=0}^{y} I(i,j)$$

P est une image intégrale. Chacun de ses éléments $P(x,y)$ contient la somme des pixels de I appartenant au rectangle délimité par les points de coordonnées $(0,0)$ et (x,y). P peut être calculée efficacement en une seule passe par la formule récursive suivante :

$$P(x,y) = P(x,y-1) + P(x-1,y) + I(x,y) - P(x-1,y-1)$$

avec $P(-1,y) = P(x,-1) = 0$.

Plus récemment, Lienhart [Lienhart 02] a proposé le calcul d'une image intégrale $P^{45°}$ orientée à 45° (figure D.3.2). Celle-ci permet de calculer la somme des pixels appartenant à un rectangle lui aussi orienté à 45°. Le calcul de $P_{45°}$ est plus lent que celui de P car il doit être réalisé en deux passes. Le calcul de la somme de la zone rectangulaire $S^{45°}_{x_1,x_2,y_1,y_2}$ est par contre toujours aussi rapide.



(a) Somme $P(x,y)$ sur la zone orienté à 45° $(0,0)$ (x,y).

(b) Somme $S^{45°}$ sur un rectangle orienté à 45°.

FIGURE D.3.2: Calcul de la somme d'une zone rectangulaire orientée à 45° à partir d'une image intégrale$P^{45°}$.

Evaluation des performances

Tous les tests présentés dans ce paragraphe ont été effectués sur des images couleur. Ils ont été effectués sur un ordinateur portable HP équipé d'un processeur (dual core) Intel T2400 à 1.83GHz avec 2Go de RAM, sous Windows XP SP3. Le code utilisé est une version parallélisée des opérateurs (*via* SharpVision). La génération des images intégrales n'est pas parallélisée (c'est une opération intrinsèquement sérielle : chaque élément dépend du précédent).

Les filtres centre-périphérie sont réalisés par une différence de filtres effectuant un moyennage (pondéré ou non), nous avons donc testé différents opérateurs de flou et comparé leurs performances respectives :

- flou moyenneur par image intégrale. Cette opération s'effectue en deux étapes : génération de l'image intégrale, puis flou à partir de l'image intégrale ;
- flou moyenneur « classique » séparable ;
- flou moyenneur optimisé. Cette version du filtre moyenneur exploite le fait que les masques de calcul du flou de deux pixels voisins possèdent de nombreuses valeurs communes. On peut alors économiser de nombreux calculs en réutilisant ces valeurs. Cette méthode est très rapide pour les filtres de grande taille, mais possède un coût constant de mise en œuvre la rendant moins efficace sur une combinaison de filtre et image de petite taille (tableau D.3) ;
- flou gaussien.

Les temps de calcul présentés au tableau D.2 montrent que le filtrage à base d'images intégrales est particulièrement intéressant pour des tailles de filtres importantes, mais reste dans tous les cas de figure plus rapide que les filtres moyenneurs « classiques » et gaussiens.

Dans le cas d'un traitement multi-résolutions, les filtres utilisés sont généralement de petite taille (typiquement 3x3 ou 9x9). Dans ce cas, le filtrage par images intégrales est bien plus efficace que les filtres séparables classiques, mais le filtre moyenneur « optimisé » reste une bonne option.

Deux avantages font cependant pencher la balance en faveur des images intégrales :

- la parallélisation du code : la construction des images intégrales n'est pas parallélisée (même pour les différents canaux d'une image couleur). Cependant, dans le cas du traitement des différentes caractéristiques de notre système de vision (intensité, couleur, orientation, mouvement), celles-ci pourront être traitées en parallèle. On gagnera alors du temps sur la génération des images intégrales. Pour 4 caractéristiques, des filtres 5x5, et une image 384x256, on aura par exemple sur un processeur *quad-core* :
  - 19 + 4*14 = 75 ms dans le cas du filtrage par image intégrale (les 4 images intégrales étant générées en parallèle).
  - 4*40 = 160 ms dans le cas du filtre moyenneur optimisé.
- la réutilisation des images intégrales pour tous les niveaux de la pyramide. Pour un traitement appliqué à plusieurs résolutions d'une pyramide, l'image intégrale ne sera calculée qu'une seule fois. Pour une image 768x512, des filtres 5x5 et deux niveaux de pyramide on aura :
  - 73+55+14 = 142 ms dans le cas du filtrage par image intégrale.
  - 137 + 40 = 177 ms dans le cas du filtre moyenneur optimisé.

En prenant en compte ces différents paramètres, le filtrage par image intégrale devient particulièrement intéressant.

	3x3	5x5	11x11	23x23
Image intégrale	126(70+56)	128 (73+55)	129 (72+57)	134(77+57)
Moyenneur « classique »	157	173	254	416
Moyenneur optimisé	137	137	142	148
Moyenneur gaussien	157	173	254	416

TABLE D.2: Temps de calcul (en millisecondes) des différentes méthodes de flou pour différentes tailles de filtre. L'image source a une taille de 768x512.

	3x3	5x5	11x11	23x23
Image intégrale	33(19+14)	33(19+14)	33(19+14)	35(21+14)
Moyenneur « classique »	37	45	69	110
Moyenneur optimisé	40	40	40	40
Moyenneur gaussien	37	45	69	110

TABLE D.3: Temps de calcul (en millisecondes) des différentes méthodes de flou pour différentes tailles de filtre. L'image source a une taille de 384x256.

D.3.2 Influence de la taille des images

Les temps de calcul moyens de notre modèle ont été calculés pour 250 itérations sur deux configurations :

- Configuration 1 : portable HP équipé d'un processeur (dual core) Intel T2400 à 1.83GHz avec 2Go de RAM, sous Windows XP SP3.
- Configuration 2 : ordinateur Shuttle équipé d'un processeur (quad core) Intel Q9950 à 2.83GHz avec 4Go de RAM, sous Windows Vista SP1.

Ces temps de calcul ont été mesurés pour trois cas d'utilisation de l'algorithme :

- sans flou : c'est la configuration par défaut de l'algorithme, aucun flou rétinien n'est utilisé. Les cartes de caractéristiques, de singularité et le système proies / prédateurs sont recalculés à chaque image. C'est la configuration utilisée le plus souvent pour le traitement des vidéos ;
- avec flou : même configuration et utilisation que précédemment. On utilise cette fois le flou rétinien, permettant d'améliorer la qualité des résultats, mais également les temps de calcul ;
- proies / prédateurs seul : seul le système proies / prédateurs est mis à jour. C'est la configuration couramment utilisée pour traiter les images fixes, après une première itération en configuration « sans flou ».

Dans son paramétrage par défaut, notre modèle effectue ses calculs multi-résolutions sur les niveaux 1 à N-2 de la pyramide (avec N le nombre maximal de niveaux calculables jusqu'à une taille de 1x1). En utilisant un nombre de niveaux de résolution variable, on

garantit une analyse à toutes les échelles quelle que soit la taille de l'image.

	160x120	256x192	320x240	512x384	640x480	800x600	1024x768
Nombre de niveaux	3	4	4	5	5	6	6

TABLE D.4: Nombre de niveaux de résolution calculés en fonction de la taille des images.

En observant les tableaux D.5 et D.6, on constate que pour la première configuration, les temps de calcul restent raisonnables jusqu'à une résolution de 320x240. Au-delà, il devient difficile de parler de temps réel. Pour la deuxième configuration cette limite est repoussée à 512x384.

	160x120	256x192	320x240	512x384	640x480	800x600	1024x768
Sans flou	19	50	77	199	301	489	787
Avec flou	16	39	57	148	229	412	584
Proies / prédateurs	5	8	12	30	47	70	97

TABLE D.5: Temps de calcul en millisecondes sur portable HP

	160x120	256x192	320x240	512x384	640x480	800x600	1024x768
Sans flou	10	23	36	92	136	234	333
Avec flou	7	19	29	77	118	208	269
Proies / prédateurs	1	3	5	14	21	29	38

TABLE D.6: Temps de calcul en millisecondes sur ordinateur Shuttle

A titre de comparaison, le tableau D.7 présente les temps de calcul de l'algorithme de Laurent Itti [Itti 98], mesurés par Achanta [Achanta 08] sur une machine équipée d'un processeur Intel double-coeur à 2.26 GHz possédant 1Go de RAM. On constate que notre algorithme est dix fois plus rapide, bien que nous ayons calculé plus de niveaux de résolution.

Dans [Frintrop 05b], Frintrop annonce des temps de calcul de 50ms et 190ms pour des images de taille respective 400x300 et 800x600 (en utilisant un processeur à 2.8 GHz). Notre algorithme semble alors plus lent. La comparaison n'est cependant pas équitable car pour une image 800x600 notre algorithme travaille sur 6 niveaux de résolution contre 3 pour Frintrop. En contraignant notre algorithme à travailler sur 3 niveaux, on obtient des temps de calcul plus favorables : 34ms pour une image 400x300 et 123ms pour une image de 800x600 (sur le Shuttle à 2.83GHz).

	320x240	640x480	800x600	1024x768
Temps de calcul	750	2540	4400	7500

TABLE D.7: Temps de calcul en millisecondes du modèle d'Itti [Itti 98].

D.3.3 Influence de la parallélisation

Les tableaux D.8 et D.9 montrent l'influence de la parallélisation des traitements de notre modèle sur les ordinateurs multi-coeurs utilisés pour nos tests. Les gains moyens constatés sont les suivants :

- passage de 1 coeur à 2 coeurs : 1.5x plus rapide en moyenne. Le taux d'occupation du CPU passe de 100% à environ 85% ;
- passage de 1 coeur à 4 coeurs : 1.7x plus rapide en moyenne. Le taux d'occupation du CPU passe de 100% à environ 70% ;

	T 2400 - 1 cœur		T 2400 - 2 cœurs	
	Temps	% CPU	Temps	% CPU
Sans flou	304	100%	199	93%
Avec flou	229	100%	148	89%
Proies / prédateurs	42	100%	30	75%

TABLE D.8: Influence de la parallélisation sur le temps de calcul, pour une taille d'image de 512x384. Pour le portable HP.

	Q9550 - 1 cœur		Q9550 - 2 cœurs		Q9550 - 4 cœurs	
	Temps	% CPU	Temps	% CPU	Temps	% CPU
Sans flou	168	100%	109	84%	92	75%
Avec flou	127	100%	87	80%	77	70%
Proies / prédateurs	23	100%	17	76%	14	60%

TABLE D.9: Influence de la parallélisation sur le temps de calcul, pour une taille d'image de 512x384. Pour l'ordinateur Shuttle.

La parallélisation est intéressante, puisque l'on augmente significativement la vitesse de traitement de notre modèle. Cependant, les gains obtenus sont assez éloignés des gains théoriques possibles (1.5x contre 2x en théorie en passant de 1 à 2 cœurs et 1.7x contre 4x en théorie en passant de 1 à 4 cœurs). On peut se demander pourquoi les gains sont aussi limités, surtout dans le cas du passage à 4 cœurs. La réponse est double :

- bien qu'une grande partie du code soit parallélisée (grâce à l'utilisation de la librairie *SharpVision*), il reste des traitements séquentiels dans notre modèle. De plus, *SharpVision* désactive d'elle-même la parallélisation pour les traitements effectués

sur de petites images car celle-ci n'est pas efficace quand la quantité de données à traiter est trop peu importante (le processeur passe plus de temps à synchroniser les threads qu'à effectuer les calculs). La combinaison de ces deux facteurs empêche une parallélisation massive ;

– les algorithmes utilisés dans notre modèle ont une complexité arithmétique assez basse. Par contre, les opérations de filtrage et de mise à jour du système proies / prédateurs effectuent des accès mémoire massifs. Le cache et le bus mémoire étant partagés par les différents cœurs, les accès mémoire deviennent rapidement un facteur limitant. Ajouter plus de cœurs n'améliore alors plus les performances de manière aussi importante que prévue.

Annexe E

Bases d'images

Nous présentons ici les 3 bases d'images utilisées lors de nos évaluations de performances.

E.1 Bruce

Cette base comporte 120 images couleur, représentant des photographies typiques d'un environnement urbain d'intérieur (lieux de vie, objets) ou d'extérieur (rues, jardins, véhicules, bâtiments). Certaines images contiennent des objets très saillants, d'autres non. La base est disponible sur le site web de son auteur (`http://www-sop.inria.fr/members/Neil.Bruce/`).

Le protocole d'acquisition de la vérité terrain est le suivant :
- présentation des images dans un ordre aléatoire pendant 4 secondes ;
- un masque est affiché entre deux images successives ;
- les sujets sont positionnés à 75cm d'un moniteur CRT 21 de 21 pouces de diagonale ;
- aucune instruction particulière n'est donnée aux sujets, à part d'observer les images ;
- l'appareil d'*eye-tracking* est un dispositif « sur table » ;
- les données sont collectées sur un ensemble de 20 personnes.

L'ensemble des différentes images est représenté figures E.1.1 et E.1.2.

E.2 Le Meur

Cette base contient 26 images couleur, dont certaines issues d'expérimentations effectuées par Laurent Itti. Les types d'images sont variés : scènes de sport, animaux, bâtiments, scènes d'intérieur, paysages. Elles sont disponibles sur le site web de leur

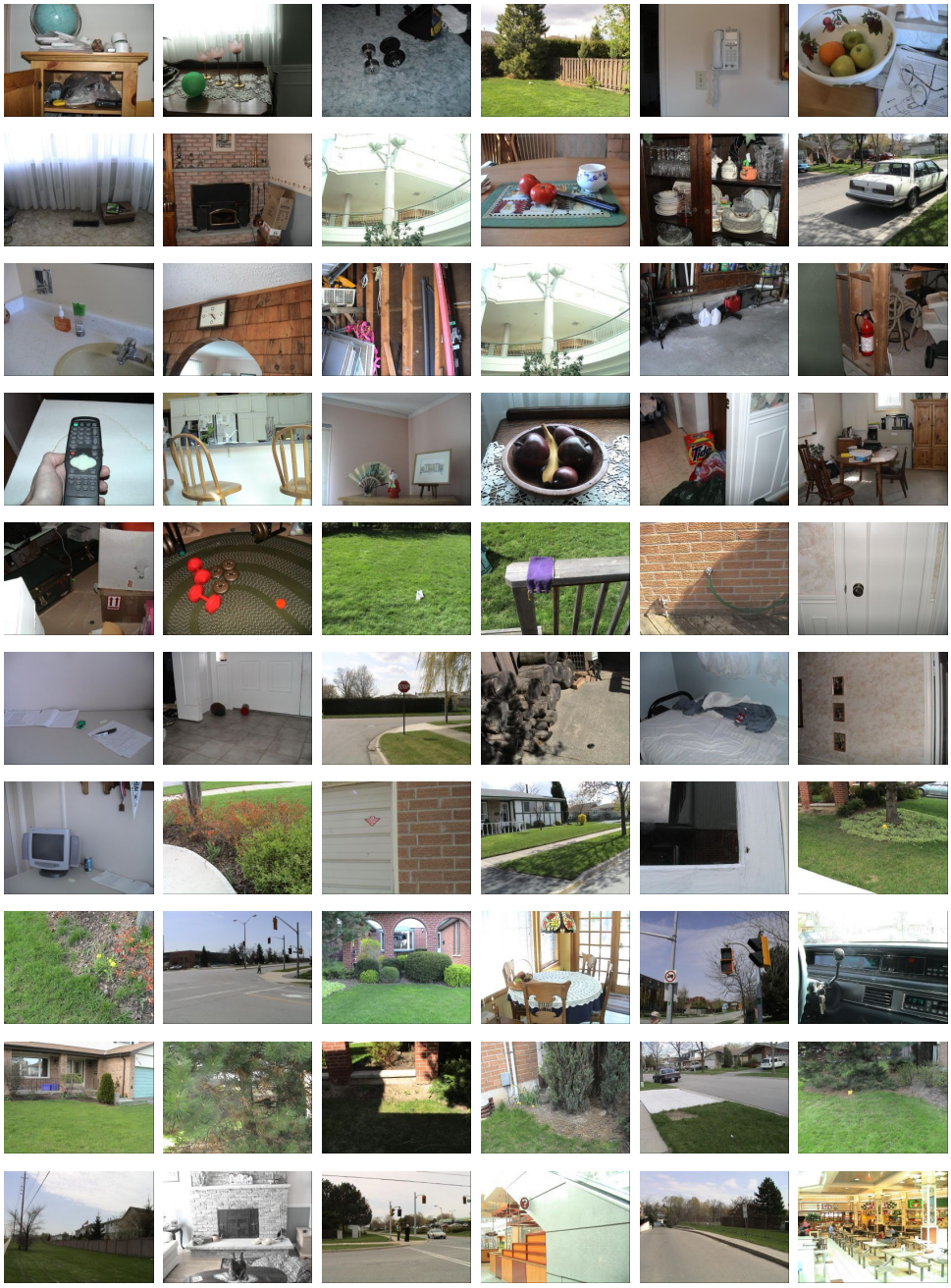


FIGURE E.1.1: Images de la base Bruce (1ère partie).

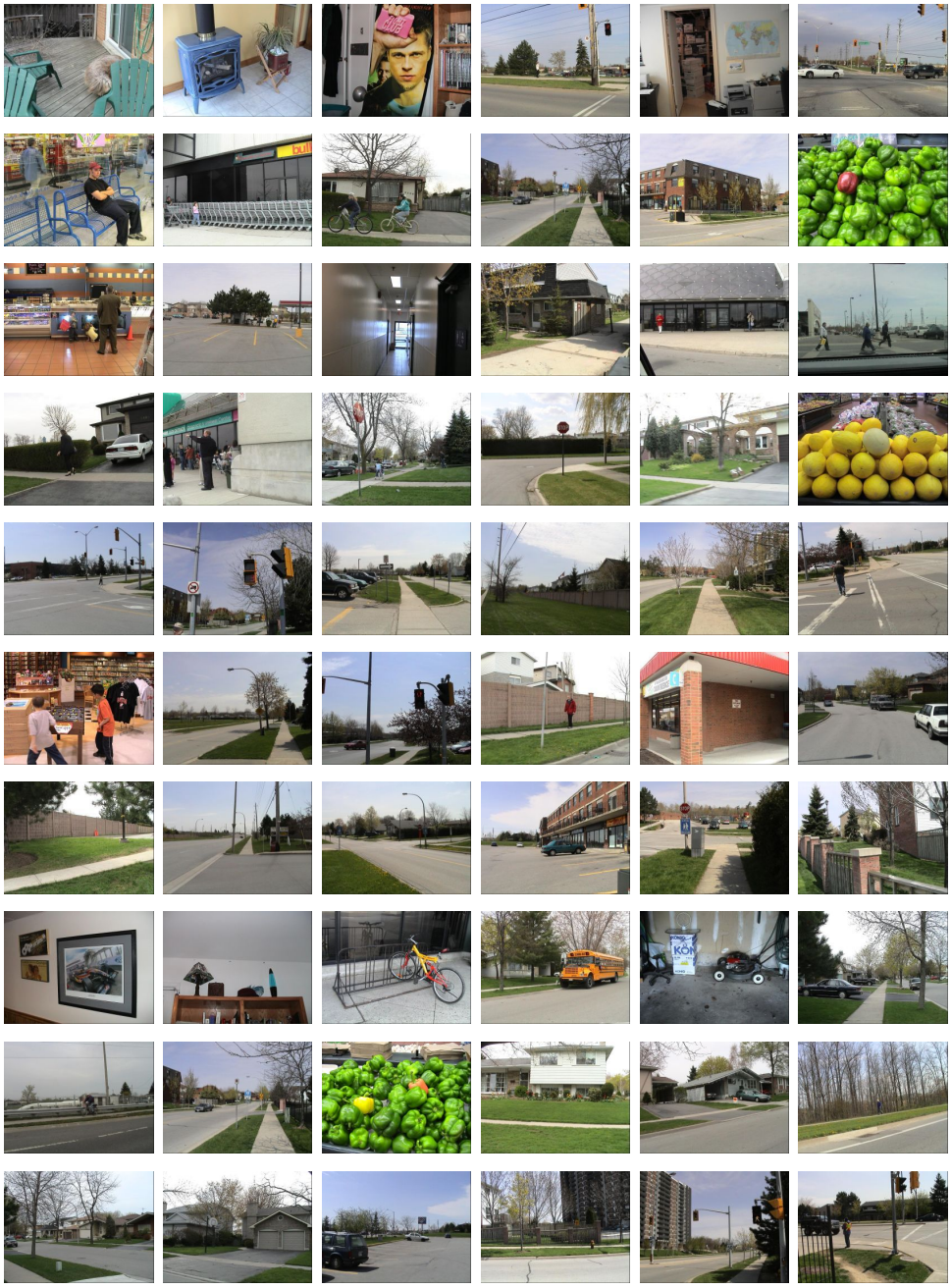


FIGURE E.1.2: Images de la base Bruce (2ème partie).

auteur (`http://www.irisa.fr/temics/staff/lemeur/visualAttention/`).

Le protocole d'acquisition de la vérité terrain est le suivant :
- présentation des images pendant 15 secondes ;
- les sujets sont positionnés à $4H$ d'un moniteur ayant un écran de hauteur H et de résolution 800x600 ;
- la fréquence d'acquisition des données est de 50Hz ;
- aucune instruction particulière n'est donnée aux sujets, à part d'observer les images ;
- l'appareil d'*eye-tracking* est un dispositif « sur table » ;
- les données sont collectées sur un ensemble de 40 personnes, cependant les données contenant des erreurs ayant été supprimées de la base, ce nombre varie en fonction des images (entre 22 et 40) ;

L'ensemble des différentes images est représenté figure E.2.1.

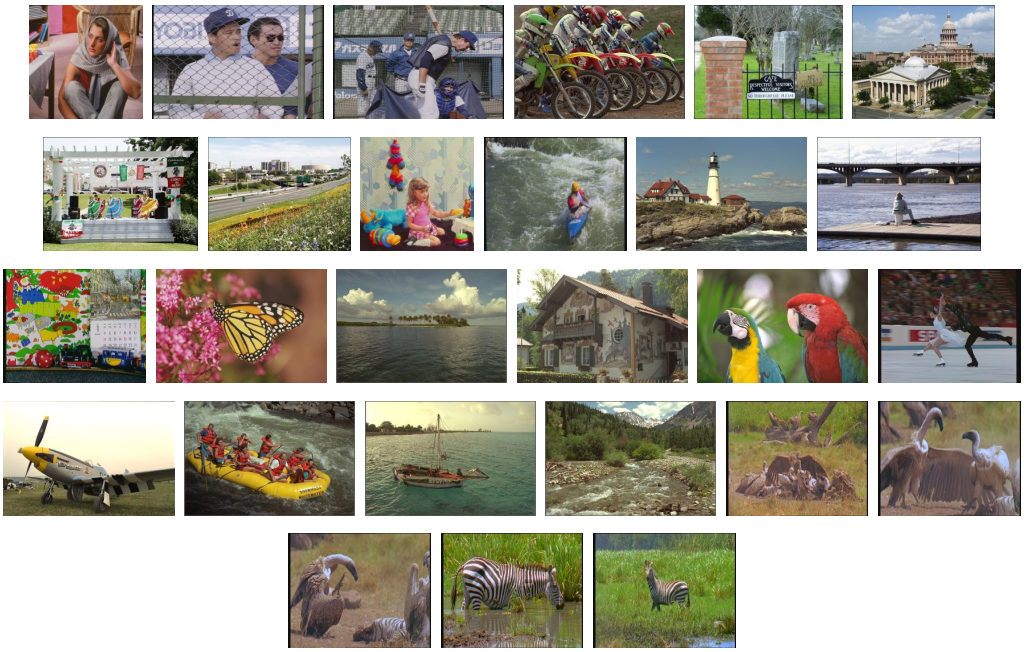


FIGURE E.2.1: Images de la base Le Meur.

E.3 Flick'r

Cette base a été utilisée uniquement dans les expériences de comparaison subjective. Aucune donnée de vérité terrain n'est donc disponible. Les images ont été téléchargées sur le site Flick'r (http ://www.flickr.com) parmi les images ayant une licence *creative commons* « attribution ». Cette licence permet une exploitation libre, tant que l'on

respecte la paternité des images. Six catégories ont été définies. Les images correspondant à chacune d'elles sont consultables figure E.3.1.

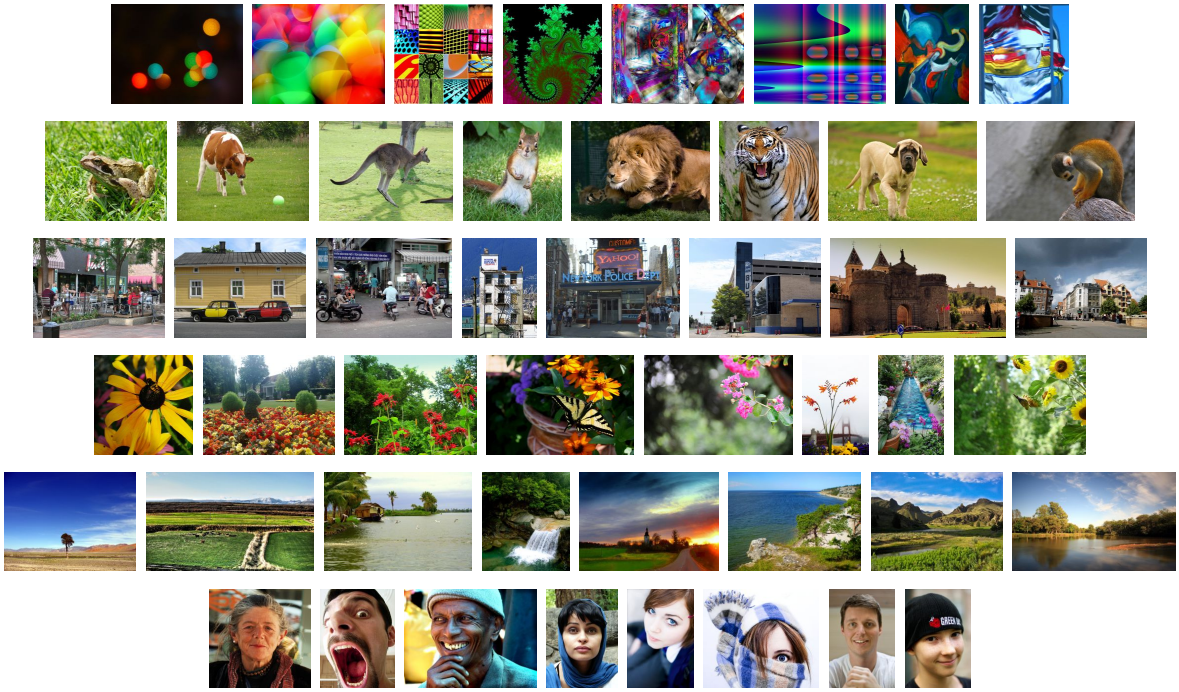


FIGURE E.3.1: Images de la base Flick'r. Chaque ligne représente une des 6 catégories (de haut en bas) : abstrait, animaux, ville, fleurs, paysages et visages.

Bibliographie

[Achanta 08] Radhakrishna Achanta, Francisco Estrada, Patricia Wils & Sabine Süsstrunk. *Salient region detection and segmentation.* In 6th International Conference on Computer Vision Systems, ICVS, pages 66–75, Berlin, Heidelberg, 2008. Springer. www

[Ahmad 92] Subutai Ahmad. *VISIT : An efficient computational model of human visual attention.* University of Illinois at Urbana-Champaign, Champaign, IL, no. 510, 1992. www

[Allport 87] D. A Allport. Selection for action : Some behavioral and neurophysiological considerations of attention and action, pages 395–419. Lawrence Erlbaum Associates, Hillsdale, NJ, 1987.

[Avraham 10] Tamar Avraham & Michael Lindenbaum. *Esaliency (extended saliency) : meaningful attention using stochastic image modeling.* IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 4, pages 693–708, April 2010. www

[Aziz 06] M.Z. Aziz, B. Mertsching, M. Salah, E.-N. Shafik & R. Stemmer. *Evaluation of Visual Attention Models for Robots.* In Fourth IEEE International Conference on Computer Vision Systems (ICVS'06), numéro Icvs, pages 20–20. Ieee, 2006. www

[Aziz 08a] M. Aziz & B Mertsching. *Visual search in static and dynamic scenes using fine-grain top-down visual attention.* In 6th International Conference on Computer Vision Systems, ICVS, volume vol5008, pages 3–12. Springer, 2008. www

[Aziz 08b] MZ Aziz & Bärbel Mertsching. *Fast and robust generation of feature maps for region-based visual attention.* IEEE transactions on image processing : a publication of the IEEE Signal Processing Society, vol. 17, no. 5, pages 633–44, May 2008. www

[Aziz 09a] M. Aziz & B. Mertsching. *Early Clustering Approach towards Modeling of Bottom-Up Visual Attention.* In KI 2009 : Advances in Artificial Intelligence, volume 1289, pages 315–322. Springer, 2009. www

[Aziz 09b] M.Z. Aziz & B. Mertsching. *Towards Standardization of Evaluation Metrics and Methods for Visual Attention Models.* In Attention in Cognitive Systems, pages 227–241. Springer, 2009. www

[Baldi 05] P. Baldi & Laurent Itti. *Attention : Bits versus Wows.* In 2005 International Conference on Neural Networks and Brain, numéro 1, pages 56–61. Ieee, 2005. www

[Bamidele 04] a Bamidele, FWM Stentiford & J Morphett. *An attention-based approach to content-based image retrieval.* BT Technology Journal, vol. 22, no. 3, pages 151–160, July 2004. www

[Ban 04] S Ban. *A face detection using biologically motivated bottom-up saliency map model and top-down perception model.* Neurocomputing, vol. 56, pages 475–480, January 2004. www

[Ban 06] S.W. Ban & Minho Lee. *Selective attention-based novelty scene detection in dynamic environments.* Neurocomputing, vol. 69, no. 13-15, pages 1723–1727, 2006. www

[Bednar 02] James A Bednar. *Learning to See : Genetic and Environmental Influences on Visual Development.* Phd, University of Texas, 2002.

[Behnke 03] Sven Behnke. Hierarchical neural networks for image interpretation, volume 2766. Springer-Verlag New York Inc, 2003. www

[Belardinelli 09] Anna Belardinelli, Fiora Pirri & Andrea Carbone. *Motion Saliency Maps from Spatiotemporal Filtering.* In Lecture Notes In Artificial Intelligence, pages 112–123. Springer, 2009. www

[Benoit 10] a. Benoit, a. Caplier, B. Durette & J. Herault. *Using human visual system modeling for bio-inspired low level image processing.* Computer Vision and Image Understanding, 2010. www

[Berthoz 09] Alain Berthoz. La simplexité. Paris, odile jaco edition, 2009.

[Bremond 06] Roland Bremond, Jean-Philippe Tarel, Hicham Choukour & Marion Deugnier. *La saillance visuelle des objets routiers, un indicateur de la visibilité routière.* In Journées des Sciences de l'Ingénieur, pages 2–7, Marne-la-Vallée, 2006.

[Broadbent 58] D. E. Broadbent. Perception and communication. Pergamon Press, Elmsford, NY, US, 1958.

[Brodu 07] Nicolas Brodu. *Practical Investigations of Complex Systems.* Phd, Concordia University, 2007.

[Bruce 03] B. Bruce & E. Jernigan. *Evolutionary design of context-free attentional operators.* In proc. ICIP'03, pages 0–3. Citeseer, 2003. www

[Bruce 08] N.D.B. Bruce & J.K. Tsotsos. *Spatiotemporal Saliency : Towards a Hierarchical Representation of Visual Saliency.* In Proc. 5th Int. Workshop on Attention in Cognitive Systems, pages 98–111. Springer, 2008. www

[Bruce 09] N D B Bruce & J K Tsotsos. *Saliency, attention, and visual search : An information theoretic approach.* Journal of Vision, vol. 9, no. 3, page 5, 2009.

[Bundesen 87] C Bundesen. *Visual attention : race models for selection from multi-element displays.* Psychol. Res., vol. 49, pages 113–121, 1987.

[Bundesen 98] C Bundesen. *A computational theory of visual attention.* Philosophical transactions of the Royal Society of London. Series B, Biological sciences, vol. 353, no. 1373, pages 1271–81, August 1998. www

[Bur 07a] A Bur, P Wurtz, RM Miiri & H. *Dynamic visual attention : competitive versus motion priority scheme.* In ICVS Workshop on Computational Attention and Applications, pages 1–10, 2007. www

[Bur 07b] a. Bur, P. Wurtz, R. M. Müri & H. Hügli. *Motion integration in visual attention models for predicting simple dynamic scenes.* In Proceedings of SPIE, numéro 47, pages 649219–649219–11. Spie, 2007. www

[Bur 07c] Alexandre Bur & Heinz Hugli. *Optimal cue combination for saliency computation : A comparison with human vision.* Lecture Notes in Computer Science, vol. 4528, pages 109–118, 2007. www

[Camus 07] Mickaël Camus. *Système auto-adaptatif générique pour le contrôle de robots ou d'entités logicielles.* Thèse de doctorat, Université Pierre et marie Curie, 2007.

[Carmi 06] Ran Carmi & Laurent Itti. *Visual causes versus correlates of attentional selection in dynamic scenes.* Vision research, vol. 46, no. 26, pages 4333–45, 2006. www

[Chalupa 03] L.M. Chalupa & J.S. Werner, editeurs. The Visual Neurosciences. MIT Press, 1ère éditi edition, 2003.

[Chamaret 09] Christel Chamaret & O Le Meur. *Attention-based video reframing : validation using eye-tracking.* In Pattern Recognition, 2008. ICPR, pages 1–4. Ieee, December 2009. www

[Chanceaux 09] Myriam Chanceaux, A. Guérin-Dugué, Benoit Lemaire & T. Baccino. *Simulations cognitives de trajets oculomoteurs lors*

d'une recherche d'information sur des pages numériques. In Proceedings of the 21st International Conference on Association Francophone d'Interaction Homme-Machine, pages 163–172. ACM, 2009. www

[Chauvin 03] Alan Chauvin. *Perception des scènes naturelles : étude et simulation du rôle de l'amplitude, de la phase et de la saillance dans la catégorisation et l'exploration des scènes naturelles.* Thèse de doctorat, Laboratoire LIS, Grenoble, 2003. www

[Choe 01] Yoonsuck Choe. *Perceptual Grouping in a Self-Organizing Map of Spiking Neurons.* Phd, University of Texas, 2001.

[Choi 06] Sang-bok Choi, Bum-soo Jung & Sang-woo Ban. *Biologically motivated vergence control system using human-like selective attention model.* Neurocomputing, vol. 69, no. 4-6, pages 537–558, January 2006. www

[Courboulay 02] Vincent Courboulay. *Une nouvelle approche variationnelle du traitement d'images. Application à la coopération détection-reconstruction.* Thèse de doctorat, Université de La Rochelle, 2002. www

[Crow 84] Franklin Crow. *Summed-area tables for texture mapping.* Computer Graphics, vol. 18, no. 3, pages 207–212, 1984.

[Dankers 07] Andrew Dankers, Nick Barnes & Alex Zelinsky. *A reactive vision system : Active-dynamic saliency.* In 5th International Conference on Computer Vision Systems (ICVS), numéro Icvs, Bielefeld, Germany, 2007. Applied Computer Science Group. www

[Deco 04] G Deco. *A Neurodynamical cortical model of visual attention and invariant object recognition.* Vision Research, vol. 44, no. 6, pages 621–642, 2004. www

[Desimone 95] R Desimone & J Duncan. *Neural mechanisms of selective visual attention.* Annual review of neuroscience, vol. 18, pages 193–222, January 1995. www

[Desolneux 01] A. Desolneux, Lionel Moisan & J.M. Morel. *Partial gestalts*, 2001. `http://www.cmla.ens-cachan.fr/fileadmin/Documentation/Prepublications/2001/CMLA2001-22.pdf`

[Desolneux 04] A. Desolneux, L. Moisan & J.M. Morel. Gestalt theory and computer vision, chapitre 3, pages 71–101. Springer, Berlin, Heidelberg, 2004. www

[Desolneux 08] Agnés Desolneux, Lionel Moisan & Jean-Michel Morel. From Gestalt Theory to Image Analysis, volume 34 of *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, NY, interdisci edition, 2008. www

[Deutsch 63] J. a. Deutsch & D. Deutsch. *Attention : Some theoretical considerations.* Psychological Review, vol. 70, no. 1, pages 51–60, 1963. www

[Dong 06] Le Dong, S.W. Ban & Minho Lee. *Biologically Inspired Selective Attention Model Using Human Interest.* International Journal of Information Technology, vol. 12, no. 2, pages 140–148, 2006. www

[Draper 05] B Draper & A Lionelle. *Evaluation of selective attention under similarity transformations.* Computer Vision and Image Understanding, vol. 100, no. 1-2, pages 152–171, October 2005. www

[Draper 07] B.A. Draper. *A Biomimetic Vision Architecture.* In 5th International Conference on Computer Vision Systems (ICVS), Bielefeld, Germany, 2007. Applied Computer Science Group. www

[Driver 01] J Driver. *A selective review of selective attention research from the past century.* British Journal of Psychology, vol. 92, no. 1, pages 53–78, 2001. www

[Duncan 84] John Duncan. *Selective attention and the organization of visual information.* Journal of Experimental Psychology : General, vol. 113, no. 4, pages 501–517, 1984. www

[Eklundh 01] JO Eklundh & H Christensen. Computer vision : Past and future, pages 328–340. Springer-Verlag, Berlin, 2001. www

[Eliasmith 95] Chris Eliasmith. *Mind as a dynamical system.* Thèse de master, University of Waterloo, 1995.

[Eriksen 85] Charles W. Eriksen & Yei-yu Yeh. *Allocation of attention in the visual field.* Journal of Experimental Psychology : Human Perception and Performance, vol. 11, no. 5, pages 583–597, 1985. www

[Fay 05] Rebecca Fay, Ulrich Kaufmann, Andreas Knoblauch, H. Markert & G. Palm. *Combining visual attention, object recognition and associative information processing in a neurobotic system.* Biomimetic Neural Learning for Intelligent Robots, pages 118–143, 2005. www

[Fecteau 06] Jillian H Fecteau & Douglas P Munoz. *Salience, relevance, and firing : a priority map for target selection.* Trends in cognitive sciences, vol. 10, no. 8, pages 382–90, August 2006. www

[Fix 08] Jérémy Fix. *Mécanismes numériques et distribués de lanticipation motrice.* Thèse de doctorat, Université Henri Poincaré - Nancy 1, 2008.

[Forsythe 09] Alexandra Forsythe. *Visual Complexity : Is That All There Is ?* Complexity, pages 158–166, 2009.

[Fox 07] Michael D. Fox, Abraham Z. Snyder, Justin L. Vincent & Marcus E. Raichle. *Intrinsic Fluctuations within Cortical Systems Account for Intertrial Variability in Human Behavior.* Neuron, vol. 56, no. 1, pages 171–184, October 2007. www

[Frintrop 05a] Simone Frintrop. *VOCUS :AVisual Attention System for Object Detection and Goal-Directed Search.* Phd, University of Bonn, 2005.

[Frintrop 05b] Simone Frintrop, Gerriet Backer & Erich Rome. *Selecting what is important : Training visual attention.* In 28th Annual German Conference on AI (KI), pages 351–366, Koblenz, Germany, 2005. Springer Verlag. www

[Frintrop 07] Simone Frintrop, Maria Klodt & Erich Rome. *A real-time visual attention system using integral images.* In 5th International Conference on Computer Vision Systems (ICVS), Bielefeld, Germany, 2007. Applied Computer Science Group. www

[Gaborski 03] Roger S. Gaborski, Vishal S. Vaingankar, Vineet S. Chaoji & Ankur M. Teredesai. *Venus : A system for novelty detection in video streams with learning.* In Proceedings of the 17th International FLAIRS Conference, numéro 2000, pages 1–5, Miami, 2003. www

[Garbay 00] C. Garbay. Architectures logicielles et contrôle dans les systèmes de vision, chapitre 7, pages 197–251. Traité IC2, Paris, France, hermès edition, 2000.

[Geerinck 09] Thomas Geerinck, Hichem Sahli, David Henderickx, Iris Vanhamel & Valentin Enescu. *Modeling Attention and Perceptual Grouping to Salient Objects.* In Attention in Cognitive Systems, page 166. Springer London, Limited, 2009. www

[Gilles 96] Sebastien Gilles. *Description and experimentation of image matching using mutual information*, 1996. `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.41.1000&rep=rep1&type=pdf`

[Gottlieb 07] Jacqueline Gottlieb. *From thought to action : the parietal cortex as a bridge between perception, action, and cognition.* Neuron, vol. 53, no. 1, pages 9–16, January 2007. www

[Grossberg 97] Stephen Grossberg, Ennio Mingolla & W.D. Ross. *Visual brain and visual perception : How does the cortex do perceptual grouping ?* Trends in Neurosciences, vol. 20, no. 3, pages 106–111, March 1997. www

[Guillaume 37] Paul Guillaume. La Psychologie de la forme. Flammarion, Paris, 1937.

[Hall 02] Daniela Hall, Bastian Leibe & Bernt Schiele. *Saliency of interest points under scale changes.* In British Machine Vision Conference, pages 646–655, Cardiff, 2002. www

[Hamker 05a] F Hamker. *Modeling Attention : From Computational Neuroscience to Computer Vision.* In Attention and Performance in Computational Vision, pages 118–132. Springer Berlin / Heidelberg, January 2005. www

[Hamker 05b] F Hamker. *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, no. 1-2, pages 64–106, 2005. www

[Hannagan 06] Thomas Hannagan. *Modèles computationnels de l'attention visuelle*, August 2006.

[Hansen 10] Dan Witzner Hansen & Qiang Ji. *In the eye of the beholder : a survey of models for eyes and gaze.* IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 3, pages 478–500, March 2010. www

[Harel 09] Jonathan Harel & Christof Koch. *On the Optimality of Spatial Attention for Object Detection.* In Attention in Cognitive Systems, pages 1–14. Springer, 2009. www

[Heckbert 86] P.S. Heckbert. *Filtering by repeated integration.* In Proceedings of the 13th annual conference on Computer graphics and interactive techniques, volume 20, pages 315–321. ACM, 1986. www

[Hérault 07] Jeanny Hérault & Barthélémy Durette. *Modeling Visual Perception.* In Francisco Sandoval, Alberto Prieto, Joan Cabestany & Manuel Graña, editeurs, International Work-Conference on Artificial Neural Networks, pages 662–675, San Sebastian, Spain, 2007. Springer Berlin / Heidelberg.

[Hu 09] Yiqun Hu, Deepu Rajan & Liang-Tien Chia. *Attention-from-motion : A factorization approach for detecting attention objects in motion.* Computer Vision and Image Understanding, vol. 113, no. 3, pages 319–331, March 2009. www

[Hua 05] Xian-sheng Hua, Lie Lu, Hong-jiang Zhang & Haidian District. *A Generic Framework of User Attention Model and Its Application in Video Summarization.* IEEE Transaction on Multimedia, vol. 7, no. 5, pages 907–919, 2005. www

[Hubel 95] D H Hubel. Eye, brain and vision. Scientific American Library, No 22, 2nd editio edition, 1995.

[Huffman 52] David Huffman. *A Method for the Construction of Minimum-Redundancy Codes*. Proceedings of the IRE, vol. 40, no. 9, pages 1098–1101, September 1952. www

[Idema 05] Timon Idema. *The behaviour and attractiveness of the Lotka-Volterra equations*. Phd, Universiteit Leiden, 2005. www

[Itti 98] Laurent Itti, Christof Koch, E. Niebur & Others. *A model of saliency-based visual attention for rapid scene analysis*. IEEE Transactions on pattern analysis and machine intelligence, vol. 20, no. 11, pages 1254–1259, 1998. www

[Itti 00] Laurent Itti. *Models of Bottom-Up and Top-Down Visual Attention*. Phd, California Institute of Technology, 2000.

[Itti 01a] L Itti & C Koch. *Feature combination strategies for saliency-based visual attention systems*. Journal of Electronic Imaging, vol. 10, pages 161–169, 2001. www

[Itti 01b] Laurent Itti & Christof Koch. *Computational modelling of visual attention*. Nature Reviews Neuroscience, vol. 2, no. 3, pages 194–204, 2001. www

[Itti 04] Laurent Itti. *Automatic foveation for video compression using a neurobiological model of visual attention*. IEEE transactions on image processing : a publication of the IEEE Signal Processing Society, vol. 13, no. 10, pages 1304–18, October 2004. www

[Itti 05a] Laurent Itti & P. Baldi. *A Principled Approach to Detecting Surprising Events in Video*. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), numéro June, pages 631–637. Ieee, 2005. www

[Itti 05b] Laurent Itti, Geraint Rees & J.K. Tsotsos, editeurs. Neurobiology of attention. Academic Press, 1st editio edition, 2005. www

[James 90] William James. The principles of psychology. Dover Publications, volume 1 edition, 1890.

[Ji 08] Zhengping Ji & Juyang Weng. *Where-What Network 1 : Where and What Assist Each Other Through Top-down Connections*. Neuron, pages 61–66, 2008.

[Kadir 01] Timor Kadir & Michael Brady. *Saliency, scale and image description*. International Journal of Computer Vision, vol. 45, no. 2, pages 83–105, 2001. www

[Klein 80] R. Klein. Does oculomotor readiness mediate cognitive control of visual attention ?, pages 259–276. Academic Press, New York, 1980.

[Koch 85] C. Koch & S. Ullman. *Shifts in selective visual attention : towards the underlying neural circuitry.* Hum Neurobiology, vol. 4, no. 4, pages 219–27, 1985. www

[Koch 09] Christof Koch. *Decoding What People See from Where They Look : Predicting Visual Stimuli from Scanpaths.* In Attention in Cognitive Systems, page 15. Springer London, Limited, 2009. www

[Kootstra 08] Gert Kootstra, Arco Nederveen & Bart De Boer. *Paying Attention to symmetry.* In Mark Everingham, Chris Needham & Roberto Fraile, editeurs, 19th British Machine Vision Conference, volume 284, Leeds, May 2008. University of Leeds.

[Kustov 96] A. A. Kustov & D. L. Robinson. *Shared neural control of attentional shifts and eye movements.* Nature, vol. 384, no. 6604, pages 74–77, 1996. www

[LaBerge 89] David LaBerge & Vincent Brown. *Theory of attentional operations in shape identification.* Psychological Review, vol. 96, no. 1, pages 101–124, 1989. www

[Laberge 95] D Laberge. Attentional Processing : The Brain's Art of Mindfulness. Harvard University Press, 1 edition edition, 1995.

[Lam 02] Nina Siu-Ngan Lam, Hong-lie Qiu, Dale a. Quattrochi & Charles W. Emerson. *An Evaluation of Fractal Methods for Characterizing Image Complexity.* Cartography and Geographic Information Science, vol. 29, no. 1, pages 25–35, January 2002. www

[Landragin 04] F. Landragin. *Saillance physique et saillance cognitive.* CORELA, vol. 2, pages 1–25, 2004. www

[Larochelle 00] Mélanie Larochelle & Charles Robitaille. *L'attention : un phénomène aux multiples déficits.* Psychologie Quebec, vol. 17, no. 6, pages 19–23, 2000.

[Le Meur 05a] O. Le Meur. *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, 2005.

[Le Meur 05b] O. Le Meur, Patrick Le Callet, Dominique Barba, D. Thoreau & F. France. *Modélisation spatio-temporelle de l'attention visuelle.* In CORESA 2005, Rennes, France, 2005. www

[Le Meur 06] O. Le Meur, Xavier Castellan, Patrick Le Callet & Dominique Barba. *Efficient saliency-based repurposing method.* In IEEE International Conference on Image Processing, pages 421–424, Atlanta, USA, 2006. www

[Le Meur 09] O. Le Meur & P. Le Callet. *What we see is most likely to be what matters : Visual attention and applications.* In International Conference on Image Processing, Cairo, Egypt, 2009. www

[Lee 05] S Lee, S Choi, Minho Lee & H Yang. *Non-uniform image compression using a biologically motivated selective attention model.* Neurocomputing, vol. 67, pages 350–356, 2005. www

[Lesser 98] Mike Lesser & Murray Dinah. *Mind as a dynamical system : Implications for autism.* In In Psychobiology of autism : current research & practice, 1998.

[Li 01] Z Li. *Computational design and nonlinear dynamics of a recurrent network model of the primary visual cortex.* Neural computation, vol. 13, no. 8, pages 1749–80, August 2001. www

[Li 02] Z Li. *A saliency map in primary visual cortex.* Trends in Cognitive Sciences, vol. 6, no. 1, pages 9–16, January 2002. www

[Lienhart 02] R. Lienhart & J. Maydt. *An extended set of haar-like features for rapid object detection.* In IEEE ICIP, volume 1, pages 900–903. Citeseer, 2002. www

[Liu 06a] F. Liu & M. Gleicher. *Region enhanced scale-invariant saliency detection.* In Proceedings of IEEE ICME, pages 1–4. Citeseer, 2006. www

[Liu 06b] Haoting Liu, Jianqun Yang & Zhehao Wei. *Moving Object Tracking and Vision Navigation Based on Selective Attention Mechanism.* In 2006 IEEE International Conference on Robotics and Biomimetics, pages 1500–1505, Kunming, December 2006. Ieee. www

[Lloyd 01] S. Lloyd. *Measures of complexity : a nonexhaustive list.* IEEE Control Systems Magazine, vol. 21, no. 4, pages 7–8, August 2001. www

[Lopez 06] M Lopez, A Fernandezcaballero, M Fernandez, J Mira & A Delgado. *Motion features to enhance scene segmentation in active visual attention.* Pattern Recognition Letters, vol. 27, no. 5, pages 469–478, 2006. www

[Luck 97] S J Luck, L Chelazzi, S a Hillyard & R Desimone. *Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex.* Journal of neurophysiology, vol. 77, no. 1, pages 24–42, January 1997. www

[Ma 03] Yu-Fei Ma & Hong-Jiang Zhang. *Contrast-based image attention analysis by using fuzzy growing.* In Proceedings of the eleventh ACM international conference on Multimedia - MULTIMEDIA

'03, page 374, New York, New York, USA, 2003. ACM Press. www

[Maki 00] a Maki. *Attentional Scene Segmentation : Integrating Depth and Motion*. Computer Vision and Image Understanding, vol. 78, no. 3, pages 351–373, June 2000. www

[Mancas 07] Matei Mancas. *Computational Attention : Towards attentive computers*. Phd, Faculté Polytechnique de Mons, 2007.

[Mancas 09] Matei Mancas. *Relative Influence of Bottom-Up and Top-Down Attention*. In Attention in Cognitive Systems, volume 5395, pages 212–226. Springer, 2009. www

[Marat 10] Sophie Marat. *Modèles de saillance visuelle par fusion d'informatios sur la luminance, le mouvement et les visages pour la prédiction de mouvement oculaire lors de l'exploration de vidéos*. Thèse de doctorat, Institut polytechnique de Grenoble, 2010. www

[Marr 82] D. Marr. Vision : a computational investigation into the human representation and processing of visual information. W. H. Freeman, San Francisco, 1982.

[Mazer 02] James A. Mazer & Jack L. Gallant. *Evidence for perceptual saliency maps in area V4 during freeviewing visual search*. Journal of Vision, vol. 2, no. 7, 2002.

[Mesulam 81] M M Mesulam. *A cortical network for directed attention and unilateral neglect*. Ann. NeuroL, vol. 10, pages 309–25, 1981.

[Meur 07] O Le Meur, P Le Callet & D Barba. *Construction d' images miniatures avec recadrage automatique basé sur un modèle perceptuel bio-inspiré*. Traitement du Signal, vol. 24, no. 5, pages 323–335, 2007. www

[Michalke 10] Thomas Michalke, Jannik Fritsch & Christian Goerick. *A biologically-inspired vision architecture for resource-constrained intelligent vehicles*. Computer Vision and Image Understanding, 2010. www

[Milanese 94] R. Milanese, H. Wechsler, S. Gill, J.-M. Bost & T. Pun. *Integration of bottom-up and top-down cues for visual attention using non-linear relaxation*. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pages 781–785. IEEE Comput. Soc. Press, 1994. www

[Mole 09] Christopher Mole. *Attention*, January 2009. `http://plato.stanford.edu/archives/fall2009/entries/attention/`

[Mozer 98] Michael C. Mozer & Mark Sitton. *Computational modeling of spatial attention.* Attention, pages 341–393, 1998. www

[Mulhern 08] Gerry Mulhern & Martin Sawey. *Confounds in pictorial sets : The role of complexity and familiarity in basic-level picture processing.* Behavior Research Methods, vol. 40, no. 1, pages 116–129, 2008. www

[Murray 03a] J.D. Murray. Mathematical biology : An introduction. Springer Verlag, Berlin, Heidelberg, 2003. www

[Murray 03b] J.D. Murray. Mathematical Biology : Spatial models and biomedical applications. Springer Verlag, 2003. www

[Murray 05] Dinah Murray, Mike Lesser & Wendy Lawson. *Attention , monotropism and the diagnostic criteria for.* Autism, vol. 9, no. 2, pages 139–156, 2005.

[Nakayama 89] Ken Nakayama & Manfred Mackeben. *Sustained and transient components of focal visual attention.* Vision Research, vol. 29, no. 11, pages 1631–1647, 1989.

[Navalpakkam 05a] Vidhya Navalpakkam, Michael Arbib & Laurent Itti. Attention and scene understanding, pages 197–203. Numeéro December 2004. ACADEMIC PRESS, 2005. www

[Navalpakkam 05b] Vidhya Navalpakkam & Laurent Itti. *Modeling the influence of task on attention.* Vision research, vol. 45, no. 2, pages 205–31, January 2005. www

[Navalpakkam 06] V. Navalpakkam & Laurent Itti. *Top-down attention selection is fine grained.* Journal of Vision, vol. 6, no. 11, page 4, 2006. www

[Neumann 87] Odmar Neumann. Beyond capacity : A functional view of attention. Perspectives on perception and action., pages 361–394. Lawrence Erlbaum Associates, Hillsdale, NJ, England, 1987.

[Orabona 08] Francesco Orabona, Giorgio Metta & Giulio Sandini. *A Proto-object based visual attention model.* In Lucas Paletta, editeur, Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint (WAPCV), pages 198–215, Berlin, Heidelberg, 2008. Springer. www

[Ouerhani 03a] Nabil Ouerhani. *Visual Attention : From Bio-Inspired Modeling to Real-Time Implementation.* Thèse de doctorat, Université de Neuchâtel, 2003.

[Ouerhani 03b] Nabil Ouerhani & Heinz Hugli. *A model of dynamic visual attention for object tracking in natural image sequences.* Lecture Notes in Computer Science, pages 702–709, 2003. www

[Ould Mohamed 07] Abdalahi Ould Mohamed, Matthieu Perreira Da Silva & Vincent Courboulay. *A history of eye gaze tracking Abdallahi Ould Mohamed*, 2007. `http://hal.archives-ouvertes.fr/docs/00/21/59/67/PDF/Rapport_interne_1.pdf`

[Paletta 08] Lucas Paletta & J.K. Tsotsos. *Preface of Attention in Cognitive Systems*. In R. Goebel, J. Siekmann & W. Wahlster, editeurs, Attention in Cognitive Systems, page Preface. Springer Berlin / Heidelberg, 2008.

[Palmer 99] Stephen E Palmer & Les De. *Les théories contemporaines de la perception de Gestalt*. Intellectica, vol. 28, no. 1, pages 53–91, 1999.

[Park 02] S.J. Park, K.H. An & Minho Lee. *Saliency map model with adaptive masking based on independent component analysis*. Neurocomputing, vol. 49, no. 1, pages 417–422, 2002. www

[Perkiö 09] Jukka Perkiö & Aapo Hyvärinen. *Modelling image complexity by independent component analysis , with application to content-based image retrieval*. In 19th International Conference on Artificial Neural Networks : Part II, pages 1–11, Limassol, Cyprus, 2009. Springer Berlin / Heidelberg.

[Perreira Da Silva 07] Matthieu Perreira Da Silva, Vincent Courboulay & Armelle Prigent. *Gameplay experience based on a gaze tracking system*. In COGAIN 2007 : Gaze-based Creativity, Interacting with Games and On-line Communities, pages 1–4, Leicester, 2007. www

[Perreira Da Silva 08] Matthieu Perreira Da Silva, Vincent Courboulay, Armelle Prigent & Pascal Estraillier. *Real-time face tracking for attention aware adaptive games*. In Computer Vision Systems, volume 5008 of *Lecture Notes in Computer Science*, pages 99–108, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. www

[Perreira Da Silva 09a] Matthieu Perreira Da Silva, Vincent Courboulay, Armelle Prigent & Pascal Estraillier. *Attention visuelle et systèmes proies/prédateurs*. In Images), XXIIe Colloque GRETSI (traitement Du Signal Et Des, Dijon, France, 2009. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images. www

[Perreira Da Silva 09b] Matthieu Perreira Da Silva, Vincent Courboulay, Armelle Prigent & Pascal Estraillier. *Fast, low resource, head detection and tracking for interactive applications*. PsychNology Journal, vol. 7, no. 3, pages 243–264, 2009. www

[Perreira Da Silva 10a] Matthieu Perreira Da Silva, V Courboulay, A Prigent & P Estraillier. *Evaluation of preys / predators systems for visual at-*

tention simulation. In VISAPP 2010 - International Conference on Computer Vision Theory and Applications, pages 275–282, Angers, 2010. INSTICC. www

[Perreira Da Silva 10b] Matthieu Perreira Da Silva, Vincent Courboulay, Armelle Prigent & Pascal Estraillier. *Système temps réel de simulation d' attention visuelle : application aux images et séquences d'images.* In RFIA 2010, Caen, France, 2010. www

[Peters 90] I. Peters & R.N. Strickland. *Image complexity metrics for automatic target recognizers.* In Automatic Target Recognizer System and Technology Conference, pages 1–17. Citeseer, 1990. www

[Peters 05] RJ Peters, A Iyer, Laurent Itti & C Koch. *Components of bottom-up gaze allocation in natural images.* Vision Research, vol. 45, pages 2397–2416, 2005. www

[Petitot 08] Jean Petitot. Neurogéométrie de la vision. Paris, France, les éditio edition, 2008.

[Petrou 06] M. Petrou & M.E. Tabacchi. *On the Evaluation of Images Complexity : A Fuzzy Approach.* In Fuzzy logic and applications : 6th international workshop, WILF 2005, Crema, Italy, September 15-17, 2005 : revised selected papers, page 305. Springer-Verlag New York Inc, 2006. www

[Posner 78] Posner, Nissen M. & W. M., Ogden. Attended and unattended processing modes : the role of set for spatial location, pages 137–157. Laurence Erlbaum, Hillsdale, New Jersey, 1978.

[Posner 80] M.I. Posner. *Orienting of attention.* The Quarterly Journal of Experimental Psychology, vol. 32, no. 1, pages 3–25, 1980. www

[Posner 90] M I Posner & S E Petersen. *The attention system of the human brain.* Annual review of neuroscience, vol. 13, pages 25–42, January 1990. www

[Privitera 00] C.M. Privitera & L.W. Stark. *Algorithms for defining visual regions-of-interest : Comparison with eye fixations.* IEEE Transactions on pattern analysis and machine intelligence, vol. 22, no. 9, pages 970–982, 2000. www

[Rapantzikos 03] K Rapantzikos & N Tsapatsoulis. *On the implementation of visual attention architectures.* In Tales of the Disappearing Computer„ Santorini, 2003. www

[Rebhan 08] Sven Rebhan, F. Röhrbein, Julian Eggert & E. Körner. *Attention modulation using short-and long-term knowledge.* In 6th International Conference (ICVS), pages 151–160, Santorini, Greece, 2008. Springer. www

[Rempulski 09] Nicolas Rempulski, Armelle Prigent, Pascal Estraillier, Vincent Courboulay & Matthieu Perreira Da Silva. *Adaptive Storytelling Based On Model-Checking Approaches*. International Journal of Intelligent Games and Simulation (IJIGS), vol. 5, no. 2, pages 33—-41, 2009. www

[Rensink 97] Ronald A Rensink, J Kevin O Regan & James J Clark. *To See or Not to See : The Need for Attention to Perceive Changes in Scenes*. Psychological Science, vol. 8, no. 5, pages 1–6, 1997.

[Rensink 00] R A Rensink. *The dynamic representation of scenes*. Visual Cognition, vol. 7, pages 17–42, 2000.

[Rigau 05] J Rigau, M Feixas & M Sbert. *An information-theoretic framework for image complexity*. Computational Aesthetics 2005, page 177, 2005. www

[Rissanen 78] J. Rissanen. *Modeling by shortest data description*. Automatica, vol. 14, pages 465–471, 1978.

[Rizzolatti 87] G Rizzolatti, L Riggio, I Dascola & C Umiltá. *Reorienting attention across the horizontal and vertical meridians : evidence in favor of a premotor theory of attention.* Neuropsychologia, vol. 25, no. 1A, pages 31–40, January 1987. www

[Robinson 92] D Robinson. *The pulvinar and visual salience*. Trends in Neurosciences, vol. 15, no. 4, pages 127–132, April 1992. www

[Rodieck 03] Robert W. Rodieck. La vision. De Boeck, Bruxelles, neuroscien edition, 2003.

[Roelfsema 06] Pieter R Roelfsema. *Cortical algorithms for perceptual grouping.* Annual review of neuroscience, vol. 29, no. March, pages 203–27, 2006. www

[Rolls 06] Edmund T Rolls & Simon M Stringer. *Invariant visual object recognition : a model, with lighting invariance.* Journal of physiology, Paris, vol. 100, no. 1-3, pages 43–62, 2006. www

[Rosenthal 99] Victor Rosenthal & Y.M. Visetti. *Sens et temps de la Gestalt*. Intellectica, vol. 28, pages 147–227, 1999. www

[Rybak 98] I a Rybak, V I Gusakova, A V Golovan, L N Podladchikova & N A Shevtsova. *A model of attention-guided visual perception and recognition.* Vision research, vol. 38, no. 15-16, pages 2387–400, August 1998. www

[Sabatini 03] S. Sabatini & Fabio Solari. *An Early Cognitive Approach to Visual Motion Analysis*. AI*IA 2003 : Advances in Artificial Intelligence, pages 385–397, 2003. www

[Salvucci 00] Dario D. Salvucci & Joseph H. Goldberg. *Identifying fixations and saccades in eye-tracking protocols.* Proceedings of the symposium on Eye tracking research & applications - ETRA '00, pages 71–78, 2000. www

[Sasaki 07] Yuka Sasaki. *Processing local signals into global patterns.* Current opinion in neurobiology, vol. 17, no. 2, pages 132–9, 2007. www

[Sayood 00] Khalid Sayood. Predictive coding, chapitre 6, pages 139–177. Morgan Kaufmann Publishers, 2nd edition, 2000.

[Schneider 77] W Schneider & R M Shiffrin. *Controlled and automatic human information processing : I. Detection, search, and attention.* Psychological Review, vol. 84, pages 1–66, 1977.

[Sela 97] Gal Sela & Martin D Levine. *Real-time attention for robotic vision.* Real-Time Imaging, vol. 3, pages 173–194, 1997.

[Shic 06] Frederick Shic & Brian Scassellati. *A Behavioral Analysis of Computational Models of Visual Attention.* International Journal of Computer Vision, vol. 73, no. 2, pages 159–177, September 2006. www

[Siagian 07] Christian Siagian & Laurent Itti. *Rapid biologically-inspired scene classification using features shared with visual attention.* IEEE transactions on pattern analysis and machine intelligence, vol. 29, no. 2, pages 300–12, February 2007. www

[Snodgrass 80] J G Snodgrass & M Vanderwart. *A standardized set of 260 pictures : norms for name agreement, image agreement, familiarity, and visual complexity.* Journal of experimental psychology. Human learning and memory, vol. 6, no. 2, pages 174–215, March 1980. www

[Spratling 04] M W Spratling & M H Johnson. *A feedback model of visual attention.* Journal of cognitive neuroscience, vol. 16, no. 2, pages 219–37, March 2004. www

[Stringer 00] S M Stringer & E T Rolls. *Position invariant recognition in the visual system with cluttered environments.* Neural networks : the official journal of the International Neural Network Society, vol. 13, no. 3, pages 305–15, April 2000. www

[Styles 06] Elizabeth A Styles. The Psychology of Attention. Psychology Press, New York, NY, 2nd editio edition, 2006.

[Sun 03] Y Sun & R Fisher. *Object-based visual attention for computer vision.* Artificial Intelligence, vol. 146, no. 1, pages 77–123, 2003. www

[Sun 08] Y Sun, R Fisher, F Wang & H Gomes. *A computer vision model for visual-object-based attention and eye movements.* Computer Vision and Image Understanding, vol. 112, no. 2, pages 126–142, November 2008. www

[Tatler 05] Benjamin W Tatler, Roland J Baddeley & Iain D Gilchrist. *Visual correlates of fixation selection : effects of scale and time.* Vision research, vol. 45, no. 5, pages 643–59, March 2005. www

[Tatler 07] Benjamin W Tatler. *The central fixation bias in scene viewing : Selecting an optimal viewing position independently of motor biases and image feature distributions.* Journal of Vision, vol. 7, pages 1–17, 2007.

[Thompson 05] Kirk G. Thompson & Narcisse P. Bichot. A visual salience map in the primate frontal eye field, volume 147, pages 251–262. Elsevier, 2005. www

[Toh 05] AM Toh, Roberto Togneri & Sven Nordholm. *Spectral entropy as speech features for speech recognition.* Proceedings of PEECS, no. 1, 2005. www

[Tollari 05a] Sabrina Tollari & Hervé Glotin. *Sélection adaptative des descripteurs visuels et dérivation de métadescripteurs contextuels dépendant du mot-clé pour l'indexation automatique d'images.* In Atelier MetSI, 2005. www

[Tollari 05b] Sabrina Tollari, Hervé Glotin & Jacques Le Maitre. *Enhancement of Textual Images Classification using Segmented Visual Contents for Image Search Engine.* Multimedia Tools and Applications, vol. 25, no. 3, pages 407–415, 2005.

[Torralba 06] Antonio Torralba, Aude Oliva, Monica S Castelhano & John M Henderson. *Contextual guidance of eye movements and attention in real-world scenes : the role of global features in object search.* Psychological review, vol. 113, no. 4, pages 766–86, October 2006. www

[Treisman 60] Anne Treisman. *Contextual cues in selective listening.* Quarterly Journal of Experimental Psychology, vol. 12, pages 242–248, 1960.

[Treisman 69] Anne Treisman. *Strategies and models of selective attention.* Psychological Review, vol. 76, pages 282–299, 1969.

[Treisman 80] Anne Treisman & Garry Gelade. *A Feature-Integration Theory of Attention.* Cognitive Psychology, vol. 136, no. 12, pages 97–136, 1980.

[Tsotsos 90] John.K. Tsotsos. *Analysing vision at the complexity level.* Behavioral. and. Brain. Sciences., vol. 13, pages 423–469, 1990.

[Tsotsos 95] J.K. Tsotsos, S.M. Culhane, W.Y. Kei Wai, Y. Lai, Neal Davis & F. Nuflo. *Modeling visual attention via selective tuning.* Artificial intelligence, vol. 78, no. 1-2, pages 507–545, 1995. www

[Tsotsos 05a] J.K. Tsotsos, Laurent Itti & G. Rees. A Brief and Selective History of Attention. Elsevier Press, 2005.

[Tsotsos 05b] J.K. Tsotsos, Y Liu, J Martineztrujillo, M Pomplun, E Simine & K Zhou. *Attending to visual motion.* Computer Vision and Image Understanding, vol. 100, no. 1-2, pages 3–40, 2005. www

[Tsotsos 07] John K Tsotsos. *A selective History of Visual Attention*, 2007.

[van der Heijden 97] a H van der Heijden & S Bem. *Successive approximations to an adequate model of attention.* Consciousness and cognition, vol. 6, no. 2-3, pages 413–28, 1997. www

[Van Rullen 05] R Van Rullen & Christof Koch. Visual Attention and Visual Awareness, volume 91125, chapitre 3, pages 65–83. Elsevier, handbook o edition, 2005. www

[VanRullen 03] Rufin VanRullen. *Visual saliency and spike timing in the ventral visual pathway.* Journal of physiology, Paris, vol. 97, no. 2-3, pages 365–77, 2003. www

[VanRullen 04] R. VanRullen & SJ Thorpe. Perception, decision, attention visuelles : ce que les potentiels evoques nous apprennent sur le fonctionnement du systeme visuel, chapitre 5, pages 95–121. Lavoisier, Paris, France, hermes edition, 2004. www

[Vieira Neto 07] H Vieira Neto & Ulrich Nehmzow. *Visual novelty detection with automatic scale selection.* Robotics and Autonomous Systems, vol. 55, no. 9, pages 693–701, 2007. www

[Vieville 05] Thierry Vieville. *Quelques Idées sur le Concept d'Adaptation en Vision par Ordinateur*, 2005. `http://www.loria.fr/~vthierry/cours/adaptation.html`

[Vijayakumar 01] S. Vijayakumar, J. Conradt, T. Shibata & S. Schaal. *Overt visual attention for a humanoid robot.* In Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems, numéro Iros, pages 2332–2337. Ieee, 2001. www

[Viola 02] Paul Viola & Michael Jones. *Robust real-time object detection.* International Journal of Computer Vision, vol. 57, no. 2, pages 137–154, 2002. www

[Vitay 05] Julien Vitay, N.P. Rougier & F. Alexandre. A distributed model of spatial visual attention, pages 54–72. Springer, 2005. www

[Volterra 28] Vito Volterra. *Variations and fluctuations of the number of individuals in animal species living together.* ICES Journal of Marine Science, vol. 3, no. 1, pages 3–51, 1928.

[Walther 05] D Walther. *Selective visual attention enables learning and recognition of multiple objects in cluttered scenes.* Computer Vision and Image Understanding, vol. 100, no. 1-2, pages 41–63, October 2005.

[Walther 06a] Dirk Walther. *Interactions of Visual Attention and Object Recognition : Computational Modeling, Algorithms, and Psychophysics.* Phd, California Institue of Technology, 2006.

[Walther 06b] Dirk Walther & Christof Koch. *Modeling attention to salient proto-objects.* Neural networks : the official journal of the International Neural Network Society, vol. 19, no. 9, pages 1395–407, 2006. www

[Wolfe 89] J M Wolfe, K R Cave & S L Franzel. *Guided search : an alternative to the feature integration model for visual search.* Journal of experimental psychology. Human perception and performance, vol. 15, no. 3, pages 419–33, August 1989. www

[Wolfe 00] J.M. Wolfe. Visual attention, volume 5, pages 335–386. Academic Press, San Diego, CA, 2nd edition, 2000.

[Wolfe 04] J.M. Wolfe & T.S. Horowitz. *What attributes guide the deployment of visual attention and how do they do it ?* Nature Reviews Neuroscience, vol. 5, no. 6, pages 495–501, 2004. www

[Wooding 02] David S Wooding. *Eye movements of large populations : II. Deriving regions of interest, coverage, and similarity using fixation maps.* Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc, vol. 34, no. 4, pages 518–28, November 2002. www

[Wörgötter 04] F. Wörgötter, N. Krüger, Nicolas Pugeault, Dirk Calow, Markus Lappe, Karl Pauwels, M. Van Hulle, Sovira Tan & Alan Johnston. *Early cognitive vision : Using Gestalt-laws for task-dependent, active image-processing.* Natural computing, vol. 3, no. 3, pages 293–321, 2004. www

[Yaghmaee 10] Farzin Yaghmaee & Mansour Jamzad. *Estimating Watermarking Capacity in Gray scale Images based on Image Complexity.* EURASIP Journal on Advances in Signal Processing, vol. 2010, page Article ID 851920, 2010. www

[Yarbus 67] Alfred L. Yarbus. Eye Movements and Vision. Plenum Press, new York, 1967.

[Yu 07] Zhiwen Yu & Hau-San Wong. *A Rule Based Technique for Extraction of Visual Attention Regions Based on Real-Time Clustering.* IEEE Transactions on Multimedia, vol. 9, no. 4, pages 766–784, June 2007. www

[Zach 07] C Zach, T Pock & H Bischof. *A duality based approach for real-time TV-L 1 optical flow*. In Proceedings of the 29th DAGM conference on Pattern recognition, volume 1, pages 214–223. Springer-Verlag, 2007. www

[Zhang 05] Long-fei Zhang, Yuan-da Cao, Ming-jie Zhang & Yi-zhuo Wang. *Object Tracking Based on Visual Attention Model and Particle Filter*. Journal of Information Technology, vol. 11, no. 9, pages 109–118, 2005.

[Zhang 08] Jing Zhang, Li Zhuo & Lansun Shen. *Regions of Interest extraction based on visual attention model and watershed segmentation*. In 2008 International Conference on Neural Networks and Signal Processing, pages 375–378. Ieee, June 2008. www

[Zilberstein 96] Shlomo Zilberstein. *Using anytime algorithms in intelligent systems*. AI Magazine, vol. 17, no. 3, pages 73–83, 1996.

[Ziv 77] J. Ziv & a. Lempel. *A universal algorithm for sequential data compression*. IEEE Transactions on Information Theory, vol. 23, no. 3, pages 337–343, May 1977. www

[Zou 05] Qi Zou, Siwei Luo & Jianyu Li. *Selective attention guided perceptual grouping model*. Advances in Natural Computation, pages 867–876, 2005. www