École Centrale de Nantes Université de Nantes École des Mines de Nantes

ÉCOLE DOCTORALE S T I M
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DES MATÉRIAUX

Année 2005

Thèse de **DOCTORAT**

Spécialité : traitement du signal et des images

présentée et soutenue publiquement par

Olivier LE MEUR

Le 24 octobre 2005.
à l'École polytechnique de l'Université de Nantes

ATTENTION SELECTIVE EN VISUALISATION D'IMAGES FIXES ET ANIMEES AFFICHEES SUR ECRAN : MODELES ET EVALUATION DE PERFORMANCES - APPLICATIONS

Jury :

Président :	M. Joseph RONSIN	Professeur des Universités, INSA, Rennes
Rapporteurs :	M. Jeanny HERAULT	Professeur des Universités, INPG, Grenoble
	M. Claude LABIT	Directeur de recherche, INRIA, IRISA, Rennes
Examinateurs :	M. Dominique BARBA	Professeur des Universités, Polytech, Nantes
	M. Atilla BASKURT	Professeur des Universités, INSA, Rennes
	M. Patrick LE CALLET	Maître de Conférences, Polytech, Nantes
Membres invités :	M. Philippe GUILLOTEL	Ingénieur, THOMSON R&D, Rennes
	M. Laurent ITTI	Assistant Professor, USC, Los Angeles

Directeur de thèse : Dominique BARBA
Co-encadrant : Patrick LE CALLET
Laboratoire : IRCCyN
Adresse : École Polytechnique de l'Université de Nantes (EPUN)
Rue Christian PAUC BP 50609
44306 NANTES Cedex 3

N^o ED 366-217

A Karen, Glen et Louen

Remerciements

Les travaux de recherche rapportés dans ce mémoire sont le résultat d'une collaboration entre deux laboratoires de recherche. Le premier est le laboratoire dédié à la compression vidéo (*VCL, Video Compression Laboratory*) de Thomson R&D Rennes, dirigé par Michel Kerdranvat. Le second est le laboratoire universitaire d'accueil, image et communication vidéo (*IVC, Image and Video Communication*), de l'IRCCyN (Institut de Recherche en Communications et en Cybernétique de Nantes) auquel appartiennent Dominique Barba, directeur de cette thèse, et Patrick Le Callet, co-encadrant.

Je tiens à remercier tout d'abord et plus particulièrement ces trois personnes qui m'ont fait confiance pour ce travail de thèse.

Je remercie Joseph Ronsin, Professeur des Universités à l'INSA de Rennes, qui m'a fait l'honneur de présider ce jury.

Je remercie Jeanny Hérault, Professeur des Universités à l'INPG de Grenoble et Claude Labit, Directeur de recherche, INRIA, IRISA, Rennes, d'avoir bien voulu accepter la charge de rapporteur.

Je remercie Atilla Baskurt, Professeur des Universités à l'INSA de Lyon et Philippe Guillotel, ingénieur à Thomson R&D Rennes, d'avoir bien voulu juger ce travail. Sans oublier, Laurent Itti, Assistant Professor à l'USC de Los Angeles, qui malgré les contraintes temporelles et géographiques a accepté notre invitation.

Encore merci à Dominique et Patrick pour m'avoir conseillé et guidé avec attention et rigueur tout au long de mes travaux.

Je tiens également à exprimer ma reconnaissance aux membres des laboratoires VCL et IVC pour leur sympathie et leur contribution à ce travail. Un merci tout particulier à Dominique Thoreau qui m'a accompagné durant ces trois années.

Enfin, j'aimerais adresser un remerciement plus personnalisé à Stef et à Karen. Le premier est à considérer comme un correcteur grammatical et orthographique quasi-temps réel de mes contributions scientifiques en langue anglaise. Merci pour ton avis lapidaire sur la qualité initiale de certains passages de mes écrits ! La seconde, Karen, pour m'avoir apporté soutien et réconfort tout au long de ces trois années.

Table des matières

Table des figures

Introduction générale

L'attention visuelle

Avec ses 100 milliards de neurones, le cerveau est un organe extrêmement complexe. Son fonctionnement partage depuis des décennies la communauté scientifique : faut-il considérer le cerveau comme un ordinateur très centralisé ou comme ayant un fonctionnement global ? La notion d'ordinateur centralisé est actuellement controversée puisqu'il ne semble pas être possible de déterminer l'endroit où *se cache ce "centre", le processeur central qui séquence les traitements, les tâches*[1]... Aujourd'hui, c'est la deuxième hypothèse, fortement soutenue par F. Varela [Varela 98], qui semble être la plus idoine.
En marge de ce débat, le squelette cérébrale est à peu près bien identifié. De nombreux travaux, issus de la neurophysiologie, ont en effet permis d'identifier via des expériences judicieusement conçues des groupes de neurones fonctionnant localement en bandes, en colonnes, échangeant des informations... Ces travaux ont largement contribué à décrire le fonctionnement du système visuel, permettant à la fois de construire la topologie du cerveau, c'est à dire la localisation des différentes aires cérébrales, et d'identifier le fonctionnement intrinsèque de la plupart des cellules visuelles.
Bien que nous n'en ayons pas conscience, du fait de la très grande efficacité de traitement de notre système visuel, démontrée dans toutes nos activités quotidiennes, nos ressources sensorielles (ou mécanique) de traitement sont biologiquement limitées. Pour faire face à l'énorme quantité d'informations visuelles de notre environnement visuel, le système visuel possède la faculté de sélectionner une information pertinente localisée spatialement dans le champ visuel parmi toutes celles qui lui parviennent : on parle d'attention visuelle.

L'attention visuelle, intimement liée au mouvement oculaire, est souvent assimilée à la répétition de phases de focalisation et de sélection. La focalisation permet de concentrer les capacités de traitement et d'analyse sur une zone particulière du champ visuel. La localisation et la détection de la zone inspectée sont effectuées lors de l'étape de sélection, dans laquelle deux mécanismes interviennent : un mécanisme automatique (appelé encore exogène ou *Bottom-Up*) et un mécanisme contrôlé (appelé encore endogène ou *Top-Down*). Le premier est un traitement automatique très rapide réalisé inconsciemment. Il autorise le traitement précoce de certaines informations permettant ainsi la réalisation de tâches complexes sans solliciter les ressources attentionnelles. Le deuxième est un processus contrôlé nécessitant un effort cognitif important et nécessitant la quasi totalité

[1] Extrait de [Varela 99]

des ressources attentionnelles. Ce mécanisme est déployé lorsqu'une tâche particulière doit être effectuée : reconnaître un lieu sur une photo, chercher l'homme portant une casquette verte par exemple...
Du fait de la grande complexité des mécanismes et des inter actions, des inter dépendances existantes entre ces mécanismes, modéliser l'attention visuelle dans son ensemble est une véritable gageure. Une voie réaliste est de modéliser l'attention visuelle pré-attentive, c'est à dire celle liée au processus automatique. Contrairement au traitement contrôlé, l'aspect cognitif est ici très faible laissant alors la place à des traitements plus mécaniques et donc plus facilement reproductibles. L'enjeu est considérable aussi bien d'un point de vue intellectuel qu'économique. Intellectuel, car les difficultés inhérentes à la modélisation et à l'évaluation des performances sont loin d'être triviales. L'intérêt économique est également important puisque de nombreuses applications peuvent bénéficier de la modélisation de l'attention visuelle.

C'est pourquoi, durant les vingt cinq dernières années, les recherches dans ce domaine se sont intensifiées. Dans un cadre de modélisation, les recherches les plus influentes sont sans conteste celles de A. Treisman et G. Gelade [Treisman 80], relatives à la théorie de l'intégration des caractéristiques visuelles (*Feature Integration Theory*, FIT). Dans l'article [Treisman 80], le concept présenté repose sur l'existence supposée dans le cerveau de cartes donnant les positions spatiales les plus importantes d'un point vue attentionnel et cela pour un certain nombre de caractéristiques visuelles. Ces différentes cartes forment ensuite une carte finale (*master map*) donnant les lieux perceptuellement importants. Plus tard, en 1985, C. Koch et S. Ullman [Koch 85] présentèrent l'architecture de base d'un modèle d'attention visuelle d'inspiration biologique, basée sur les travaux de A. Treisman et G. Gelade. Le terme carte de saillance apparaît pour la première fois. Cette carte est définie par les auteurs comme une représentation de l'environnement accentuant les régions d'intérêt du champ visuel.
La théorie de l'intégration des caractéristiques visuelles et l'architecture proposée par C. Koch et S. Ullman ont joué un rôle de catalyseur dans la modélisation de l'attention visuelle. De nombreux modèles ont alors vu le jour. La taxinomie des modèles fait apparaître deux catégories : la première concerne les modèles purement basés sur des algorithmes et des outils provenant du traitement d'images. Le terme "vision computationnelle" est utilisé pour qualifier ce type de modèles, où les propriétés fonctionnelles du système visuel ne sont pas ou peu considérées. La modélisation proposée par W. Osberger et A. Maeder [Osberger 98] est l'exemple le plus démonstratif. En effet, les auteurs utilisent des propriétés du système visuel de haut niveau pour construire une carte localisant les régions d'intérêt visuel. La seconde catégorie, quant à elle, concerne les modèles d'inspiration biologique. La conception de ces modèles se base sur la modélisation des propriétés fonctionnelles du système visuel, formant un point de convergence entre plusieurs disciplines telles que la neurobiologie, les sciences cognitives, la psychophysique, le traitement d'images... Le premier modèle ayant réussi à faire le lien entre les différentes communautés est celui de L. Itti [Itti 98], issu de ses travaux de thèse. Des résultats très intéressants ont été obtenu sur des scènes naturelles. De nombreuses études se sont ensuite inspirées de ces travaux, tant dans l'amélioration de la modélisation que dans l'utilisation de ses cartes de saillance pour améliorer des applications de codage, de recherche de contenu...

Modélisation et applications

Dans cette thèse, nous nous intéressons à la modélisation de l'attention visuelle pré-attentive. Le modèle envisagé, d'inspiration biologique, doit être capable de déterminer les zones visuellement importantes d'une image ou d'une séquence d'images projetées sur un écran. Afin de répondre aux problèmes rencontrés par les modèles de l'état de l'art, nous nous proposons d'étudier plusieurs points.
Le premier est de montrer qu'il est possible d'extraire les caractéristiques visuelles précoces de façon cohérente. En effet, les modèles actuels sont tous confrontés à la comparaison de données provenant de différentes dimensions visuelles et ayant des dynamiques différentes. Peut-on comparer les valeurs de caractéristiques visuelles provenant de différentes dimensions ? Actuellement, la seule solution est de normaliser dans la même dynamique toutes les données. La philosophie envisagée est ici différente : il s'agit en effet de construire un espace psycho-visuel dans lequel toutes les données visuelles s'expriment en terme de visibilité. La modélisation des propriétés fonctionnelles de la rétine est alors nécessaire pour hiérarchiser les données suivant leur degré de visibilité. La base de la construction de l'espace psycho-visuel s'appuie sur les précédents travaux de thèse de P. Le Callet [Callet 01], L. Bedat [Bedat 98] et H. Sénane [Sénane 96] effectués au sein de l'équipe Image et Vidéo-Communications de l'IRCCyN.
La seconde question concerne la transformation des données exprimées en visibilité en données exprimées en saillance. La transformation doit permettre de définir des traitements d'inspiration biologique permettant d'identifier les zones visuellement importantes ou qui "sautent aux yeux".
Enfin, le dernier problème soulevé par les modèles de l'état de l'art concerne la fusion d'informations de saillance de haut niveau provenant de diverses dimensions. Les concepts de fusion proposés par R. Milanese [Milanese 93] sont repris pour construire une fusion cohérente des informations de saillance provenant des dimensions visuelles achromatique, chromatiques et temporelle.

Outre les problèmes inhérents à la conception des modèles actuels, l'évaluation des performances est relativement mal abordée. Dans cette étude, il ne s'agit pas d'étudier le chemin visuel emprunté par un observateur pour explorer le contenu d'une scène. L'évaluation consiste plutôt à mesurer la capacité du modèle à reproduire le comportement d'un observateur moyen, c'est à dire d'un large panel d'observateurs. Néanmoins, compte tenu de l'idiosyncrasie de la stratégie visuelle, c'est à dire que chaque observateur présente une stratégie visuelle qui lui est propre, le problème de la variabilité de la stratégie visuelle inter observateurs se pose. Il est donc intéressant de déterminer cette variabilité inter observateurs pour pouvoir se faire une opinion sur la pertinence des résultats de la modélisation. Une forte dispersion rend la modélisation du comportement d'un observateur moyen difficile. Par contre, une faible dispersion inter observateur dénote l'existence de zones attirant l'attention visuelle.
Pour rendre "robuste" l'évaluation des performances, plusieurs métriques complémentaires doivent être utilisées. Par ailleurs, une comparaison avec le modèle de L. Itti est effectuée sur les images fixes.

L'utilisation d'un tel modèle dans le traitement d'images peut être très bénéfique pour

de nombreuses applications. Par exemple, le tatouage numérique d'informations peut être piloté par ce type de modèle. Dans un contexte de tatouage invisible, l'idée consiste alors à tatouer les zones inintéressantes visuellement. L'estimation de qualité peut également voire ses performances améliorées puisqu'un artefact apparaissant sur une zone saillante est certainement plus gênant qu'un artefact apparaîssant sur une zone n'attirant pas l'attention.
Dans cette étude, nous proposons de décrire et d'évaluer les performances d'un codage vidéo piloté par une carte de saillance spatio-temporelle. Une seconde application concerne la création d'images ou de séquences d'images miniatures. Il s'agit de construire une image miniature centrée sur les zones les plus saillantes.

Organisation du mémoire

Ce mémoire est organisé en trois parties.

La première partie traite de l'anatomie du SVH (Système Visuel Humain) et de la modélisation de l'attention visuelle. Il s'agit en fait d'un état de l'art orienté vers la problématique traitée.
La deuxième partie présente notre contribution en terme de modélisation spatiale et spatio-temporelle de l'attention visuelle pré-attentive. Elle comprend également l'évaluation des performances des modèles proposés ainsi que la description des expérimentations oculométriques réalisées.
La dernière partie est consacrée à deux applications dont les performances peuvent être substantiellement améliorées par l'utilisation du modèle proposé. Une application de codage et une application de création d'images miniatures sont détaillées.

Première partie

Anatomie et modélisation du système visuel humain

Chapitre 1

Biologie du système visuel humain et l'attention visuelle

1.1 Introduction

Dans le cadre de la modélisation d'un système biologique, en l'occurrence celui lié à notre perception visuelle, la connaissance des étapes biologiques du traitement de l'information est primordiale. Cette compréhension du fonctionnement de notre système biologique est parfois délaissée au détriment d'approches plus pragmatiques, éloignées de la biologie visuelle. Pour notre étude, nous souhaitons pouvoir justifier les composantes de la modélisation en faisant un parallèle avec le système visuel humain et ainsi pouvoir argumenter nos choix, dans la limite des connaissances actuelles auxquelles nous avons pu avoir accès.
Ainsi, dans ce premier chapitre, les mécanismes et le fonctionnement du système visuel humain font l'objet d'une étude détaillée. En premier lieu, l'organisation biologique de l'oeil est décrite à partir de données neurophysiologiques. La neurophysiologie concerne l'étude des composantes et des mécanismes cérébraux, et notamment ceux de la vision. Bien qu'imparfaites, ces connaissances fournissent une idée de la structure et de l'organisation cérébrale : structure du cortex visuel en plusieurs aires spécialisées dans un certain type de traitement, mise en évidence de plusieurs voies opérant en parallèle sur des dimensions différentes (couleur, mouvement...)...
Ensuite, les propriétés dites de haut niveau sont abordées mettant ainsi l'accent sur la complexité et la difficulté de modélisation des mécanismes cérébraux mis en jeux.
Enfin, le troisième chapitre est consacré à la description de l'attention visuelle sélective, primordiale pour comprendre et appréhender notre environnement. A cause de notre limitation biologique (nos ressources sensorielles sont intrinsèquement limitées), nous ne pouvons pas traiter toutes les informations de notre environnement visuel. Des mécanismes de sélection ont donc été développés afin de pallier ce problème : les premiers, les mécanismes dits passifs, sont liés au fonctionnement de nos cellules et à leur organisation. A titre d'exemple, les cellules photoréceptrices compressent l'information visuelle d'un facteur 130 :1 avant de la transmettre aux aires corticales. En dépit de ce fort taux de compression, la quantité d'informations à traiter reste considérable. C'est pourquoi, des mécanismes de sélection actifs permettant de concentrer nos ressources sur des zones intéressantes de notre

environnement visuel sont nécessaires et incontournables. Ces mécanismes de sélection actifs sont intimement liés au mouvement oculaire que nous détaillons également. Le dernier chapitre, en guise de conclusion, tente de faire lien entre les propriétés des cellules visuelles et les mécanismes de l'attention visuelle permettant d'avoir une compréhension globale du système visuel.

1.2 Biologie et neurophysiologie du système visuel humain

La biologie et la neurophysiologie du SVH ont fait l'objet de nombreux articles et de livres. Afin de ne pas multiplier les références, le lecteur pourra se référer aux livres de P. Kowaliski [Kowaliski 90] publié en 1990 et de J. Hérault [Hérault 01] publié en 2001 faisant appel chacun à de nombreuses références.

1.2.1 La structure de l'oeil

L'oeil est l'organe de la vision ; il est d'ailleurs souvent comparé à une fenêtre sur notre cerveau. Il est de faible volume (environ de 6.5 cm^3), pèse environ 7 grammes et a la forme d'une sphère d'environ 24 mm de diamètre, complétée vers l'avant par une autre demi-sphère de 8 mm de rayon, la cornée.

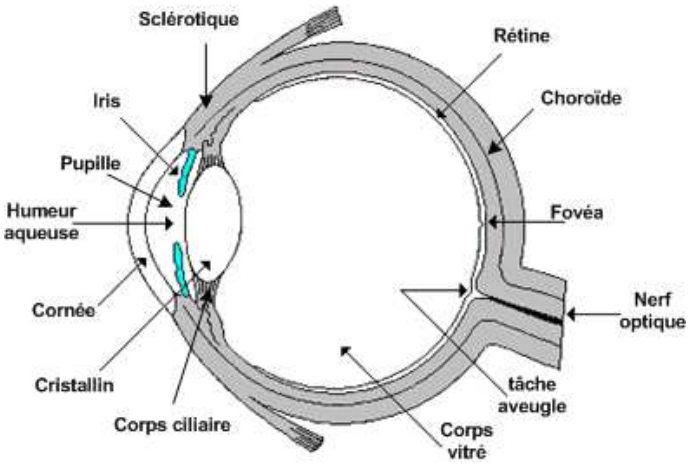

FIG. 1.1: Anatomie de l'oeil.

La paroi du globe oculaire est formée de 3 tuniques détaillées ci-après. La figure 1.1 sert de support permettant d'expliquer la composition de ces différentes tuniques :

- la tunique fibreuse, externe, se compose de la sclérotique opaque en arrière et de la cornée transparente en avant. La sclérotique forme ce qu'on appelle couramment le blanc de l'oeil. Elle est traversée par un grand nombre de petits canaux sanguins et à l'arrière par une ouverture appelée tâche aveugle où passent les fibres du nerf optique. Comme son nom l'indique, au niveau de la tâche aveugle, l'information visuelle incidente est perdue. Une expérience simple et ludique représentée à la figure 1.2 permet de mettre en exergue cette propriété biologique. Afin de conclure sur cette

description sommaire de la tunique fibreuse, citons la cornée qui constitue la lentille principale du système optique oculaire ;
- la tunique uvéale, dite aussi uvée, se compose de trois éléments : l'iris en avant, le corps ciliaire et la choroïde en arrière. L'iris est la portion la plus antérieure de l'uvée, pigmentée, donnant sa couleur à l'oeil et percée d'un trou, la pupille. Le corps ciliaire est un anneau de tissu musculaire qui produit une substance liquide appelée humeur aqueuse. Enfin, la choroïde est le tissu nourricier de l'oeil : il apporte oxygène et nutriments dont les cellules ont besoin pour leur métabolisme ;
- la tunique nerveuse se compose de la rétine, lieu de transformation du signal lumineux en signaux nerveux, décrite ultérieurement.

Ces tuniques enferment des milieux transparents, indispensables à la vision. Ces milieux transparents sont au nombre de trois et forment une lentille convergente :
- l'humeur aqueuse est un liquide transparent qui remplit l'espace ente la cornée et le cristallin. Ce liquide est continuellement renouvelé et permet avec le corps vitré de maintenir la pression oculaire. Si il y a trop de pression oculaire, il y a une mauvaise irrigation sanguine de la papille, lieu d'émergence du nerf optique ;
- le cristallin est la lentille de l'oeil qui permet d'effectuer la mise au point par sa propriété essentielle de plasticité. Il permet, par ailleurs, de focaliser la lumière sur la rétine en modifiant ses courbures lors de l'accommodation. De forme biconvexe, flexible et transparent, il est situé à l'intérieur du globe oculaire ;
- le corps vitré est une masse gélatineuse et transparente, qui contient 90% d'eau et représente 60% du volume oculaire.

FIG. 1.2: Expérience simple de mise en évidence de la tâche aveugle. Cette expérience consiste à fermer l'oeil droit et à fixer la croix du haut à distance d'environ 30 *cm* avec l'oeil gauche. Dans ces conditions et en effectuant un léger mouvement ascendant de la tête, le cercle de gauche disparaît. De la même façon, si la croix du bas est fixée, la ligne du bas apparaît pleine (Extrait de [Kandel 91]).

1.2.2 La rétine

La rétine visuelle ou nerveuse est un tissu neuronal très fin d'une épaisseur de 0.1 à 0.5 *mm* tapissant le fond de l'oeil. C'est à ce niveau que s'effectue le premier traitement de l'information. Il consiste à traduire le message lumineux venant de l'extérieur en signaux nerveux utilisables et interprétables par les neurones des aires visuelles du cerveau. Cette

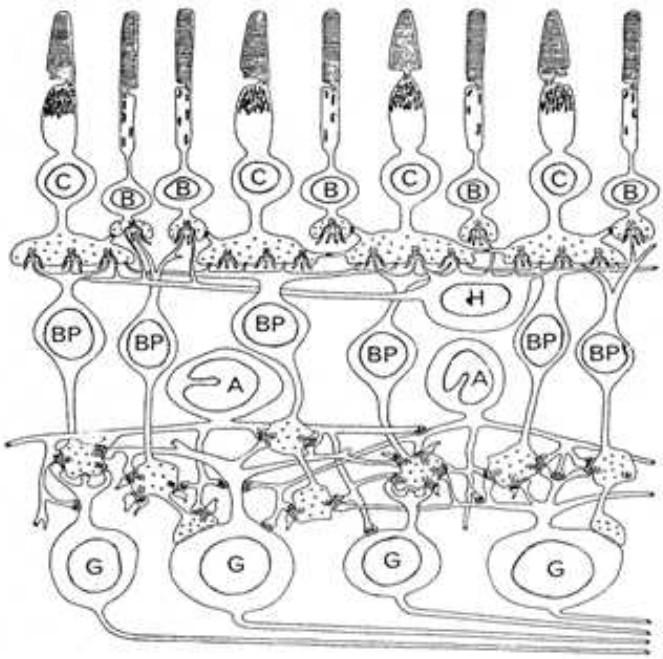

FIG. 1.3: Structure multicouches de la rétine : B bâtonnets, C cônes, H cellules horizontales, BP cellules bipolaires, A cellules amacrines et G cellules ganglionnaires (Extrait de [Kowaliski 90]).

traduction, ou plus précisément la fonction de transduction[1] de la rétine, ne s'opère pas sur toute la gamme de longueur d'ondes du signal incident mais simplement sur une faible bande de longueurs d'ondes, appelée le spectre visible, située dans l'intervalle $[400, 700]$ nanomètres.
La rétine est composée de plusieurs couches successives de neurones, comme le montre la figure 1.3. De la plus proche à la plus profonde, celles-ci correspondent successivement à la couche des ganglionnaires, la couche des cellules amacrines, la couche des cellules bipolaires, la couche des cellules horizontales et la couche des photorécepteurs.

1.2.2.1 La couche des photorécepteurs : les cônes et les bâtonnets

La couche la plus profonde par rapport à l'arrivée de l'information lumineuse est, paradoxalement, la couche des cellules photosensibles, aussi appelées photorécepteurs ou cellules sensorielles. Le signal lumineux doit en effet traverser l'ensemble des couches avant de pouvoir atteindre les photorécepteurs. Cette couche comporte environ 130 millions de cellules photosensibles, portant des noms caractéristiques de leur forme : bâtonnets et cônes.
Leur nombre n'est pas identique : le nombre de bâtonnets est, en effet, à peu près 20 fois supérieur à celui des cônes. Par ailleurs, les cônes se concentrent au centre de la rétine, appelé la fovéa, alors que les bâtonnets sont situés à la para fovéa et à la périphérie.

[1] La transduction est un processus de codage d'un signal lumineux en réponse électrique.

Comme l'indique la figure 1.4, la densité de bâtonnets augmente très rapidement lorsque la distance par rapport à la fovéa augmente. Un maximum est atteint pour une distance de 20 degrés par rapport à la fovéa.

Les cônes sont dédiés à la perception d'informations de moyennes à fortes luminances,

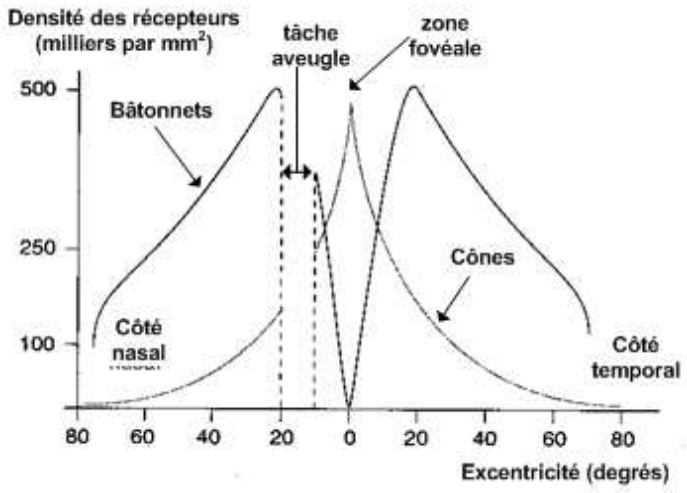

Fig. 1.4: Répartition des cellules photoréceptrices au sein de la rétine.

c'est à dire dans des conditions photopiques. Comparativement aux bâtonnets, les cônes permettent d'avoir une représentation fine d'une scène observée en conservant l'essentiel de sa résolution spatiale et temporelle ; l'acuité visuelle est élevée. Leur grande densité au centre de la rétine contribue à ce phénomène mais ce n'est pas la principale explication.

En effet, cette meilleure efficacité est causée par la façon dont l'information est distribuée par les cônes : contrairement aux bâtonnets qui distribuent l'information à plusieurs cellules réceptrices, les cônes sont reliés uniquement à une cellule, en l'occurrence une cellule bipolaire. Du point de vue de la cellule réceptrice, le système des cônes est dit non-convergent.

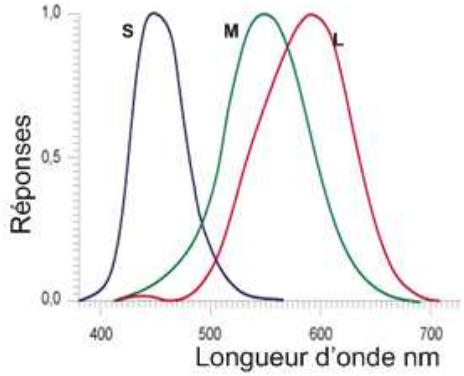

Fig. 1.5: Réponses normalisées des trois types de cônes présents au niveau de la rétine.

Des études portant sur les cônes ont permis de dégager trois grandes catégories caractérisées par des sensibilités différentes aux longueurs d'ondes : par ordre croissant de longueurs d'ondes, on trouve d'abord les cônes dits S (pour *small*) dont la sensibilité maximale est située autour de 420 *nm* (longueur d'onde proche de celle produisant la couleur bleue). Les deux autres catégories, notées M et L (pour respectivement *medium* et *large*), ont leur maximum de sensibilité autour de 531 *nm* (vert) et de 558 *nm* (rouge) respectivement. La figure 1.5 donne les réponses normalisées de ces trois types de cônes.

Les bâtonnets, quant à eux, sont dédiés à la perception d'informations basses luminances, c'est à dire dans des conditions scotopiques. Comme précédemment cité, les bâtonnets transmettent leurs réponses à plusieurs cellules. Ce maillage revient donc à un lissage de l'information incidente mettant en exergue la faible capacité des bâtonnets à restituer une bonne résolution spatiale et temporelle. L'intérêt de ces cellules ne réside donc pas dans la bonne restitution de l'information mais plutôt dans leur capacité de détection d'un évènement survenant en vision périphérique.
Le tableau 1.1 dresse la liste des différences et des caractéristiques majeures des cellules photoréceptrices.

Tab. 1.1: Différences et caractéristiques des cellules photoréceptrices, les cônes et les bâtonnets.

	Bâtonnets	Cônes
Répartition spatiale	périphérie rétinienne	fovéa
Population	100 millions	5 millions
Gamme de fonctionnement	scotopique	photopique
Résolution spatiale	faible	élevée
Résolution temporelle	faible	élevée
Capacité de détection en vision périphérique	élevée	faible
Information couleur	non	oui (cônes S,M,L)
Maillage	convergent	non-convergent

1.2.2.2 La couche des cellules horizontales

Les cellules horizontales, situées au plus proche des cellules photosensibles de la rétine, sont activées par des contacts synaptiques et par diffusion chimique provenant des photorécepteurs. L'activation se fait latéralement, c'est à dire que les cellules horizontales exploitent les signaux provenant d'un grand nombre de cellules photoréceptrices comme le montre la figure 1.6. Ces cellules jouent un rôle important puisqu'elles propagent latéralement le signal dans la rétine, modulant ainsi la réponse des photorécepteurs. Du fait qu'elles connectent plusieurs photorécepteurs, les cellules horizontales effectuent un lissage de l'information transmise par les cônes, conforme à la notion d'économie des systèmes biologiques [Hérault 01]. Par ailleurs, ces cellules influencent le comportement inhibiteur des cellules bipolaires. Ce phénomène que nous décrirons ultérieurement est communément

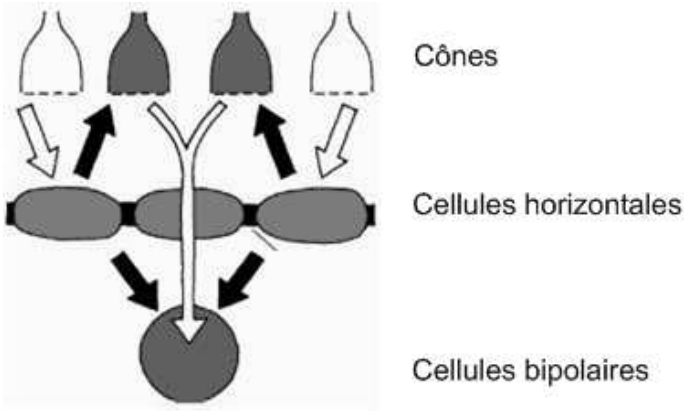

FIG. 1.6: Influence des cellules horizontales sur le comportement des cellules bipolaires et contre-réaction sur le comportement des cônes.

appelé interaction antagoniste du centre sur le pourtour des cellules bipolaires. Comme pour les cônes, il existe trois types de cellules horizontales présentant une préférence chromatique (bleu, vert, rouge) donnant naissance à des antagonismes chromatiques rouge-vert (les signaux des cônes M s'opposent à ceux des cônes L) et jaune-bleu (les signaux des cônes S s'opposent à la somme des signaux issus des cônes M et L).
Cette propriété très importante avait été mise en évidence par Ewald H. Hering en 1878. En effet, il introduisit l'idée de couleurs opposées et donc des antagonismes chromatiques à partir de simples expérimentations psychophysiques. Il nota que certaines nuances de couleurs ne sont jamais perçues simultanément. Une couleur n'est jamais décrite comme un rouge-vert ou un jaune-bleu. Il émit l'hypothèse qu'il existait des mécanismes opposés les uns aux autres.Ce n'est que bien plus tard que cette observation de mécanismes antagonistes fut confirmée [Valois 58], [Jameson 55].

1.2.2.3 La couche des cellules bipolaires

Les cellules bipolaires sont le moyen le plus direct pour véhiculer l'information des photorécepteurs aux cellules ganglionnaires, schématisant ainsi deux pôles (le premier lié aux cellules horizontales, le deuxième aux cellules ganglionnaires). Deux sortes de cellules bipolaires existent : celles qui relient directement des bâtonnets à une cellule ganglionnaire et celles qui relient un cône à une cellule ganglionnaire. Dans cette dernière catégorie, on trouve deux sous catégories : les cellules bipolaires ayant un centre dit ON et les cellules bipolaires ayant un centre dit OFF. Une cellule bipolaire de type ON répond maximalement lorsque le signal incident correspond à un spot de lumière entouré d'un pourtour sombre. Réciproquement, une cellule bipolaire OFF répond maximalement à un centre sombre sur un fond clair. L'information antagoniste du pourtour est véhiculée, comme nous l'avions évoqué dans le paragraphe précédent, par les cellules horizontales. Ce mécanisme est appelé mécanisme d'inhibition latérale.

1.2.2.4 La couche des cellules amacrines

Les cellules amacrines permettent d'effectuer l'interface entre les cellules bipolaires et ganglionnaires. Ces cellules, tout comme les cellules horizontales, jouent un grand rôle dans la propagation latérale de l'information. Par ailleurs, contrairement aux cellules bipolaires, elles sont très sensibles aux variations temporelles. Il existe plus de quarante types différents de cellules amacrines. Cette grande variété provient certainement de la différence de conception existant entre la rétine et les traitements post-rétiniens. Dans ces derniers, la prise en compte d'effets externes permettant de moduler la réponse est possible et largement utilisée. Par contre, au niveau de la rétine, les traitements sont plus mécaniques : les photorécepteurs ne reçoivent pas d'informations externes pouvant modifier leurs comportements. Le réseau des cellules amacrines permet de compenser ce manque d'informations [Wässle 91]. Ainsi, afin d'obtenir la meilleure perception, ces cellules permettent au système visuel de s'adapter à la composition spectrale de la scène. Par exemple, la composition spectrale de la lumière varie suivant les moments de la journée : le matin ainsi que le soir, le spectre est plutôt orienté sur le rouge alors qu'en milieu de journée, des longueurs d'ondes plus courtes sont majoritaires. Notre système visuel s'adapte naturellement à ces changements et compense ces variations grâce aux cellules amacrines.

1.2.2.5 la couche des ganglionnaires

Les cellules ganglionnaires constituent l'étage de sortie de la rétine. Elles reçoivent en entrée des informations provenant des cellules bipolaires et amacrines. Les réponses des cellules ganglionnaires, qui représentent la traduction de l'image perçue en message neuronal, sont ensuite délivrées aux aires visuelles supérieures.
Les premières expérimentations sur ces cellules ont été effectuées en 1938 par K. Hartline. En 1967, il reçut ainsi que R. Granit le prix Nobel de médecine pour le premier enregistrement électrique de réponses de cellules ganglionnaires à une stimulation lumineuse. C'est en 1953 que H. Barlow découvrit que les cellules ganglionnaires répondaient à une stimulation lumineuse d'une façon très particulière. Ces cellules répondent en effet de façon maximale lorsqu'une zone spécifique de la rétine est stimulée. Cette découverte a fait naître la notion fondamentale de *champ récepteur* qui représente la zone de la rétine à stimuler pour obtenir une réponse maximale. Cela signifie également que toutes stimulations effectuées en dehors du champ récepteur n'influent pas sur le comportement de la cellule visée.
Les champs récepteurs des cellules ganglionnaires présentent trois grandes caractéristiques :

- ils sont circulaires ;
- ils sont constitués de deux parties, un centre et un pourtour (un anneau entourant le centre). Lors d'une stimulation, ces deux parties délivrent des réponses antagonistes qui permettent soit d'annihiler soit d'amplifier la réponse finale ;
- il existe deux types de cellules ganglionnaires se différenciant par leur champ récepteur. On parle de cellules à centre ON (ayant un pourtour OFF) lorsque ces dernières répondent de façon maximale à une stimulation lumineuse du centre et d'un pourtour non stimulé. Les cellules ganglionnaires à centre OFF fournissent, quant à elles, une réponse maximale pour une stimulation lumineuse du pourtour

alors que le centre reste dans l'obscurité.

L'organisation antagoniste centre/pourtour des champs récepteurs des cellules ganglionnaires, très proches de celle des cellules bipolaires, leur confère une capacité de détection des contrastes d'informations : en effet, le centre et le pourtour se neutralisent lorsque la stimulation est uniforme sur le champ récepteur. Par contre, lorsque la stimulation est différente sur les deux parties du champ récepteur, la réponse finale est amplifiée par les deux contributions. L'oeil est donc plus sensible aux contrastes des informations qu'à leurs valeurs absolues.

Par ailleurs, nous verrons dans les chapitres suivants que cette organisation spatiale des champs récepteurs a conduit à définir *des fonctions de sensibilité aux contrastes*, plus connues sous le sigle *CSF* désignant *Contrast Sensitivity Function*. Ces CSFs modélisent la dépendance de la sensibilité visuelle en fonction de différents paramètres tels que la fréquence spatiale de la stimulation, son orientation, son intensité lumineuse ou en fonction de paramètres de plus haut niveaux tels que la distance d'observation...

Il existe deux principales catégories de cellules ganglionnaires se différenciant notamment par la taille de leurs champs récepteurs et par leurs projections dans le corps genouillé latéral, abrégé CGL (la description du CGL est effectuée dans le paragraphe 1.2.3.2). Relativement à la zone du CGL concernée, ces cellules sont notées P et M. Les cellules P (notées initialement Y suite à des expérimentations sur des chats), également appelées *midgets*, représentent 90% de la population des cellules ganglionnaires. Elles ont des champs récepteurs de taille réduite et encodent les détails fins d'une image ainsi que la plupart des informations chromatiques. Elles se situent en grande majorité au sein de la fovéa. Les cellules M (notée initialement X), appelées *parasols*, possèdent des champs récepteurs de grande taille ; elles sont insensibles à la couleur mais répondent très rapidement au mouvement. Le tableau 1.2 et la figure 1.7 résument et schématisent respectivement les principales caractéristiques de ces cellules.

Notons également que la différence des caractéristiques spatio-temporelles des cellules P et M a amené à introduire l'hypothèse de deux voies visuelles non indépendantes : la voie ventrale, "*quoi*" (ou *What* en anglais) associée à l'identification, à un processus de reconnaissance et la voie dorsale, "*où*" (ou *Where* en anglais) associée à la localisation d'un évènement. Nous retrouverons cette dichotomie lors de la présentation des mécanismes de l'attention visuelle.

Un mécanisme simplifié de coopération entre les deux voies peut être aisément décrit :

TAB. 1.2: Différences et caractéristiques des cellules ganglionnaires P et M.

	P	M
pourcentage de cellules	90%	10%
RF	petit	grand
Résolution spatiale	élevée	faible
Résolution temporelle	faible	élevée
Capacité de détection en vision périphérique	faible	élevée
Information couleur	oui	non

la localisation de l'information est calculée par la voie dorsale-où qui commande alors un

mouvement oculaire amenant l'objet au centre de la rétine pour une inspection fine via la voie ventrale-quoi.

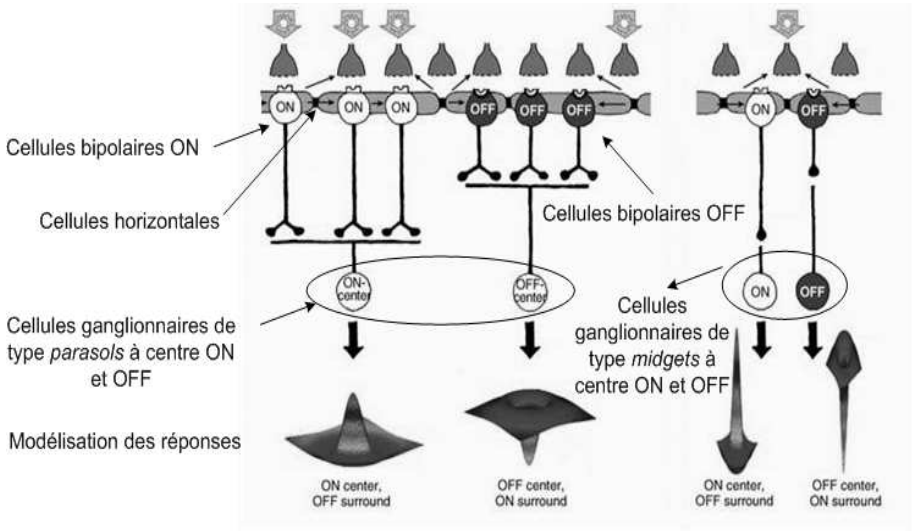

FIG. 1.7: Les différents types de cellules ganglionnaires midgets/parasols à centre ON ou OFF.

1.2.3 Les traitements post-rétiniens

1.2.3.1 De la rétine au cortex visuel primaire

Le chiasma optique, du grec khiasnos signifiant disposé en croix, est un organe qui permet d'aiguiller les informations collectées par la rétine. Ainsi, la sortie de chaque oeil est divisée en deux : les informations provenant d'une part de l'hémisphère temporal de l'oeil droit (respectivement gauche) et d'autre part de l'hémisphère nasal de l'oeil gauche (respectivement droit) se rejoignent sans se mélanger, pour être transmises au *corps genouillé latéral* droit (respectivement gauche). La figure 1.8 présente la façon dont les informations provenant de la rétine sont aiguillées vers les aires supérieures du cerveau. Ce signal stéréoscopique converge donc vers le corps genouillé latéral qui est une formation thalamique composée de six couches. Ces dernières peuvent être regroupées en fonction des réponses rétiniennes incidentes : quatre couches parvocellulaires recevant les réponses des

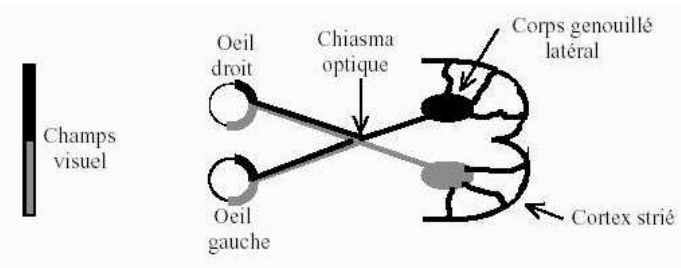

FIG. 1.8: La distribution des réponses rétiniennes aux corps génouillés.

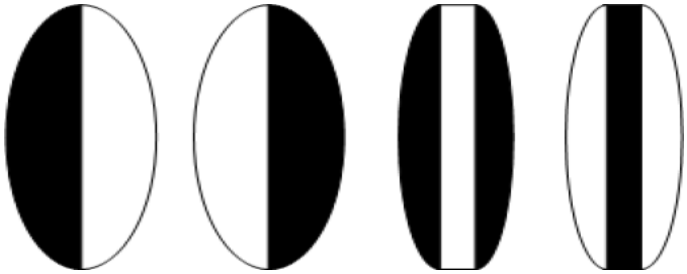

FIG. 1.9: Schématisation des différents champs récepteurs corticaux. Les régions blanches sont des zones d'excitation. Les zones noires sont des zones d'inhibition.

cellules P de la rétine et deux couches magnocellulaires recevant les réponses des cellules M. Par ailleurs, chacune des couches reçoit uniquement les réponses des neurones issues d'un seul oeil : les couches 1, 4, 6 pour un côté et 2, 3, 5 pour l'autre. Le CGL est le principal récepteur des réponses des cellules rétiniennes. A ce niveau, la représentation des réponses des différentes cellules est dite *rétinotopique*. Cela signifie que chaque zone du CGL correspond à une zone de la rétine et que les zones adjacentes à cette zone correspondent également à des zones rétiniennes adjacentes. Toutefois, la représentation rétinotopique (le terme topographique est également utilisé) est fortement non-uniforme. Cette non-uniformité provient de la forte variation de la densité des cônes et des cellules ganglionnaires de la rétine : il y a une très forte densité de cellules ganglionnaires au sein de la fovéa ; cette densité diminuant avec l'excentricité par rapport à la fovéa. Cette non-uniformité attribue donc plus d'importance au centre de la rétine qu'à sa périphérie.
Dans bien des cas, par souci de simplification, le CGL est simplement considéré comme une station relais permettant de véhiculer les réponses des cellules rétiniennes vers l'aire visuelle primaire, classiquement appelé cortex visuel primaire.

1.2.3.2 Le cortex visuel primaire (V1)

Le cortex visuel primaire, situé dans le lobe occipital à l'arrière du cerveau, reçoit les réponses provenant des CGL droit et gauche.
La différence fondamentale existant entre les cellules corticales et les autres cellules (celles de la rétine et du CGL) concerne la structure du champ récepteur. Alors que les champs récepteurs des cellules rétiniennes et celles du CGL sont globalement circulaires, les champs récepteurs des cellules corticales sont de forme elliptique ou encore étirés suivant une certaine orientation. Cette sensibilité à l'orientation, identifiée par les prix Nobel de médecine D. Hubel et T. Wiesel en 1981, constitue donc la différence majeure. La figure 1.9 donne des exemples de champs récepteurs corticaux. La taille de ces derniers varie entre 1 et 7 degrés d'angle visuel, avec une taille moyenne de 2.7 degrés [Felleman 81]. Par ailleurs, D. Hubel et T. Wiesel ont également proposé une classification des cellules corticales en deux familles : les cellules simples et les cellules complexes.

- les cellules simples se composent de deux ou trois zones étirées suivant la direction préférée des cellules. Ces zones ont des comportements antagonistes et par conséquent le champ récepteur est constitué d'une succession de zones excitatrices et inhibitrices. Les orientations préférées de ces cellules varient par plage d'environ 10° ; pour 180°, 20 types de champs récepteurs sont identifiés ;

- les cellules complexes ont également une orientation préférée mais leur champ récepteur est soit de type excitateur (ON) soit de type inhibiteur (OFF).

Les cellules simples et complexes sont regroupées suivant leur sélectivité angulaire dans des colonnes corticales dite orientées : une colonne orientée contient l'ensemble des cellules simples et complexes réagissant à la même orientation. Deux colonnes adjacentes présentent une variation de sélectivité angulaire d'environ 10°. D. Hubel et T. Wiesel ont introduit le concept d'hypercolonne qui est défini comme étant un ensemble de colonnes répondant à une région particulière de l'espace visuel pour toutes les orientations et pour les deux yeux. Ainsi, une hypercolonne est capable d'effectuer une analyse complète d'une région du champ visuel.
Enfin, afin d'être assez complet dans l'énumération des propriétés majeures des cellules corticales mais tout en gardant un niveau de détails cohérent avec nos ambitions, il est nécessaire d'introduire la notion de connections horizontales étendues (le terme anglais *long-range horizontal connection* étant plus connu et plus usité). En fait, les cellules de différentes colonnes verticales orientées sont liées par de longues connections horizontales. Ces dernières ont un rôle important puisqu'elles relient des cellules ayant les mêmes propriétés (même sélectivité angulaire) [Gilbert 89], [Ts'o 86]. Grâce à cette propriété, la réponse d'une cellule à un stimulus peut être modulée par une stimulation extérieure à son champ récepteur. Cette découverte a permis d'affiner et de préciser la définition d'un champ récepteur de la façon suivante :

- Champ récepteur classique (abrégé CRF pour le terme anglais *Classical Receptive Field*) se compose d'un centre et d'un pourtour ayant des comportements antagonistes. On rappelle que les modulations centre/pourtour les plus fréquentes et les plus documentées sont les modulations de nature suppressive. Elles sont généralement maximales lorsque le pourtour présente des caractéristiques similaires à celles du centre :
  - o sensibilité à l'orientation : les modulations du pourtour sont sensibles à l'orientation et sont généralement plus importantes lorsque l'orientation du stimulus périphérique correspond à l'orientation préférée de la cellule [Walker 99]. Lorsque les orientations diffèrent, les effets de suppression sont faibles ou disparaissent. Il existe cependant des exceptions : pour certaines cellules, ce sont les configurations obliques ou orthogonales qui induisent une suppression maximale ;
  - o sensibilité à la fréquence spatiale : les modulations sont maximales lorsque la fréquence spatiale de la stimulation du pourtour et celle de la stimulation du champ récepteur sont similaires [Walker 99] ;
  - o sensibilité au contraste du pourtour : les modulations suppressives augmentent de manière approximativement linéaire lorsque le contraste du pourtour augmente [Walker 99].
- Champ récepteur non classique (abrégé NCRF pour le terme anglais *Non Classical Receptive Field*) se compose d'un CRF et des régions distantes du CRF. Dans ce cas, les phénomènes de modulation centre pourtour ou modulations contextuelles peuvent être de deux types : soit de nature suppressive comme décrit auparavant soit de nature facilitatrice. En effet, la stimulation du pourtour induit le plus souvent une suppression de la réponse liée au stimulus central. Cependant, de nombreuses études ont révélé l'existence de modulations facilitatrices. Les modulations facilitatrices

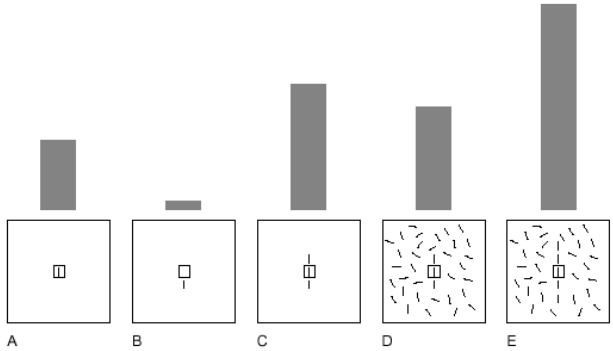

FIG. 1.10: Influence du contexte sur la réponse d'une cellule corticale : sur la dernière ligne, la forme géométrique carré représente l'étendue du CRF. Sur la première ligne, 4 configurations différentes d'effets contextuels sont représentées : (A) une réponse modérée est provoquée par la stimulation du CRF via un stimulus orienté suivant l'axe préféré (ici, vertical). (B) Le même stimulus situé en dehors du CRF engendre une très faible réponse du CRF. (C) La présence de stimuli de même orientation que le stimulus du CRF et alignés avec ce dernier provoque une augmentation sensible de la réponse mettant en évidence un effet de facilitation. (D) La présence de stimuli d'orientations quelconques à l'extérieur du CRF provoque une diminution de la réponse ; c'est l'effet de suppression. (E) Si on augmente le nombre de stimuli de même orientation que le stimulus du CRF et aligné avec ce dernier, la réponse augmente en conséquence (Extrait de M. Kapadia et al. [Kapadia 95]).

maximales sont observées lorsque le stimuli central et périphérique sont iso-orientés et alignés. Les phénomènes de nature suppressive et facilitatrice sont illustrés à la figure 1.10. Les conséquences de cette dernière propriété sont nombreuses et essentielles pour la compréhension de notre environnement visuel. En effet, les connections horizontales étendues jouent un rôle important dans le liage d'informations issues de champs récepteurs éloignés. La facilitation colinéaire participe donc activement à l'intégration des informations afin de les transmettre à notre système cognitif ; la cognition est l'ensemble des processus mentaux s'intercalant entre le stimulus et la réponse et qui permettent de transformer les informations sensorielles en code abstrait. Les sciences cognitives issues d'un courant de pensée apparu dans les années 20 en Allemagne, la Gestalt [Köhler 29],[Koffka 35],[Wertheimer 38], sont actuellement en plein essor. Nous reviendrons sur ce point dans la section 1.3.

1.2.3.3 Les aires visuelles de plus hauts niveaux (V2, V3, V4 et V5) du cortex péristrié

Outre l'aire V1, décrite dans le paragraphe précédent, le cortex visuel contient 4 autres aires visuelles. Avant de donner leurs spécificités, il faut savoir que plus on s'éloigne de V1 et plus l'aspect rétinotopique diminue. En d'autres termes, cela signifie qu'un neurone devient sensible à une plus grande partie du champ visuel. Les grandes caractéristiques des aires péristriées sont données ci-dessous :

– l'aire secondaire, V2, représente avec l'aire V1 les aires dites de bas niveau. Dans

V2, les principales caractéristiques de l'aire V1 se retrouvent ;
- l'aire V3 traite les informations relatives à la dynamique des formes. Elle reçoit des informations de directions via les cellules de la voie dorsale-où de V1 et de V2 ;
- l'aire V4 traite les informations relatives à la couleur et aux formes couleurs. Elle est insensible au mouvement. Elle reçoit des informations des cellules de la voie ventrale-quoi [Livingstone 90] ;
- l'aire V5 est directement liée à la voie dorsale-où. Ses neurones sont sélectifs au mouvement mais pas à la couleur ou à la forme. Ils présentent par ailleurs une très forte sensibilité aux contrastes.

Toutes les descriptions faites jusqu'à présent concernaient le traitement des attributs visuels dit de bas niveau. Pour construire à partir de caractéristiques visuelles a priori indépendantes une représentation sémantiques, les mécanismes d'extraction de traits caractéristiques sont couplés avec des mécanismes de plus haut niveau que nous présentons maintenant sous une forme simplifiée (sachant par ailleurs que ce type de mécanismes reste encore mal connu).

1.3 Les propriétés haut niveau du SVH et leurs difficiles modélisations

1.3.1 L'école Gestaltiste

Les propriétés de haut niveau du SVH, étudiées par l'école Gestaltiste, sont d'une incroyable complexité. Elles lient des informations bas niveau a priori indépendantes afin de construire un objet perceptuel représentant une information de haut niveau sémantique. La phrase suivante, résumant le courant de pensée de l'école Gestaltiste, explicite d'une façon concise mais claire le problème : *the whole is different from the sum of its parts.*
Cette partie n'a pas pour objectif de dresser la liste exhaustive des mécanismes de haut niveau. Les sciences cognitives, c'est à dire l'étude des processus mentaux s'intercalant entre le stimulus et la réponse permettant de transformer les informations sensorielles en code abstrait, est un domaine en expansion mais reste encore relativement méconnu. Notre objectif bien plus modeste est ici de sensibiliser le lecteur à la difficulté de modélisation du système visuel humain.

1.3.2 La perception des formes

A partir des études de l'école Gestaltiste, les principaux effets de la structuration perceptive se basent sur des phénomènes de fusions et sur des phénomènes de fissions de l'information visuelle. On distingue notamment les effets de :
- groupement par proximité : le système visuel a tendance à regrouper les éléments qui sont en proximité spatiale. Sur l'exemple très simple de la figure 1.11 (a), il y a à la fois un phénomène de fusion et un phénomène de fission. Le premier permet de regrouper et de lier des informations visuelles a priori indépendantes afin de former des objets présentant une information sémantique. Le deuxième mécanisme, quant à lui, trie, via un mécanisme de fission, les informations ;

– groupement par similarité : il est représenté à la figure 1.11 (b) ; cette fois-ci, on distingue deux ensembles : le premier constitué de deux sous ensemble de lettres A et le deuxième constitué des lettres B.

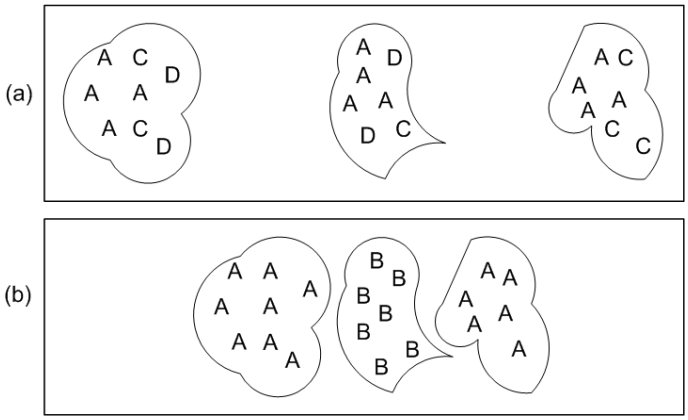

FIG. 1.11: La perception des formes : (a) groupement par proximité ; (b) groupement par similarité.

1.3.3 Phénomènes d'illusions

Les phénomènes d'illusions sont particulièrement intéressants à étudier. De nombreux chercheurs s'accordent sur le fait que les illusions ne sont ni anecdotiques, ni des comportements aberrants de notre système visuel mais bien le reflet des mécanismes inhérents à notre perception. Cette façon de considérer les phénomènes illusoires n'est pas nouvelle ; Helmholtz en 1911 disait :

The study of what are called illusions of the senses is very prominent part of the senses ; for just thoses cases which are not in accordance with reality are particularly instructive for discovering the laws of those means and processes by which normal perception originates.

La figure 1.12 extraite de [Hansen 02] présente sept illusions bien connues. Pour illustrer le fait que des illusions peuvent nous apprendre comment fonctionnent les mécanismes visuels, il suffit d'expliquer ce qui se passe sur l'illusion de la grille d'Herman (cf. figure 1.12 B) : pour cette illusion des zones noires apparaissent entre les différents motifs noirs. L'inhibition pour chaque intersection est supérieure à l'excitation. Les intersections apparaissent donc plus sombres. Cet effet est accentué en périphérie par la plus grande taille des champs récepteurs périphériques.

1.3.4 Limitations des connaissances

La complexité du SVH est en partie maîtrisée pour les parties rétiniennes et pour le cortex visuel primaire. A partir d'études neurophysiologiques et d'expérimentations psychophysiques, de nombreux mécanismes inhérents aux premiers traitements mis en jeu dans l'analyse visuelle peuvent être reproduits via des modèles mathématiques. La sensibilité aux contrastes, l'analyse ondelette du signal optique sont des exemples de traitements

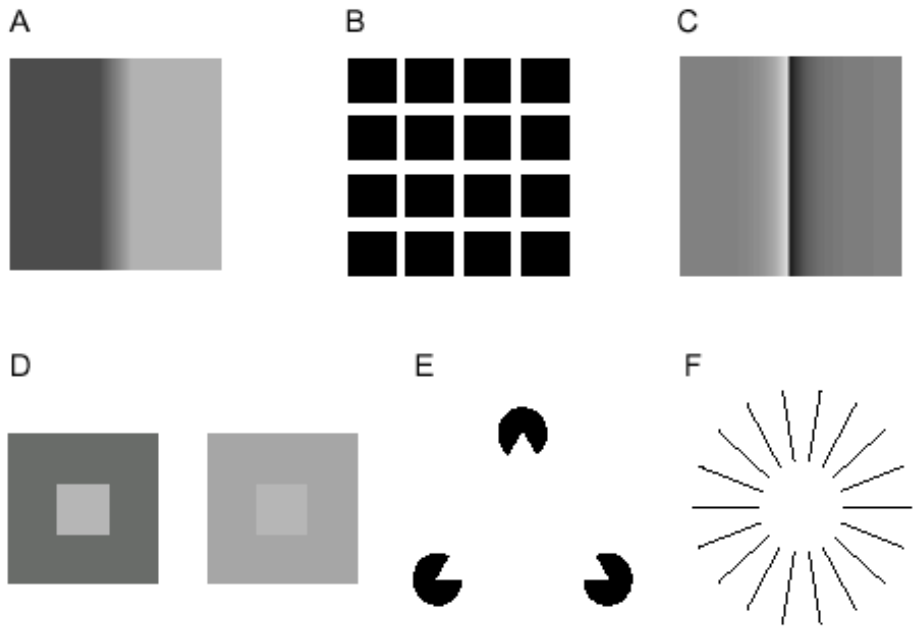

FIG. 1.12: Illusions visuelles liées à la perception de contraste et la formation de contours : (A) Bandes de Mach : le dégradé visible au centre est du à un effet d'illusion. (B) Grille de Herman : des points noirs illusoirs apparaissent entre les carrés noirs. (C) Effet de Craik-O'Brien-Cornsweet : les deux régions adjacentes au signal du centre à fort contraste ont la même luminance. (D) Contraste simultané : le même carré apparaît plus clair sur un fond noir et plus foncé sur un fond clair. (E) Triangle de Kanitza : un triangle apparaît. La luminance de ses côtés semble croître avec le temps d'observation. (F) Figure de Ehrenstein : un cercle de luminosité croissante apparaît (Extrait de [Hansen 02]).

biologiques que nous pouvons aujourd'hui appréhender.
La véritable connaissance du SVH ne s'arrête-t-elle pas à ce stade ? La connaissance des aires corticales supérieures (V2, V3...) reste encore très faible. Au plus, nous sommes capables de leur affecter un type de tâches sans vraiment pouvoir définir et caractériser précisément les mécanismes mis en jeux.
Ce système, intrinsèquement limité, est capable de traiter une quantité considérable d'informations visuelles en partie grâce à un mécanisme passif de réduction de la redondance des informations incidentes (champs récepteurs des cellules rétiniennes et corticales). Mais, bien que le facteur de compression soit de 130 : 1 entre l'information incidente et l'information transmise, un autre mécanisme, un mécanisme actif, est nécessaire pour sélectionner et pour conserver uniquement les informations les plus pertinentes.

1.4 Les mouvements oculaires et l'attention visuelle

1.4.1 Les mouvements oculaires

Afin d'optimiser les ressources de traitement de l'information visuelle inhérentes à notre système, l'homme dispose d'un certain nombre de mouvements oculaires. Bien que

nous n'ayons pas conscience, ces différents types de mouvements prennent la forme de mouvements de poursuites, de convergences, de saccades ou encore de fixations. Les deux mouvements oculaires principaux, associés à la focalisation dite *overt*, sont les saccades et paradoxalement les fixations. Ces deux types de mouvements sont décrits dans les paragraphes suivants.

1.4.1.1 Les saccades

Les saccades sont des mouvements oculaires balistiques dont la vitesse est comprise entre 100 et 700 degrés par seconde [Salvucci 99]. Ce type de mouvement permet de déplacer l'attention visuelle d'un endroit à un autre (un saut d'un point à un autre) afin de les inspecter par la partie la plus performante (en terme de résolution spatiale) de la rétine, la fovéa. Les saccades sont souvent considérées comme un mécanisme favorisant la sélection des informations visuelles pertinentes de notre champ visuel. La scrutation de notre monde visuel se fait donc par une série de sauts permettant le déplacement rapide de nos ressources sensorielles d'un point à un autre. Lorsqu'une saccade est effectuée en direction d'une cible particulière, la précision de la visée peut être soit très bonne soit mauvaise ; dans ce dernier cas, une seconde saccade ajuste le déplacement. Durant ces déplacements, notre pouvoir d'analyse est très faible signifiant que quasiment aucune information visuelle n'est traitée. Notons que le passage d'un point à un autre ne se fait pas forcément par le plus court chemin, c'est à dire la ligne droite. La trajectoire peut en effet être incurvée. Enfin, les saccades sont séparées par des phases de fixations.

1.4.1.2 Les fixations

Une phase de fixations se produit lorsque l'oeil fixe une zone de notre environnement. A première vue, l'oeil a donc une position stationnaire d'où le terme de fixation. Pourtant et paradoxalement, les fixations sont considérées comme des mouvements oculaires. L'explication est en fait très simple : lors d'une phase de fixation, l'oeil est animé d'un mouvement résiduel. Ces légers mouvements permettent de décaler la zone examinée par la fovéa afin que cette dernière soit constamment excitée. Si l'oeil était réellement stationnaire, c'est à dire en vision stabilisée, la perception visuelle disparaîtrait progressivement en raison du mécanisme inhibiteur de l'attention qui est expliqué au paragraphe 1.4.2.4.

1.4.1.3 Les autres types de mouvement

Les autres mouvements oculaires, d'importance secondaire, sont brièvement décrits ci-dessous :

- *les mouvements de poursuite* : ce type de mouvement oculaire permet de suivre un objet en mouvement. Son rôle est important puisque l'objet poursuivi du regard ayant une vitesse relative nulle (la différence entre la vitesse de l'objet et celle du regard) est alors stabilisé sur la rétine et peut donc être examiné par la fovéa avec un fort pouvoir de résolution. La vitesse angulaire limite de poursuite visuelle est d'environ 30 deg/s ;
- *les mouvements de convergence* : c'est un mouvement pour lequel les yeux se déplacent dans des directions horizontales opposées. Ce type de mouvement est utile

pour acquérir des informations visuelles d'un objet proche ou éloigné de notre regard. Par exemple, lorsqu'un objet s'approche de notre regard, les yeux ajustent leurs positions en se rapprochant pour conserver une vision binoculaire alors que, lorsque l'objet s'éloigne, les yeux ont plutôt tendance à s'écarter.

1.4.2 L'attention visuelle sélective

1.4.2.1 Définition

La première définition de l'attention visuelle fut donnée par le père de la psychologie, William James en 1890 [James 90]. Sa définition, dans sa version anglaise originale, est la suivante :

Everyone knows what attention is. It is the taking possession of 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...

Aujourd'hui, grâce aux progrès de la neurobiologie et grâce à des expériences psychophysiques savamment conçues, cette définition a évolué pour prendre une connotation biologique, plus proche de nos capacités sensorielles. Ainsi, l'attention visuelle désigne le mécanisme de sélection des informations visuelles spatio-temporelles pertinentes du monde visible. Étant donné que notre système visuel est intrinsèquement limité en capacité de traitement et que notre environnement visuel contient bien plus d'informations visuelles que nous ne pouvons effectivement traiter (une estimation du débit du nerf optique est tout de même de $10^8 - 10^9$ bits par seconde !), notre système visuel s'est adapté en mettant en place des mécanismes, des stratégies bien particulières pour réduire la quantité d'informations à traiter et pour ne conserver que les informations les plus importantes. En d'autres termes, l'attention visuelle nous permet d'utiliser de façon optimisée nos ressources biologiques ; ainsi, seule une petite partie des informations incidentes est transmise aux aires supérieures de notre cerveau [Ballard 91].
En 1993 R. Milanese [Milanese 93] puis plus tard en 1995 J. K. Tsotsos [Tsotsos 95] décrivent le mécanisme d'attention visuelle comme étant des répétitions de phases de sélection (détection et localisation) et de focalisation (mouvement oculaire ou focalisation interne).

1.4.2.2 Les mécanismes de sélection dit passifs

Les mécanismes de sélection passifs ont été largement abordés dans les paragraphes précédents. En guise de rappel, les principaux mécanismes passifs de sélection de l'information visuelle sont listés ci-dessous :

- le premier mécanisme et le plus évident concerne la transduction photoélectrique (transformation de la lumière en code interprétable par le cerveau). Cette transformation ne concerne qu'une bande étroite du spectre global de la lumière incidente, appelée la lumière visible ;
- l'information est échantillonnée par les cellules photosensibles de façon non uniforme : au centre de la rétine, c'est à dire la fovéa, la restitution de la résolution spatiale de l'information est très bonne ;

- les cellules visuelles présentent une sensibilité aux fréquences spatiales ; en d'autres termes, nous ne sommes pas en mesure d'apprécier tous les détails de notre environnement visuel avec le même degré de précision ;
- les cellules rétiniennes et corticales suppriment la redondance d'informations ; elles répondent uniquement aux contrastes.

1.4.2.3 Les mécanismes de sélection dit actifs

La métaphore du faisceau lumineux plus connue sous le terme de *spotlight of attention* [Neisser 67] a été certainement l'image la plus usitée pour illustrer le concept de l'attention visuelle : l'attention étant comparée à un faisceau lumineux illuminant des zones de notre champ visuel afin de les inspecter.
Notons, tout de suite que la focalisation d'attention, c'est à dire l'inspection d'une zone particulière, peut se faire de deux façons : une focalisation dite *overt* ou une focalisation dite *covert*. Le premier type de focalisation se manifeste directement par un mouvement oculaire. La deuxième, quant à elle, ne met pas directement en jeu un mouvement oculaire. Cette focalisation utilise la vision périphérique, comme lorsqu'on regarde du coin de l'oeil. Cette forme d'attention est particulièrement bien mise en évidence chez les malentendants [Bavelier 00, Muir 03]. En effet, des expériences oculométriques, utilisant des séquences d'images présentant une personne traduisant un discours en langage de signes, ont montré que l'attention fovéale des malentendants se portait essentiellement sur le visage de la traductrice. En dépit du fait qu'ils ne fixaient pas directement les signes, ils étaient tout à fait capables de retranscrire le discours.
Le modèle du faisceau lumineux repose sur le paradigme de la vision binaire : c'est J. Braun et D. Sagi [Braun 90, Braun 94, Braun 98] qui sont à l'origine de la théorie binaire de l'attention visuelle, bien que W. James [James 90] et K. Nakayama et al. [Nakayama 89] avaient déjà émis l'hypothèse d'au moins deux formes de mécanismes de l'attention. Ces deux mécanismes sont les suivants :

- un mécanisme exogène (pré-attentif) [Posner 80] ou plus communément appelé *Bottom-Up* sélectionnant les informations visuelles selon leur saillance. C'est un mécanisme relativement éphémère piloté par les données de notre champ visuel et faisant référence à l'attention involontaire (déplacement oculaire vers des zones capturant notre attention (*Signal driven*)). Ce mécanisme de sélection se fait donc sans aucune connaissance a priori. Les travaux récents de D. Parkhurst [Parkhurst 02] et de R. Peters [Peters 05] montrent toutefois que le mécanisme exogène guide l'attention principalement après l'apparition d'un stimulus visuel mais également bien plus tard dans le temps alors que le mécanisme endogène est supposé actif ;
- le second mécanisme est dit endogène (attentif) [Posner 80] ou *Top-Down*. Notre attention et le déplacement oculaire s'effectuent sous un contrôle volontaire et cognitif. En d'autres termes, ce mécanisme est piloté par la tâche que nous avons à effectuer (*Task-dependent*). La figure 1.13 donne, pour un même observateur et pour cinq tâches différentes à effectuer, les stratégies visuelles associées à la tâche.

Ces mécanismes, et plus particulièrement le mécanisme *Bottom-up*, nous amènent à parler de la théorie de l'intégration de caractéristiques (*Feature Integration Theory* abrégé FIT) de A. Treisman et G. Gelade [Treisman 80]. Ces travaux reposent sur des expériences de

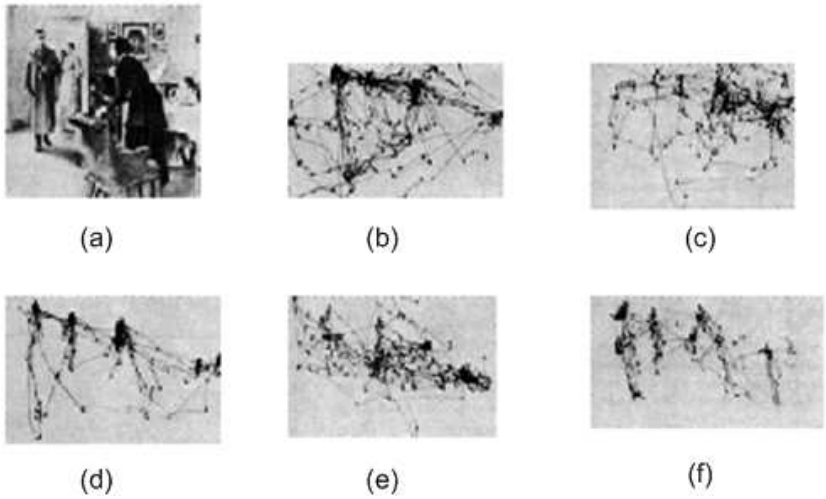

FIG. 1.13: Impact de la tâche à effectuer sur le trajet oculaire lors de la visualisation d'une image par le même observateur : (a) image originale (tableau de I. E. Repin intitulé *unexpected return*) ; (b) aucune tâche n'est donnée (*Free-viewing mode*) ; (c) première tâche : estimer le niveau social des personnages ; (d) deuxième tâche : évaluer leur âge ; (e) troisième tâche : que faisaient les personnages avant l'arrivée du visiteur ? ; (f) quatrième tâche : souvenez-vous des habits portés par les différents personnages (Extrait de [Yarbus 67]).

recherche visuelle. Le principe de ces expériences consiste à mesurer le temps de réaction nécessaire pour discriminer un objet cible enfoui parmi d'autres objets communément appelé distracteurs. Les objets peuvent être simples, c'est à dire constitués d'une seule dimension visuelle (la couleur, l'orientation, la forme...) ou composés de plusieurs dimensions (objet coloré orienté par exemple). Les expériences effectuées révèlent deux comportements distincts :

- si la cible diffère des distracteurs d'au moins une caractéristique visuelle, cas disjonctif (exemple de la figure 1.14 (a)), alors le temps de réaction nécessaire pour résoudre la recherche visuelle est constant et cela quel que soit le nombre de distracteurs. Bien souvent, on considère que la cible saute au yeux (dans la littérature scientifique, le verbe anglais *to pop-out* est très souvent utilisé) ;
- par contre, si la cible est une combinaison de caractéristiques (exemple de la figure 1.14 (b)), le temps de réaction augmente linéairement avec le nombre de distracteurs. Dans ce cas, appelé cas conjonctif, la recherche de la cible est séquentielle puisque tous les objets sont scrutés afin de déterminer la cible.

Ainsi, le cas disjonctif est à rapprocher du mécanisme *Bottom-up* qui, finalement, permet de traiter les caractéristiques visuelles d'une scène rapidement et d'une façon massivement parallèle.

Le cas conjonctif, quant à lui, est à rapprocher du mécanisme *Top-Down* qui est un mécanisme lent et traitant les informations visuelles de façon séquentielle ou série. On parle également de la dichotomie attentif/pré-attentif, représentée par les travaux de A. Treisman et G. Gelade [Treisman 80] et de J.M. Wolfe [Wolfe 89], qui supposent un premier traitement automatique sur l'ensemble du champ visuel suivi d'un traitement localisé

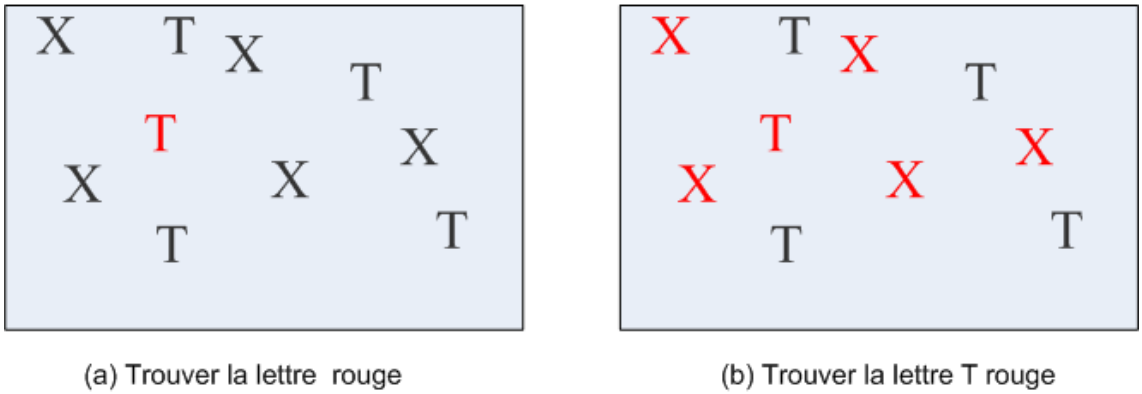

FIG. 1.14: Exemples d'expériences de recherche visuelle : (a) cas disjonctif (traitement parallèle) ; (b) cas conjonctif (traitement série).

déployé par l'observateur.

1.4.2.4 Le mécanisme inhibiteur de l'attention

La fonction première de l'attention visuelle sélective est de diriger notre regard vers des objets d'intérêt contenus dans un environnement visuel général. Comme nous venons de le voir, le mécanisme de l'attention visuelle est constitué de deux types d'attention pouvant être qualifiés de volontaire ou d'involontaire.
En marge de ces deux type d'attention, un autre mécanisme, appelé inhibition de retour en abrégé IOR (*Inhibition Of Return*), n'est pas à négliger. L'inhibition de retour consiste en effet à inhiber une zone inspectée afin d'éviter que notre attention visuelle se porte continuellement sur cette même zone. Ainsi, grâce à l'attention visuelle, nous scrutons séquentiellement les régions en fonction de leur saillance. L'inspection visuelle est par conséquent grandement facilitée par ce mécanisme. Par exemple, dans le cadre d'une recherche visuelle de type conjonctive, de type série, l'inhibition de retour est primordiale puisqu'elle évite à l'observateur de continuellement re-vérifier les mêmes objets [Klein 88, Klein 99]. D'après les études de M. Posner et Y. Cohen [Posner 84], l'inhibition de retour n'a lieu que lorsque la durée d'inspection d'une zone excède 300 ms.

1.4.3 Les caractéristiques visuelles attirant l'attention visuelle

Comme nous venons de le voir, l'humain possède une attention visuelle sélective signifiant que notre système visuel répond de façon privilégiée à un certain nombre de signaux provenant des objets et à des évènements de notre environnement.
Le signal attirant notre attention le plus évidement et le plus intuitif est certainement l'apparition soudaine d'un objet dans une scène [Yantis 96]. Bien souvent, l'exemple d'une proie et d'une espèce prédatrice est donné : en effet, la proie et le prédateur ont une grande faculté à détecter de faibles mouvements afin, respectivement, de pouvoir fuir ou de pouvoir chasser. Plus généralement, on parle de singularité locale [Treisman 80]. Un exemple classique de singularité locale "sautant aux yeux" est représenté à la figure 1.14.
Par ailleurs, l'apparence de nouveaux objets cohérent ou non avec le contexte de la scène

attire notre attention [Hillstrom 94, Henderson 99b]. Lorsque l'objet est incohérent avec la scène, J.M. Henderson et al. [Henderson 99b, Henderson 99a] montrent que les observateurs ont tendance à faire des fixations plus longues et plus fréquentes sur cet objet, saillant sémantiquement.
Enfin, différentes études [Mannan 96, Mannan 97, Reinagel 99] cherchent à estimer, à partir de points de fixation réels, les similarités des caractéristiques visuelles attirant notre regard. D'une façon générale, ces études concernent la mesure de différentes grandeurs telles que la variance normalisée par la luminosité moyenne de l'image, l'entropie et la corrélation entre le point de fixation mesuré et son voisinage. Les principales conclusions sont les suivantes :

- les mesures de contraste des régions fixées sont plus élevées que celles de régions prises au hasard. En d'autres termes, le contraste d'une zone quel qu'il soit (de luminance, de couleur, de mouvement, de texture [Parkhurst 04]...) attire notre attention même lorsque cette zone n'a rien avoir avec la tâche que l'observateur doit accomplir ;
- à partir des mesures de corrélation, ces études montrent également que les régions fixées diffèrent de leur voisinage.

Ainsi, les régions fixées présentent une plus forte hétérogénéité dans leurs attributs visuels que des zones prises au hasard [Mannan 96, Mannan 97].
Notons enfin que ces mesures tendent à montrer que le système visuel essaie de maximiser l'information à transmettre au cerveau en minimisant la redondance spatiale de l'information à transmettre.

1.5 Conclusion

La vocation de ce chapitre était de présenter la physiologie du système visuel humain et l'attention visuelle. Les différentes étapes nécessaire pour transformer un signal lumineux en un signal interprétable par le cerveau ont d'abord été décrites. Le fonctionnement général ainsi que le rôle des cônes, des bâtonnets, des cellules horizontales, bipolaires, amacrines et ganglionnaire ont été évoqués, précédent la description des traitements post-rétiniens. Les propriétés intrinsèques (on utilise également le terme de mécanismes passifs) de ces cellules sont maintenant plutôt bien appréhendées. Néanmoins, si nous arrivons aujourd'hui à maîtriser le comportement de telle ou telle cellule prise individuellement, l'enchevêtrement, les inter dépendances (entre cellules de même types et entre cellules de types différents), les voies de retour modulant les différentes informations restent encore un pan de la recherche fondamentale. A notre niveau, nous avons voulu sensibiliser le lecteur à la complexité du système visuel lorsqu'il est considéré dans son intégralité. L'introduction de la pensée Gestaliste a cet objectif. Enfin, la dernière partie était consacrée à la description de l'attention visuelle, que nous appelons également mécanismes actifs. L'attention visuelle est intimement liée aux mouvements oculaires que nous avons décrits. La description du système visuel humain n'est pas aisée car tous les mécanismes sont dépendants les uns des autres. Afin d'éclaircir ces dépendances, n'apparaissant pas explicitement dans le plan, le mécanisme de perception de l'information est présenté dans le tableau 1.3 extrait de l'article de synthèse de C. Hurst [Hurst 04]. Toutes les données de ce tableau ont été décrites individuellement mais leur intervention dans le temps n'a jamais été évoquée. Néanmoins, de nombreux éléments de réponses apparaissent en filigrane

TAB. 1.3: Liens entre les différents niveaux du système visuel.

attention pré-attentive (*Bottom-Up*)	→	attention attentive (*Top-Down*)
traitement périphérique magnocellulaire (cellules M)	→	traitement central parvocellulaire (cellules P)
mouvement oculaire de saccade	→	fixation / mouvement de poursuite

tout au long de ce chapitre et plus particulièrement dans la section 1.2.2.5. Pour faire le lien entre les différents éléments, considérons l'exemple suivant : supposons qu'une zone saillante apparaisse soudainement dans notre champ de vision périphérique, les cellules M, ayant une très bonne capacité de détection et étant très sensibles au mouvement (Cf. tableau 1.2) décrivent alors de façon grossière l'information visuelle incidente (voie dorsale "*où*" (ou *Where* en anglais) associée à la localisation d'un évènement). Un mouvement de saccade en direction de cette cible est alors initié. Une fois la cible atteinte, des mouvements oculaires de poursuite et de fixation stabilisent la cible au centre de la rétine. L'attention pré-attentive est alors activée et les cellules P transforment avec une grande précision les composantes visuelles de la cible (la voie ventrale, "*quoi*" (ou *What* en anglais) associée à l'identification). L'attention attentive peut alors démarrer. Cet exemple stéréotypé permet d'illustrer la séquentialité et la non concomitance des différents mécanismes du système visuel. Néanmoins, soyons conscients que c'est une vue très simplifiée de la perception visuelle. Nous aurons l'occasion dans les chapitres suivant de revenir sur ces notions.

Chapitre 2

Modèles associés au système visuel humain

2.1 Introduction

L'objet de ce chapitre est de présenter des modèles mathématiques permettant de simuler des propriétés du système visuel dans un contexte donné. Dans une première partie, ce sont les modèles associés aux propriétés dites de bas niveau qui sont décrits. Ces propriétés peuvent également être qualifiées de passives. Il s'agit, en fait, de reproduire le comportement des cellules visuelles via des modèles mathématiques, issus soit de données neurophysiologiques soit d'expériences psychophysiques. Les données neurophysiologiques sont nécessaires pour comprendre le fonctionnement global des mécanismes mis en jeu dans le cerveau et pour décliner des modèles algorithmiques biologiquement plausibles. Les données issues d'expériences psychophysiques, quant à elles, sont nécessaire pour identifier et évaluer les performances et les limitations de la perception visuelle. Nous verrons, par exemple, que la sensibilité au contraste est dérivée des données psychophysiques.
Dans une seconde partie, des modèles mathématiques de l'attention visuelle pré-attentive sont détaillés. L'objectif est de prédire, à partir d'attributs de bas niveau, les positions des zones visuellement importantes d'une image ou d'une séquence d'images. Dans un premier temps, la conception des modèles les plus connus et les plus influents est détaillée. Les modèles présentés forment deux catégories. La première est constituée d'une part de modèles dit empiriques (c'est à dire élaborés à partir de certaines connaissances du système visuel et de l'expérience) et d'autre part de modèles statistiques. Ces modèles sont relativement éloignés du système visuel. La seconde catégorie, quant à elle, regroupe les modèles basés sur une architecture biologiquement plausible. Cette architecture proposée par C. Koch et S. Ullman [Koch 85] est d'abord décrite introduisant la notion de carte de saillance. Les grandes caractéristiques des modèles s'inspirant de cette architecture et plus particulièrement leurs caractères innovants sont ensuite listés. La modélisation de l'attention spatio-temporelle est également abordé dans ce chapitre.
Une discussion, mettant en exergue les limites et les lacunes des modèles actuels, permet d'introduire les axes de recherches dans lesquels s'inscrivent nos travaux.

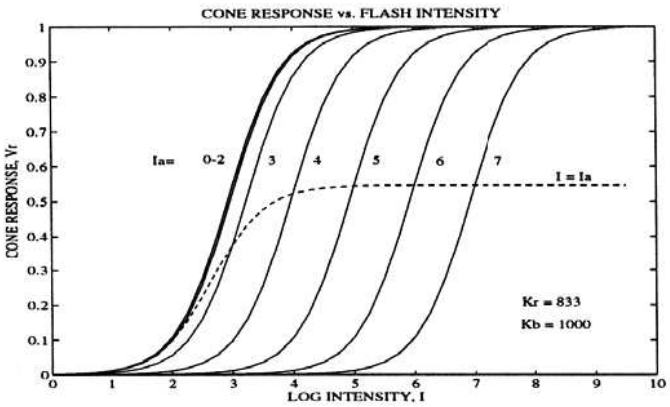

FIG. 2.1: Réponses d'un cône en fonction du logarithme de l'intensité lumineuse incidente (extrait de [Kolb 96]).

2.2 La modélisation des propriétés bas niveau

2.2.1 Le phénomène d'adaptation et la perception de l'intensité lumineuse

Face aux différentes dynamiques d'illumination auxquelles notre système visuel est confronté, un mécanisme d'adaptation est nécessaire afin de limiter la dynamique des signaux d'entrés. Le premier mécanisme d'adaptation est la variation de l'ouverture de la pupille : la pupille a un diamètre important en condition scotopique et faible en condition photopique. Un second mécanisme, moins évident que le premier dans sa mise en oeuvre, modifie le comportement des cellules photoréceptrices. En présence d'une forte illumination, les cônes saturent leurs réponses de façon à ce qu'une augmentation de l'information lumineuse supplémentaire ne génère pas une réponse en sortie des cellules photosensibles plus forte. Ce phénomènes d'adaptation est donc non linéaire et suit approximativement une loi logarithmique illustrée à la figure 2.1. Notons également, que les cellules photosensibles adaptent leur dynamique de sortie en centrant celle-ci autour de la valeur de luminance ambiante déterminée dans une région de faible taille entourant le photorécepteur stimulé.

Outre la capacité d'adaptation, la relation liant la luminance perçue et la luminance réelle n'est pas linéaire. Pour en être convaincu, une expérience simple consiste à donner à un groupe d'observateurs cent cartes représentant cent nuances de gris allant de un à cent. Les observateurs doivent sélectionner dix cartes de façon à obtenir un panel de nuance de gris le plus homogène possible allant du blanc au noir. Dans le cas idéal, l'écart entre deux cartes consécutives doit être le même.
Des expérimentations intensives réalisées par Mumsell ont permis de construire la courbe présentant la relation existante entre la luminance originale et la luminance perçue.

Sur la figure 2.2, différentes courbes apparaissent :
- la courbe associée aux expérimentations de Mumsell ;
- la courbe logarithmique qui a tendance à sur estimer la luminance perçue pour le milieu de gamme des luminances ;
- la courbe linéaire qui a tendance à sous estimer la luminance perçue ;

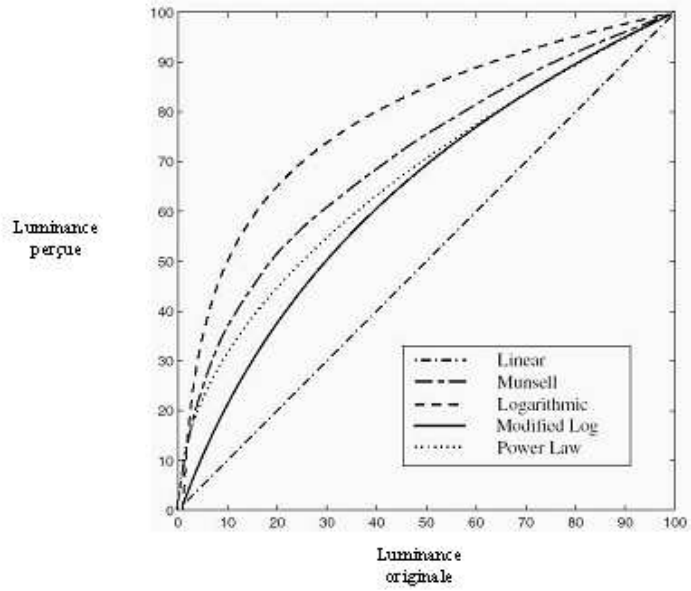

FIG. 2.2: Relation liant la luminance perçue et la luminance réelle. Comparaison avec des modèles mathématiques.

– et la fonction logarithme modifiée qui a tendance à sous estimer la luminance perçue pour les basses valeurs de luminances.

Finalement, c'est la fonction puissance plus connue sous le nom de loi Gamma qui est la relation la plus adaptée pour modéliser la transduction photoélectrique localisée au niveau de la rétine. La dynamique des valeurs faibles de niveau de gris est augmentée alors que celle des zones claires est réduite. Le contraste est donc amélioré et les détails sont accentués.
La luminance perçue, notée Lum_{per}, s'écrit donc en fonction de la luminance originale, notée Lum_{orig} :

$$Lum_{per} = a \times Lum_{orig}^{p} - offset \tag{2.1}$$

H. Bodmann et al. [Bodmann 80] définissent l'exposant p égal à 0.31 ± 0.03. Les deux autres coefficients sont utilisés pour effectuer une adaptation. Classiquement, $a = 1$ et $offset = 0$.
En règle générale et dans un contexte d'affichage d'images sur un moniteur, la fonction Gamma d'exposant $p \times n$, avec $p = 0.31$ et $n = 2.3$ est utilisée pour modéliser la transduction photoélectrique et pour compenser le comportement non linéaire des écrans.

2.2.2 La perception des couleurs

Les différentes données biologiques et psychophysiques laissent penser que le SVH sépare l'information lumineuse reçue sur la rétine en trois composantes distinctes mais pas totalement indépendantes. Ainsi, les signaux issus des trois types de cônes (L, M, S) sont combinés afin de définir les axes chromatiques psycho-visuels conduisant aux composantes couleurs psycho-visuelles que nous appellerons Cr_1 et Cr_2. Ces deux composantes

complètent la composante psycho-visuelle achromatique que nous noterons A.
Différentes modélisations de la perception des couleurs par le SVH ont été proposées, majoritairement inspirées par les données biologiques et psychophysiques et par la théorie des composantes antagonistes ; les antagonismes chromatiques rouge-vert et jaune-bleu ont été évoqués précédemment.
La plupart du temps, cela revient à définir une transformation linéaire

$$\begin{pmatrix} A \\ Cr_1 \\ Cr_2 \end{pmatrix} = [T] \times \begin{pmatrix} L \\ M \\ S \end{pmatrix} \tag{2.2}$$

La matrice T peut être déterminée de différentes façons : de façon physiologique comme l'ont proposé R. De Valois [Valois 92] et O. Faugeras [Faugeras 76]. Le premier se base sur les signaux arrivant sur les zones excitatrices et inhibitrices des champs récepteurs :

$$[T]_{DeValois} = \begin{pmatrix} 0.375 & 0.6875 & 0.0625 \\ 0.5625 & -0.7187 & 0.1562 \\ -0.8125 & 0.5938 & 0.2187 \end{pmatrix} \tag{2.3}$$

Le second utilise les différentes courbes d'absorption des différents types de cônes :

$$[T]_{Faugeras} = \begin{pmatrix} 13.63 & 8.33 & 0.42 \\ 64 & -64 & 0 \\ -5 & -5 & 10 \end{pmatrix} \tag{2.4}$$

Une autre façon consiste à déduire la matrice T à partir d'expériences psychophysiques. Citons le modèle de M. Webster [Webster 90], de P. Flanagan [Flanagan 90] et celui de J. Krauskopf [Krauskopf 82].
La matrice T définie par J. Krauskopf nous intéresse particulièrement car c'est cet espace de représentation que nous utiliserons dans cette étude :

$$[T]_{Krauskopf} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ -0.5 & -0.5 & 1 \end{pmatrix} \tag{2.5}$$

2.2.3 La sensibilité aux contrastes

La sensibilité aux contrastes est une propriété clé de notre système visuel. Elle est basée sur notre capacité à détecter des différences de luminances, de couleurs, de mouvement... En se plaçant dans le domaine de luminance, des expériences psychophysiques (effectuées à partir de stimuli appropriés) permettent de mesurer notre sensibilité, dans un contexte strict (signaux simples, environnement maîtrisé...). Pour des signaux stationnaires, l'expression de contraste donnée par Michelson est la plupart du temps utilisée :

$$C = \frac{L_{max} - L_{min}}{L_{max} + L_{min}} \tag{2.6}$$

où, L_{max} et L_{min} représentent respectivement la luminance maximum et minimum du stimulus utilisé.

Le contraste minimal sera dans le contexte particulier étudié (fréquence spatiale donnée, distance de visualisation donnée...), la plus petite valeur de contraste susceptible d'entraîner la détection du stimulus par le sujet.
Le seuil de sensibilité au contraste ou le seuil de visibilité, toujours dans le même contexte, sera l'inverse de la valeur du contraste minimal définie ci-dessus. Cette sensibilité sera donc d'autant plus élevée que le contraste détecté sera faible et inversement.
La luminance des stimuli n'est pas le seul facteur influant sur le seuil de sensibilité. D'autres caractéristiques sont essentielles et la taxinomie suivante peut être faite :

- des caractéristiques de bas niveau telles que la luminance de fond, la fréquence spatiale, la couleur, la fréquence temporelle, l'orientation du stimulus...
- des caractéristiques de haut niveau telles que l'âge des sujets, la distance d'observation...

Ces dépendances sont généralement modélisées par une fonction de sensibilité au contraste (FSC, mais plus connu sous l'abréviation anglaise CSF pour *Contrast Sensitivity Function*). Dans la suite, des CSFs modélisant la sensibilité visuelle aux contrastes spatiaux et aux contrastes spatio-temporels sont présentées.

2.2.3.1 Sensibilité aux contrastes spatiaux (luminance et couleur)

Dans la littérature, un ensemble assez complet de courbes de fonction de sensibilité pour des signaux achromatiques, pour différentes configurations de stimuli, est fourni par E. Peli et al. [Peli 93]. Un article récent de P. Barten [Barten 04], l'un des précurseur dans la modélisation de CSFs, propose une formulation plus complète des CSFs, en terme de dépendance à de nombreux paramètres.
Un exemple de CSF classique isotrope, proposée par J. Mannos et D. Sakrison [Mannos 74] est représenté à la figure 2.3. Sa formulation est la suivante :

$$CSF(f) = 2.6 \times (0.0192 + 0.114f)exp(-(0.114f)^{1.1}) \tag{2.7}$$

où, f est exprimé en cycle par degré (cpd).

Cette courbe de sensibilité montre que nous sommes plus sensibles aux fréquences spatiales intermédiaires (entre $6-10$ *cpd*) qu'aux basses et aux hautes fréquences. Au delà de 50 *cpd*, l'oeil ne détecte plus rien. L'oeil a donc un comportement passe bande vis à vis des fréquences spatiales de la composante luminance.

En d'autres termes, pour des zones de faibles fréquences spatiales, l'apparition d'un signal (une dégradation ou un artefact de codage vidéo par exemple) même à faible contraste risque d'être gênante pour l'observateur. A contrario, pour des zones de fréquences spatiales élevées, l'apparition d'un signal ayant un contraste moyen provoquera moins de gêne.
Nous présentons ici une autre fonction de sensibilité au contraste de luminance, celle de S. Daly [Daly 93]. Cette fonction anisotrope nous intéresse particulièrement puisque nous l'utiliserons dans la suite de nos travaux. Elle exprime la sensibilité en fonction de la fréquence radiale w en cycle par degré, l'orientation θ en degrés, le niveau d'adaptation en luminance l en cd/m^2, la surface de l'image s en degré^2, la distance d'observation d en mètre et e l'excentricité en degré. Son expression est la suivante :

$$S_A(w, \theta, l, s, d, e) = P \times min\left(S(\frac{w}{bw_a \times bw_e \times bw_\theta}, l, s), S(w, l, s)\right) \tag{2.8}$$

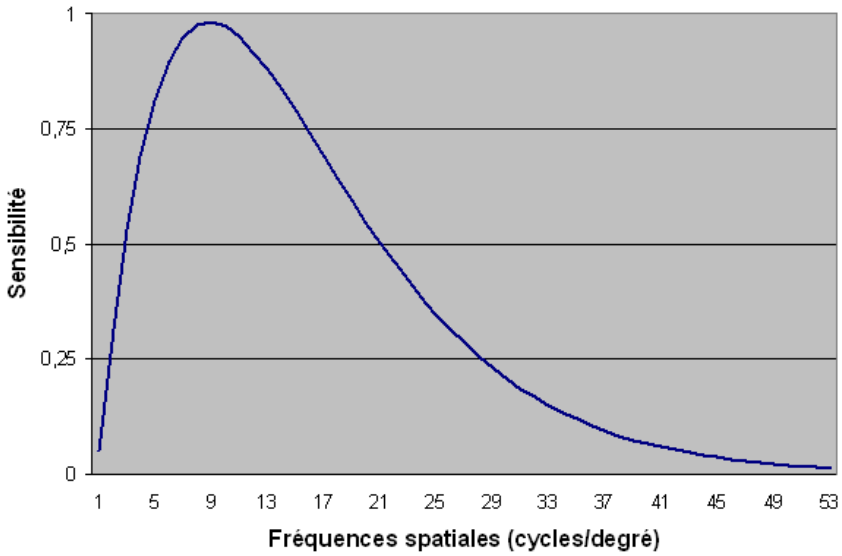

FIG. 2.3: Fonction normalisée de sensibilité au contraste proposée par J. Mannos et D. Sakrison [Mannos 74].

où, le paramètre P désigne la sensibilité maximale ; cette dernière varie d'un observateur à un autre mais la valeur 250 utilisée par S. Daly a été conservée. Les paramètres bw_a, bw_e et bw_θ modélisent respectivement l'influence de la distance, de l'excentricité et de l'orientation. Ils sont donnés par :

$$bw_a = 0.856 \times d^{0.14} \tag{2.9}$$

$$bw_e = \frac{1}{1 + 0.24 \times e} \tag{2.10}$$

$$bw_\theta = 0.15 \times cos(4\theta) + 0.85 \tag{2.11}$$

$S(w, l, s)$, quant à elle, est définie par :

$$S(w, l, s) = ((3.23(w^2 s)^{-0.3})^5 + 1)^{-\frac{1}{5}} \times A_l \times 0.9w \times e^{-(B_l 0.9w)} \sqrt{1 + 0.006 \times e^{B_l 0.9w}} \tag{2.12}$$

où,
le paramètre A_l est donné par : $A_l = 0.801(1 + 0.7/l)^{-0.2}$,
le paramètre B_l est donné par : $B_l = 0.3(1 + 100/l)^{0.15}$.

Classiquement, on utilise les valeurs de paramètres suivants : $P = 250$, $l = 100$ et $e = 0$. La valeur nulle pour ce dernier paramètre est issue d'une hypothèse qui considère que toute l'image est vue et inspectée par la zone fovéale de la rétine.

En ce qui concerne la couleur, des études ont montré que notre sélectivité était plus importante ; la courbe de sensibilité est proche de celle d'un filtre passe bas. La modélisation par P. Le Callet [Callet 01] des CSFs associées aux composantes Cr_1 et Cr_2, a permis de déterminer les fréquences de coupure des filtres passes bas. Elles sont respectivement de

5.5 et 4.1 *cpd*. Pour la composante Cr_1, l'expression de la CSF est la suivante :

$$S_{Cr_1}(w,\theta) = \frac{33}{1+(\frac{w}{5.52}^{1.72})}(1+0.27sin(2\theta)) \tag{2.13}$$

Concernant la composante Cr_2, l'expression de la CSF est la suivante :

$$S_{Cr_2}(w,\theta) = \frac{5}{1+(\frac{w}{4.12}^{1.64})}(1-0.24sin(2\theta)) \tag{2.14}$$

La figure 2.4 présente les courbes de sensibilité des CSFs associées aux composantes Cr_1 et Cr_2 pour une orientation θ nulle.

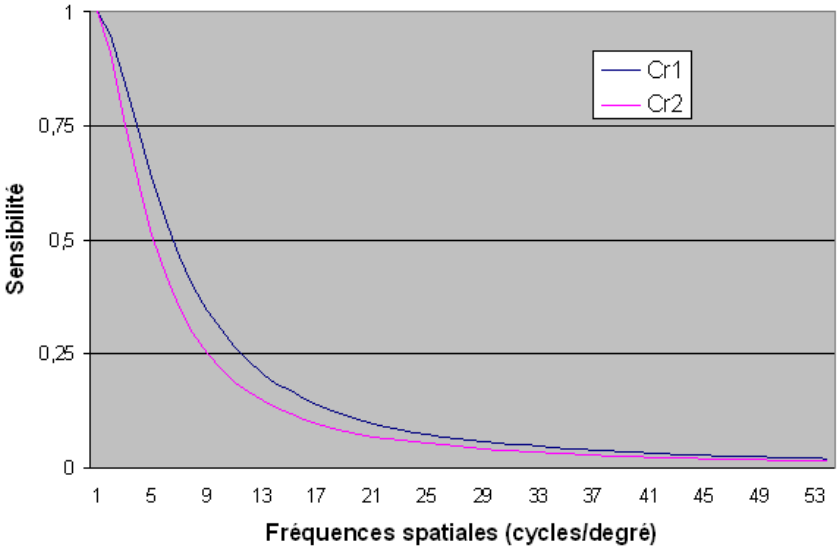

FIG. 2.4: Fonctions normalisées de sensibilité au contraste proposées par P. Le Callet [Callet 01] pour les composantes couleurs (notées Cr_1 et Cr_2) et pour une orientation $\theta = 0^\circ$.

2.2.3.2 Sensibilité aux contrastes spatio-temporels

Par analogie au domaine spatial pour lequel des fonctions de sensibilités aux contrastes sont définies, il est possible de définir des CSFs temporelles.
Lorsque le contraste de la cible varie sinusoïdalement à certaines fréquences temporelles, des faibles variations d'amplitudes sont invisibles ou plutôt ne sont pas discernables par notre système visuel. Par contre, lorsque le contraste augmente, la cible peut devenir visible.
Les premières expérimentations, menées par De Lande en 1958 permettant d'obtenir des CSFs temporelles ont montré que :

- le maximum de sensibilité (200) était obtenu pour un fort contraste et pour une fréquence temporelle de l'ordre de $8Hz$;
- au delà de la fréquence de $8Hz$, la sensibilité décroît très rapidement. La fréquence temporelle pour laquelle la sensibilité est égale à 1 est obtenue pour une fréquence

comprise entre 50 et $70Hz$. Cette fréquence appelée CFF (*Critical Flicker Frequency*), représente la transition (au niveau de nos sens) entre un scintillement de lumière et une lumière continue. En d'autres termes, au delà de cette fréquence, la variation temporelle de la lumière est perçue par note système visuel comme une lumière continue ;
- la sensibilité décroît également pour les basses fréquences temporelles mais cette décroissance se fait de façon modérée. Aux très basses fréquences, la sensibilité est d'environ 50.

Bien qu'il soit commode de considérer la CSF temporelle indépendamment des autres dimensions visuelles, il apparaît cependant que la sensibilité temporelle est intimement liée à la fréquence spatiale du stimulus visuel. La séparabilité des deux dimensions (spatiale et temporelle) serait obtenue si la sensibilité était obtenue via le produit des fonctions de sensibilités $s_u(u, \theta)$ et $s_w(w)$ représentant respectivement la sensibilité aux contrastes spatiaux et la sensibilité aux contrastes temporels :

$$s(u, w) = s_u(u, \theta) s_w(w) \tag{2.15}$$

où, u, w et θ représentent respectivement la fréquence spatiale, la fréquence temporelle et l'orientation.
Pour la vision humaine, la sensibilité spatio-temporelle n'est pas une fonction séparable. De nombreuses études ont montré qu'il y avait une forte interaction entre la perception spatiale et temporelle. Les facteurs influençant la sensibilité temporelle sont notamment :
- la taille de la cible : une cible de taille importante tend à réduire la sensibilité temporelle à basse fréquence temporelle. A haute fréquence, la sensibilité est quasiment inchangée ;
- les contours : une cible présentant des contours contrastés augmente la sensibilité à basses fréquences temporelles. Il n'y a pas d'effet sur les hautes fréquences.

En résumé, pour les fortes fréquences spatiales et temporelles, les deux fonctions de sensibilité au contraste peuvent être considérées comme indépendantes. La formule (2.15) est donc utilisable. Par contre, pour les faibles fréquences spatiales et/ou temporelles, la fonction n'est pas séparable.

2.2.3.3 Limitations des CSFs

Ces fonctions sont largement utilisées aujourd'hui. Néanmoins, il y a un certain nombre de limitations souvent négligées. Le problème majeur réside dans la prise en compte de la capacité de notre système à s'adapter à la luminance de fond entraînant une modification de notre sensibilité visuelle (Cf. figure 2.5). Puisque les CSFs sont déterminées pour une luminance de fond constante durant les expérimentations, elles ne prennent pas en compte ce phénomène d'adaptation. L'estimation de notre sensibilité visuelle peut donc être biaisée. Par ailleurs, les caractéristiques du panel d'observateurs utilisés peuvent engendrer des différences dans l'obtention des CSFs [Valois 88]. Enfin, la limitation la plus sévère consiste à considérer le système visuel comme un système mono-canal dont les caractéristiques seraient uniquement données par une CSF. Or, il est clair que le système visuel est un système multi-canal avec des canaux non indépendants. Les interactions à

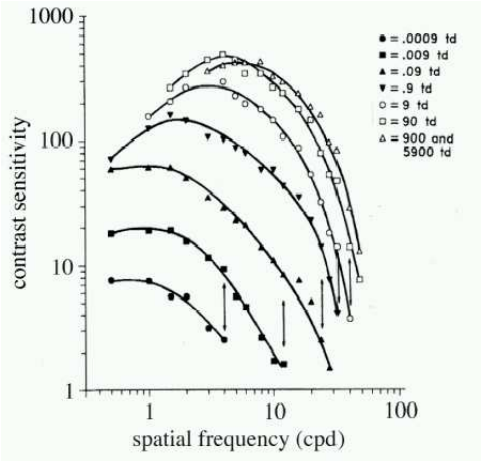

FIG. 2.5: CSF pour différentes luminance de fond. Les niveaux de luminance sont exprimés en Troland (Td) (Trolands = luminance en $cd/m^2 \times tailledelapupille$) (Extrait de [Pattanaik 89]).

l'intérieur d'un canal et entre les différents canaux, plus communément appelé effets de masquage, ne peuvent être convenablement abordés qu'en considérant la structure multi-résolution du système visuel.

2.2.4 Classification des cellules corticales via une organisation multi-canal en fonction de la fréquence spatiale et de leur sélectivité angulaire

Comme nous l'avons précédemment introduit, les cellules corticales présentent une sélectivité en fréquence et en orientation. Les premières études et notamment celle de D. Hubel et T. Wiesel ont montré que les orientations préférées de ces cellules variaient par plage de 10°. Aujourd'hui, le pavage fréquentiel du SVH, c'est à dire les caractéristiques de la décomposition de l'information visuelle en un ensemble de canaux, est décrit en terme de sélectivité radiale, notée r (de 1 à 2 octaves) et en terme de sélectivité angulaire, notée θ (20 à 60°). La modélisation de ces comportements nécessite donc l'utilisation de représentations multi-résolutions simulant les différentes populations de cellules corticales (plus précisément, c'est la taille des champs récepteurs qui est simulée par les différents niveaux de résolution). Plusieurs représentations multi-résolutions sont décrites ci-après.

2.2.4.1 Transformée Cortex

A. Watson [Watson 87] a le premier proposé une transformation appelée transformée Cortex. Il décompose le signal en 5 couronnes de fréquences radiales ayant chacune une largeur de bande d'une octave et présentant une sélectivité angulaire constante de 45° (sauf pour la couronne des très basses fréquences).
Cette décomposition en canaux perceptuels utilise les filtres Cortex construits à l'aide de deux autres filtres appelés filtre DoM (filtre à sélectivité radiale) et filtre Fan (filtre à sélectivité angulaire). Étant donné que ces filtres sont utilisés dans nos travaux, ils seront

détaillés au paragraphe 2.3.4 de la partie II.

Les filtres DoM, filtres passe-bandes radiaux, sont obtenus par soustraction de deux filtres passe-bas à symétrie circulaire et de fréquences de coupures différentes, appelés filtre $Mesa$. Ils sont modélisés dans le domaine des fréquences spatiales radiales par la convolution d'une fonction porte de gain unité par un filtre gaussien.
Le filtres DoM décrit précédemment permettent de construire des couronnes. Il s'agit maintenant de découper les couronnes en fonction de leur sélectivité angulaire. Pour cela, les filtres directionnels Fan sont utilisés et construits à partir de filtres de bissection. Ils se calculent en effectuant la convolution d'un échelon bidimensionnel orienté avec une gaussienne. Enfin, un filtre Cortex est obtenu en multipliant un filtre Fan par un filtre DoM.
A partir d'expériences psychophysiques sur la luminance et la couleur effectuées au travers de différente thèses [Sénane 96], [Bedat 98] et [Callet 01], les largeurs de bande des canaux ainsi que leurs sélectivités angulaires pour des conditions normalisées de visualisation, correspondant à une distance d'observation égale à 6 fois la hauteur de l'écran, ont été mesurées. Ces expériences ont permis d'affiner les paramètres des filtres Cortex proposés par A. Watson. Dans nos travaux de modélisation, nous utiliserons cette décomposition. Ces paramètres et ces caractéristiques seront donc donnés précisément ultérieurement.

2.2.4.2 Filtres de Gabor

Les filtres de Gabor (passe-bande orienté) sont également largement utilisés pour définir le pavage fréquentiel du SVH. Il est en effet admis que les formes des champs récepteurs corticaux peuvent être approximées par des ondelettes bidimensionnelles de Gabor couvrant un certain nombres de bandes de fréquences et d'orientations [Marcelja 80, Daugman 80]. Les fonctions de Gabor ont été définies à partir du principe d'incertitude, stipulant qu'une fonction ne peut être correctement localisée à la fois dans les domaines spatial et fréquentiel. Dans ce cas, les fonctions de Gabor sont le meilleur compromis. Leur avantage, comparativement à des approches plus classiques telles que Fourier, réside donc dans leur bonne localisation tant spatiale que fréquentielle. L'inconvénient des filtres de Gabor est leur sélectivité angulaire relativement faible. Ainsi, pour reproduire le pavage fréquentiel du SVH, il serait nécessaire en pratique de considérer un grand nombre d'orientations.

2.2.4.3 Autres filtres

Mis à part les filtres précédemment cités, d'autres représentations multi-résolutions, plus directes à implanter mais moins fidèles à notre système visuel, peuvent être utilisées. Des approches de transformations pyramidales classiques, initialement proposées par P. Burt et al. [Burt 83], présentent une bonne localisation spatiale. Les sélectivités fréquentielle et angulaire, par contre, sont mauvaises. Une analyse ondelette classique où la variation en résolution d'un niveau à l'autre est définie selon une loi en puissance de 2 peut également être utilisée. L'intérêt majeur est d'avoir ici une excellente localisation spatiale pour tous les niveaux de résolution. La localisation fréquentielle et la sélectivité angulaire réduite à 0, 45, 90 et 135 degrés sont des défauts majeurs de cette approche.

2.2.5 Les effets de masquage visuels spatiaux

2.2.5.1 Définition et illustrations

Précédemment, nous avons présenté le concept de fonction de sensibilité aux contrastes. Ces fonctions traduisent le fait que les cellules visuelles ne répondent à un stimulus que si ce dernier présente un contraste supérieur à une valeur seuil appelé seuil de visibilité.
Malheureusement, cette modélisation mono-canal du système visuel qui reste cohérente pour des signaux simples ne l'est plus lorsque les signaux sont complexes. En d'autres termes, le seuil de visibilité ne dépend pas que de la valeur de contraste du stimulus mais aussi de l'environnement dans lequel il se situe. Une décomposition multi-résolution semble inévitable pour pouvoir espérer modéliser correctement cette modulation, communément appelée effet de masquage visuel. Il exprime soit la réduction (*masking effect*) de la visibilité d'un stimulus (appelé signal masqué) par un autre stimulus (appelé signal masquant) soit l'augmentation de sa visibilité (*pedestal effect*).
Une illustration simple mais pédagogiquement efficace du masquage visuel peut se faire par analogie avec le masquage audio : soit deux sonorités, une faible et une forte, auxquelles nous sommes sensibles si on les considère de façon indépendante. Si on juxtapose les deux sonorités, nous ne serons sensibles uniquement à la sonorité la plus forte qui se comportera alors comme un signal masquant vis à vis de la sonorité la plus faible. Toutefois, précisons que si les fréquences des deux sonorités sont très différentes, les deux sons seront perçus.
Concernant le masquage visuel, plus complexe, des études ont montré qu'il existait trois types d'effet de masquage [Bonds 89, Heeger 93, Foley 94, Callet 01] :

- le plus important est incontestablement le masquage intra canal se traduisant par une interaction entre stimuli de mêmes caractéristiques (fréquence, orientation...). La modélisation proposée par S. Daly [Daly 93] de ce type de masquage sera décrite dans le paragraphe 2.2.5.3 ;
- le masquage entre stimuli de caractéristiques différentes c'est à dire n'appartenant pas au même canal. Cet effet est appelé masquage inter-canal ;
- le masquage entre différentes composantes (l'impact de la couleur sur la perception de structure de contraste de luminance par exemple).

Cette brève énumération montre une nouvelle fois l'intérêt de procéder à une décomposition en canaux visuels.

2.2.5.2 Courbes caractéristiques du masquage

Les courbes caractéristiques du masquage sont appelées TVC (*Target contrast threshold Versus masker Contrast* traduit en français par seuil du contraste cible par rapport au contraste masquant). Le masquage intra canal peut être décrit par la caractéristique de la figure 2.6 avec :

- Δ : le seuil différentiel de visibilité du stimulus en présence d'un stimulus masquant de contraste C_M ;
- Δ_0 : le seuil différentiel de visibilité du stimulus en absence de stimulus masquant (il est donné par la CSF : $\Delta_0 = \frac{1}{CSF}$).

Lorsque $C_M < C_{M_0}$, le signal masquant ne perturbe pas la perception du signal. A l'opposé, lorsque $C_M > C_{M_0}$, la perception devient plus difficile. La valeur de C_{M_0} dépend des formes des signaux masqué et masquant.

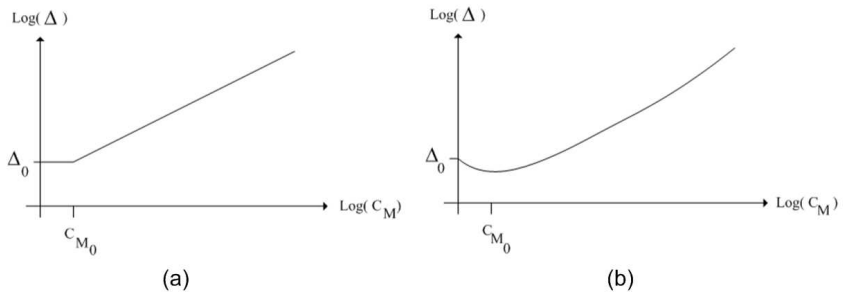

FIG. 2.6: Caractéristique du masquage d'un stimulus par un autre : (a) sans facilitation, (b) avec facilitation.

La courbe de masquage, illustrée sur la figure 2.6 (b), laisse apparaître une zone de facilitation. Lorsque C_M est inférieur à C_{M_0}, la courbe présente un affaissement signifiant que le signal masquant aide à la perception du stimulus. Comme souligné par J. Foley et al [Foley 94], un phénomène de facilitation est possible lorsque les orientations des signaux masqué et masquant sont proches.

2.2.5.3 Le masquage intra canal proposé par S. Daly

Le modèle de masquage de S. Daly [Daly 93] est un modèle simplifié dans le sens où seules les interactions de masquage intra canal (la facilitation n'est pas prise en compte) sont modélisées. Il s'applique sur une décomposition en sous bandes perceptuelles. L'avantage majeur de ce modèle réside dans le fait qu'il résulte d'un nombre conséquent de résultats expérimentaux.
Dans ce modèle, la modulation du seuil différentiel de visibilité, c'est à dire dans ce cas son élévation, est calculée en chaque site s de chaque sous bande perceptuelle (ρ, θ) par la relation suivante :

$$Elevation_{\rho,\theta}(s) = (1 + (k_1 \times (k_2 \times |f_{\rho,\theta}(s)|)^l)^b)^{\frac{1}{b}} \tag{2.16}$$

avec :
- $f_{\rho,\theta}(s)$: valeur au site s de la sous bande (ρ, θ). (ρ, θ) correspondent respectivement à la bande de fréquences et à la gamme d'orientations de la sous bande considérée ;
- $k_1 = 0.0153$;
- $k_2 = 392.5$;
- l, b : constantes dépendant de la sous bande.

Comme il ne s'agit que d'un calcul d'élévation du seuil différentiel de visibilité et non d'un nouveau seuil, la CSF de Daly est appliquée avant le masquage. Le nouveau seuil différentiel de visibilité incluant l'effet de masquage est donc :

$$Seuil_{\rho,\theta}^{Masquage}(s) = Seuil_{\rho,\theta}^{CSF}(s) \times Elevation_{\rho,\theta}(s) \tag{2.17}$$

2.2.6 Modélisation des réponses des cellules corticales

Il existe plusieurs niveaux de modélisation des réponses des cellules corticales allant du plus simple au plus complexe. Les modèles peuvent, en effet, être limités au champ

récepteur ou tenir compte d'une zone s'étendant au delà du champ récepteur (via les connexions horizontales étendue). Par ailleurs, les réponses sont obtenues directement ou via un mécanisme prenant en compte différentes contre-réactions (modèle récursif). On ne détaille ici que les modèles simples. Le lecteur pourra se référer à [Hansen 02] (page 40 à 47) pour obtenir de plus amples informations.
Le point commun entre ces différents modèles est l'utilisation soit de différence de fonctions gaussiennes orientées ou non soit de fonctions de Gabor bidimensionnelles. Ces filtres présentent l'avantage de pouvoir modéliser à la fois la partie excitatrice mais aussi la partie inhibitrice des champs récepteurs. Nous aurons l'occasion de revenir sur cette modélisation puisque le modèle que nous proposerons simulera le comportement des cellules corticales.

2.2.7 Décomposition temporelle de l'information

Tout comme la dimension spatiale se caractérise par un ensemble de sous bandes (ou canaux) sélectives en orientation et en fréquence spatiale, des études ont montré qu'il existait dans notre système visuel des cellules corticales sensibles à certaines fréquences temporelles [Valois 00]. Approximativement, 20 à 25% de la population des cellules corticales simples ont une réponse transitoire, de courte durée à un stimulus temporel adapté. Cette réponse, illustrée par la courbe appelée *strongly phasic* de la figure 2.7 (b), se caractérise par une forte excitation suivie d'une forte inhibition. Le reste de la population a un comportement tout à fait différent. En effet, leurs réponses, provoquées par une stimulation adaptée, se caractérisent par une forte excitation suivie d'une faible inhibition. Par ailleurs, le temps nécessaire pour atteindre le maximum de l'excitation est relativement important (courbe appelée *weakly phasic* de la figure 2.7 (b)). Notons également que les deux réponses sont quasiment en quadrature de phase, d'où la non concomittance de leurs réponses.
En 2000, R. De Valois et al. [Valois 00] ont fait le rapprochement entre les propriétés des cellules précédemment citées et celles des couches magnocellulaires (M) et parvocellulaires (P). Le tableau 1.2 (présenté dans le chapitre 1) rappelle les différentes caractéristiques de ces cellules M et P. On constate que les cellules M présentent une forte capacité de détection s'appariant très bien avec les réponses de style *strongly phasic* ou *transient*. Par contre, les cellules P ont une faible capacité de détection (temps de réponse important à un stimulus) s'appariant cette fois-ci avec les réponses de style *weakly phasic* ou *sustained*.
La figure 2.7 (a) donne un autre exemple de filtres modélisant les deux canaux temporels de notre système visuel. Notons, tout de même, que le nombre de canaux n'est pas encore très bien déterminé. Si il ne fait aucun doute sur le canal *sustained*, caractérisé par un filtre passe bas temporel, il n'en est pas de même pour le canal *transient*. Doit-on considérer un ou plusieurs canaux dédiés aux fréquences temporelles supérieures [Hammett 92], [Fredericksen 97, Fredericksen 98] ? La question reste encore ouverte aujourd'hui même si la tendance est à considérer un canal *transient* et un canal *sustained*.

2.2.8 Les effets de masquage visuel temporel

Comparativement au masquage spatial, les effets de masquage temporel sont moins connus et la littérature les concernant est bien moins abondante.
Avant de décrire deux phénomènes de masquage temporel, rappelons que ce dernier module le seuil de visibilité d'un signal en fonction d'évènements temporels. Tout comme le

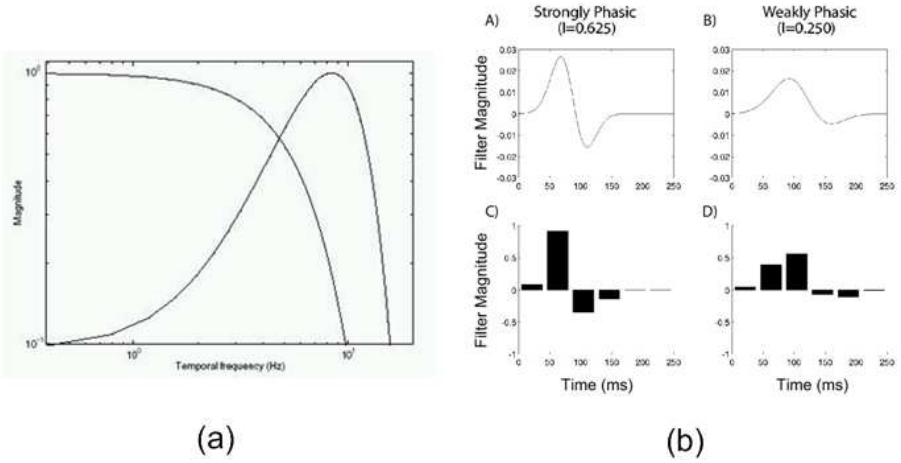

FIG. 2.7: Exemples de modèles de décomposition temporelle de l'information : (a) filtres passe bas et passe bande proposés par [Lambrecht 96] ; (b) sur la première ligne, réponses impulsionnelles des filtres idéaux de décomposition temporelle en deux canaux et filtres transversaux associés sur la seconde ligne (extrait de [Parkhurst 02]).

masquage spatial, cette modulation peut faciliter ou atténuer la visibilité d'un stimulus.
Le premier type de masquage, qui est le plus intuitif, est le masquage avant (*forward* en anglais) apparaissant après une discontinuité temporelle (typiquement un changement de plan dans un contexte vidéo) [Seyler 59, Seyler 65]. La perception des images situées après le changement de plan se trouve alors réduite pendant une durée d'environ 100 millisecondes (4 images en $25Hz$) [Tam 95].
Le deuxième type de masquage, plus surprenant à première vue, est le masquage arrière (*backward* en anglais) correspondant à une réduction de la visibilité d'images avant le changement de plan. Ce type de masquage est plus éphémère que le précédent puisque sa durée est de l'ordre de 10 millisecondes. Ce type de masquage provient certainement du temps de traitement nécessaire à notre système visuel pour intégrer les informations visuelles.
Des exemples d'implantation de masquage temporel sont décrits dans les articles [Girod 89] et [Watson 98].

2.3 Etat de l'art de la modélisation de l'attention visuelle pré-attentive

Confronté à la quantité d'information à traiter, le système visuel a développé des stratégies bien particulières. L'une d'elles est l'attention visuelle permettant de concentrer les ressources sensorielles sur des zones particulières de notre environnement visuel. La concentration de ces ressources de traitement est soit effectuée de façon volontaire soit de façon involontaire. Nous nous intéressons ici à l'attention visuelle pré-attentive, c'est à dire aux zones de notre environnement visuel qui attirent notre attention de façon "involontaire". Plusieurs types de modélisation de l'attention visuelle pré-attentive sont présentés dans la suite du document.

2.3.1 Modèles empiriques et statistiques

2.3.1.1 Modèles empiriques

Les modèles que nous appelons ici empiriques se basent sur des méthodes de traitements d'images, relativement éloignées des propriétés du SVH. Nous ne donnons ici que les caractéristiques principales des modèles les plus connus.
W. Osberger et A. Maeder [Osberger 98] déterminent la carte d'importance d'une image donnée en effectuant une segmentation de la source en régions homogènes au sens d'un certain critère. Une taxinomie des régions est ensuite réalisée. Cette hiérarchisation est basée sur une combinaison de critères très intuitifs tels que la taille, le contraste, la forme... Évidement, la qualité de cette taxinomie dépend fortement de la qualité de la segmentation. Dans le même ordre d'idée, J. Luo et A. Singhal [Luo 00] définissent un ensemble d'éléments visuels susceptibles d'attirer l'attention. Ces derniers sont extraits afin de piloter une segmentation favorisant l'apparition de régions d'intérêt. Enfin, citons les méthodes de X. Marichal et al. [Marichal 96] et de J. Zhao et al. [Zhao 96] qui dérivent des premières méthodes exposées. Ces différentes méthodes n'ont pas été validées via des expérimentations oculométriques.
Une autre étude, influente et très connue, est celle de C. Privitera et L. Stark [Privitera 00]. Ces derniers ont évalué la pertinence de 10 algorithmes de traitement d'images pour la détection de régions d'intérêt. L'évaluation de la pertinence se fait par comparaison des régions d'intérêt dites algorithmiques et humaines. Ils montrent que chacun des algorithmes testés est pertinent pour un ensemble restreint d'images et présente de mauvais résultats pour les autres images. Ce résultat suggère qu'il n'est pas envisageable d'utiliser un seul algorithme de traitement d'images pour prédire des régions d'intérêt de façon fiable.

2.3.1.2 Modèle statistique

Le modèle statistique le plus connu est celui proposé par A. Oliva et al. [Oliva 03]. Bien que, a priori purement statistique, ce type de modélisation utilise une propriété du SVH vérifiée par différents travaux et citée dans le paragraphe 1.4.3. Le principe consiste à dire que la capacité d'attraction d'une zone est inversement proportionnelle à sa probabilité d'apparition. En d'autres termes, notre attention visuelle est attirée par les zones en contraste avec leurs voisinages.
Dans cette approche, chaque site (pixel ou autre...) de l'image source se caractérise via un vecteur $v_l(x) = \{v_l(x,k)\}_{k=1,..,N}$ de dimension N obtenu via une décomposition hiérarchique. A partir de la propriété énoncée préalablement (la saillance est d'autant plus importante que les mesures locales sont incongrues), la carte de saillance S est obtenue en prenant l'inverse de la probabilité d'apparition d'un élément :

$$S(x) = \frac{1}{p(v_l)} \tag{2.18}$$

La probabilité $p(v_l)$ est approximée par une densité de probabilité Gaussienne :

$$p(v_l) = \frac{1}{(2\pi)^{N/2}\,|X|^{N/2}} exp\left\{-1/2(v_l-\mu)^T X^{-1}(v_l-\mu)\right\} \tag{2.19}$$

Dans l'article [Oliva 03], les auteurs montrent que les performances de ce modèle sont comparables à celles du modèle de L. Itti décrit dans la section suivante.

2.3.2 Modèle psycho-visuel

2.3.2.1 L'architecture de base [Koch 84],[Koch 85]

C'est en 1984, à partir des précédents travaux de A. Treisman et G. Gelade [Treisman 80] sur la théorie de l'intégration des caractéristiques visuelles (*Feature Integration Theory*), que C. Koch et S. Ullman [Koch 85] définirent le premier modèle d'attention visuelle basé sur une architecture biologiquement plausible. La figure 2.8 présente cette architecture devenue une référence.
A partir de l'image source à étudier, un ensemble de caractéristique visuelle pré-attentive est extrait. Les caractéristiques visuelles pré-attentives correspondent à des primitives visuelles précocement extraites par le système visuel humain. Il n'y a pas vraiment de liste exhaustive de ces caractéristiques mais il existe un consensus au sujet d'un certain nombre d'entre elles [Wolfe 04] : le contraste, l'orientation, la direction du mouvement...
Les caractéristiques pré-attentives peuvent être identifiées via des expériences psychophysiques en mesurant le temps de réaction nécessaire pour détecter une cible parmi un ensemble d'éléments perturbateurs (la figure 1.14 présente un exemple d'expériences psychophysiques et le paragraphe 1.4.2.1 définit la notion de temps de réaction).
D'une façon massivement parallèle, ces caractéristiques pré-attentives sont extraites générant ainsi des cartes de caractérisation respectant la topologie de la source. Des mécanismes d'inhibition latérale, simulant les champs récepteurs des cellules visuelles, permettent d'isoler les zones de l'image présentant des caractéristiques visuelles différentes de leurs voisinages. Ces mécanismes sont à apparenter à des détecteurs de contraste. Enfin, la carte de saillance, qui est l'un des caractères les plus novateurs de ces travaux, est construite en combinant les différentes cartes de caractérisation. La carte de saillance encode la saillance ou le pouvoir attracteur pour toutes les positions de la scène visuelle. La définition originelle donnée par C. Koch et S. Ullman est la suivante : la carte de saillance est une représentation de l'environnement accentuant les régions d'intérêt du champ visuel.

2.3.2.2 Le modèle de L. Itti [Itti 98]

Le modèle développé par L. Itti, C. Koch et E. Niebur est sans conteste le modèle le plus connu. Cette célébrité relative s'explique d'une part par le nombre de parutions scientifiques (dont [Itti 98, Itti 00a, Itti 00b, Itti 01a]) et d'autre part par le fait que les codes sources du modèle soient disponibles sur Internet.
Comme le montre la figure 2.9, ce modèle est basé sur l'architecture proposée par Koch et Ullman. Étant donné que ce modèle est indéniablement le modèle le plus utilisé, la bonne compréhension du fonctionnement de ce modèle est nécessaire. Par ailleurs, nous comparerons nos travaux avec ceux de L. Itti. Par conséquent, nous détaillons ses grandes étapes :

- la première étape consiste à créer trois canaux à partir d'une image (r, g, b) :
  1. un canal lié à l'intensité : $I = (r + g + b)/3$

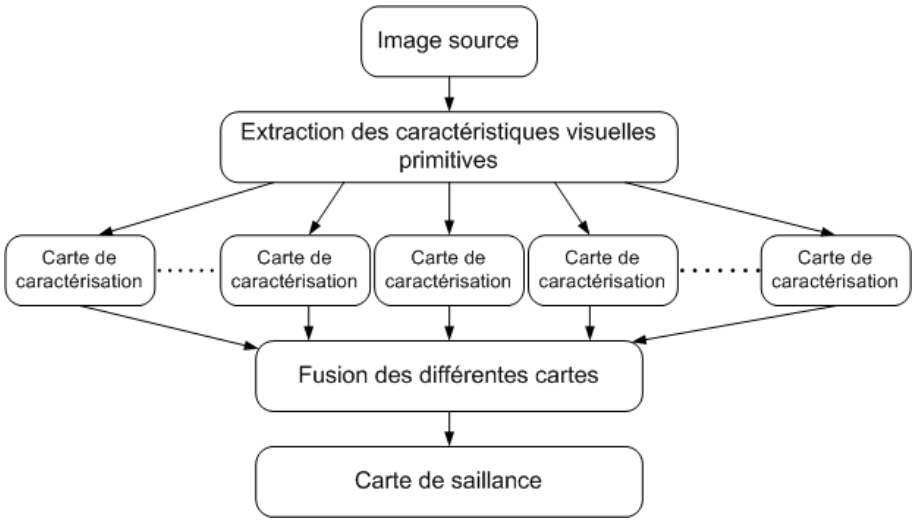

FIG. 2.8: Architecture biologiquement plausible d'un modèle d'attention visuelle proposée par C. Koch et S. Ullman [Koch 85].

2. un canal couleur composé de quatre composantes, lié à la théorie des couleurs antogonistes : rouge $R = r - (g + b)/2$, vert $G = g - (r + b)/2$, bleu $B = b - (g + r)/2$ et jaune $Y = (g + r)/2 - |r - g|/2 - b$.
3. un canal, dédié aux composantes orientées, est obtenu à partir du canal intensité I et d'une pyramide de Gabor orienté $O(\sigma, \theta)$, où σ indique le niveau de la pyramide et $\theta \in \{0, 45, 90, 135\}$ l'orientation exprimée en degré.

– une décomposition hiérarchique sur 9 niveaux via des pyramides Gaussiennes [Burt 83] est effectuée sur chaque composante. Ces pyramides sont censées représenter une approximation du pavage fréquentiel des cellules visuelles ;
– un mécanisme de centre / pourtour permet ensuite d'extraire des différents niveaux de la pyramide les informations pertinentes contrastant avec leur voisinage.
Pour le canal intensité, le contraste d'intensité est calculé en effectuant une différence entre les valeurs d'un niveau c de résolution fine avec un niveau s de résolution plus grossier :

$$\mathcal{I}(c, s) = |I(c) \ominus I(s)| \tag{2.20}$$

où, $c \in \{2, 3, 4\}$ et $s = c + \delta$ avec $\delta \in \{3, 4\}$. L'opérateur $\ominus$ représente l'opérateur de différenciation inter niveaux ; ce dernier réalise en fait une interpolation du niveau grossier afin de pouvoir faire une différence point à point entre deux niveaux de même résolution. Pour le canal de luminance, 6 cartes sont calculées ($\mathcal{I}(2, 5)$,$\mathcal{I}(2, 6)$,$\mathcal{I}(3, 6)$,$\mathcal{I}(3, 7)$,$\mathcal{I}(4, 7)$ et $\mathcal{I}(4, 8)$).
Pour le canal couleur, deux ensembles de cartes (12 au total) sont construits représentant respectivement la double opposition rouge/vert et vert/rouge, noté $\mathcal{RG}$ et la double opposition bleu/jaune et jaune/bleu, noté $\mathcal{BY}$:

$$\mathcal{RG}(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))| \tag{2.21}$$

$$\mathcal{BY}(c, s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))| \tag{2.22}$$

Pour le canal orientation, la convolution de la pyramide Gaussienne issue de la composante intensité I par la pyramide de Gabor orientée fournit 24 cartes :

$$\mathcal{O}(c, s, \theta) = |O(c, \theta) \ominus O(s, \theta)| \tag{2.23}$$

– une étape de normalisation, notée $\mathcal{N}(.)$, est menée sur chaque carte indépendamment les unes des autres pour créer une carte de saillance par canal. Dans [Itti 01b], trois méthodes sont proposées :
  1. la sommation naïve consiste à normaliser toutes les cartes sur la même dynamique et à moyenner la somme des différentes cartes ;
  2. l'amplification non linéaire globale dépendant du contenu de la carte consiste à multiplier la carte normalisée dans une dynamique $[0..M]$ par le facteur $(M - \overline{m})^2$. M représente le maximum global de la carte et $\overline{m}$ représente la moyenne de toutes les autres maximum locaux. Par conséquent, lorsque la carte normalisée contient uniquement quelques pics de saillance, ces derniers sont amplifiés. Par contre, lorsque la distribution de contraste est quasi uniforme, les valeurs de saillance sont atténuées ;

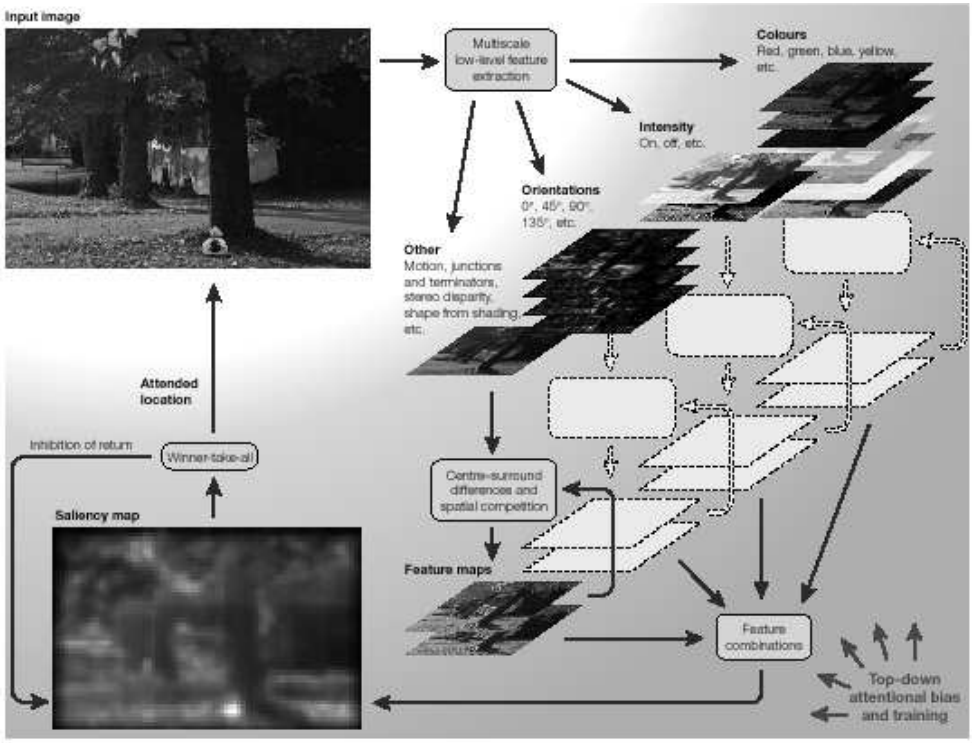

FIG. 2.9: Le modèle de L. Itti et C. Koch (extrait de [Itti 01a]).

  3. enfin, la dernière méthode de normalisation se veut plus proche de la réalité biologique en tentant d'une part de reproduire le comportement des champs récepteurs non classiques et d'autre part de modéliser les connections horizontales étendues (*long-range connections* en anglais).
     Tout d'abord, chaque carte est normalisée en utilisant le maximum global entre

zéro et un. Ensuite, chaque carte est filtrée via une différence de Gaussienne bidimensionnelle :

$$\mathcal{DOG}(x,y) = \frac{c_{ex}^2}{2\pi\sigma_{ex}^2} exp\left\{-\frac{x^2+y^2}{2\sigma_{ex}^2}\right\} - \frac{c_{inh}^2}{2\pi\sigma_{inh}^2} exp\left\{-\frac{x^2+y^2}{2\sigma_{inh}^2}\right\} \tag{2.24}$$

avec, $\sigma_{ex} = 0.02$ et $\sigma_{inh} = 0.25$ de la largeur de l'image, $c_{ex} = 0.5$ et $c_{inh} = 1.5$. La valeur C_{inh} est fixée à 0.02, si les cartes sont normalisées dans l'intervalle [0..1].
Le résultat est ajouté à la source et seules les valeurs positives sont considérées. La formulation analytique de cette normalisation est donnée ci-dessous :

$$\mathcal{M}^{k+1} = \left|\mathcal{M}^k + \mathcal{M}^k * \mathcal{DOG} - C_{inh}\right|_{\geq 0} \tag{2.25}$$

avec, $\mathcal{M}$ représente une carte de caractéristique. $\mathcal{DOG}$ représente la différence de Gaussienne 2D et l'opérateur $|.|_{\geq 0}$ annule les valeurs négatives. La valeur de k indique l'indice d'itération ; 10 itérations sont nécessaires pour obtenir un bon résultat.

- à partir des cartes normalisées, une carte de saillance est construite pour chaque canal. Ils suggèrent en fait que, pour un canal donné, toutes les cartes soient mises à l'échelle du niveau 4 de la pyramide et cumulées point à point. Ils obtiennent ainsi $\overline{\mathcal{I}}$, $\overline{\mathcal{C}}$ et $\overline{\mathcal{O}}$ respectivement pour la carte de saillance en intensité, en couleur et en orientation ;
- finalement, la carte de saillance finale $\mathcal{S}$ est obtenue de la façon suivante :

$$\mathcal{S} = \frac{1}{3}\left(\mathcal{N}(\overline{\mathcal{I}}) + \mathcal{N}(\overline{\mathcal{C}}) + \mathcal{N}(\overline{\mathcal{O}})\right) \tag{2.26}$$

D'après les travaux de L. Itti, ce sont les deux dernières méthodes qui offrent les meilleurs résultats.

La figure 2.10 donne un exemple des cartes de saillance calculées sur une image. Bien que ce modèle donne en règle générale de bons résultats, il reste critiquable sur plusieurs points. Tout d'abord, de nombreuses valeurs ne sont pas véritablement justifiées : le passage de RGB vers un espace de couleurs antagonistes est-il issu de tests particuliers ? Comment sont déterminées les valeurs de σ_{ex}, σ_{inh}, c_{ex}, c_{inh} définissant la DoG ? Outre ces aspects, le point le plus controversé, et celui sur lequel nous souhaitons nous attarder, concerne le calcul de la carte de saillance d'un canal. En effet, pour déterminer cette carte de saillance, tous les niveaux de la pyramide doivent avoir la même résolution que celle du niveau 4. Par ailleurs, en dépit du fait que la perception d'une image dépend fortement de la distance de visualisation séparant l'observateur de la source, la distance de visualisation n'est pas un paramètre clairement explicité du modèle. Enfin, notons pour conclure qu'il faut au moins 3 étapes (au plus 7) de normalisation pour aboutir à la carte de saillance finale. Il est clair que la grande difficulté de la modélisation de l'attention visuelle concerne la combinaison de caractéristiques visuelles de natures différentes. L'obtention d'une valeur de saillance sous le forme d'un scalaire doit bien évidemment passer par des étapes de normalisation. Dans ce modèle, trois étapes de normalisation sont déjà nécessaires pour calculer la carte de saillance d'une seule caractéristique, sachant qu'aucune interaction inter caractéristique visuelle n'est envisagée.

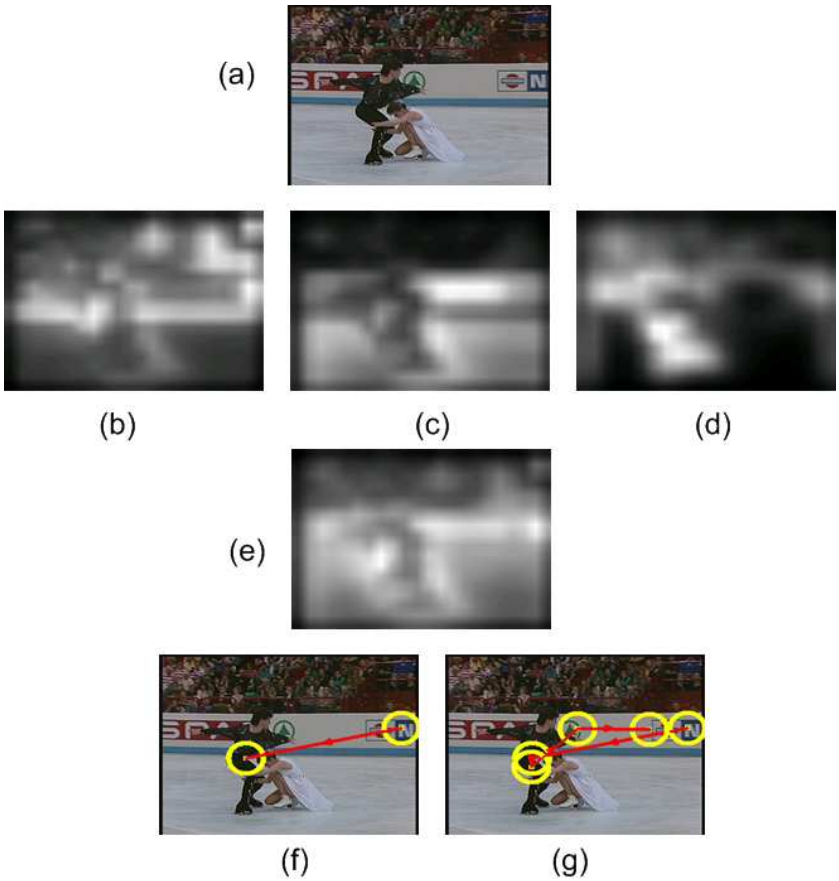

FIG. 2.10: Exemple d'obtention d'une carte de saillance sur une image donnée (a) ; les cartes de saillances couleur, intensité et orientation sont données respectivement en (b), (c) et (d) ; la carte de saillance finale est donnée en (e) alors que (f) et (g) représentent respectivement les deux et les cinq points de fixation les plus saillants.

2.3.2.3 Autres modèles psycho-visuels

Le modèle de R. Milanese [Milanese 92, Milanese 93], suit l'architecture de C. Koch et S. Ullman. Des caractéristiques visuelles sont d'abord extraites de l'image analysée : on retrouve les classiques canaux intensité et couleur, à ces derniers viennent s'ajouter l'amplitude des contours et la courbure. L'ensemble de ces caractéristiques est ensuite filtré par un filtre passe bande orienté. L'intérêt de ce modèle réside essentiellement dans la façon de construire la carte de saillance finale à partir des cartes de saillance de chaque canal. L'algorithme de fusion défini par R. Milanese doit répondre aux principes suivants :

1. la carte de saillance finale doit être un "résumé"[1] de toutes les cartes de saillance intermédiaires ;
2. la carte de saillance doit isoler clairement les régions perceptuellement importantes. Idéalement, elle doit être binaire ;
3. les régions perceptuellement importantes doivent correspondre aux objets les plus

[1] ce sont les termes exactes utilisés par l'auteur (page 117 de [Milanese 93]).

saillants physiquement et sémantiquement. Les composantes de chaque objet sont de moindre importance ;

4. les formes des régions détectées doivent être compactes et approximer la forme d'un objet.

Le mécanisme de fusion, envisagé par R. Milanese, est basé sur la moyenne des cartes de saillance. Bien évidemment, ce mécanisme ne satisfait que la première condition. R. Milanese introduit donc un ensemble de prétraitement pour satisfaire les autres conditions. Le premier consiste à appliquer un filtre passe bas isotrope Gaussien sur les différentes cartes de saillance afin de répondre aux exigences (3) et (4). L'application d'un filtre passe bas permet de lisser l'information et de réduire l'écart entre deux pics de saillance. Ensuite, une procédure de relaxation est effectuée en utilisant la descente de gradient d'une fonction énergétique. Cette dernière est définie par la combinaison de plusieurs mesures d'incohérences :

- l'incohérence inter-carte représente les incohérences locales des différentes cartes de saillance ;
- l'incohérence intra-carte représente la mesure de la compacité de la carte (pour répondre aux exigences (3) et (4)). Cette mesure est minimisée pour une carte de saillance uniforme ;
- l'activité globale est une mesure permettant de pénaliser les cartes présentant une très forte activité.

Après la phase d'optimisation, la carte de saillance est seuillée pour former une carte binaire. Notons également que ce modèle intègre une mémoire des positions des lieux saillants visités afin de piloter un mécanisme d'inhibition de retour.

Le modèle de A. Chauvin [Chauvin 00, Chauvin 02] se différencie de l'implantation proposée par L. Itti sur plusieurs points. L'extraction des primitives visuelles pré-attentives s'effectue à partir d'une base de 32 ondelettes de Gabor couvrant 4 bandes de fréquences et 8 orientations. Une étape de normalisation de type inhibition par division (plus connu sous le terme anglais *divise inhibition*) des différents canaux obtenus est réalisée par bandes de fréquences. Ensuite, afin de modéliser les modulations facilitatrices renforçant les contours alignés et colinéaires, les canaux sont filtrés par le produit d'une gaussienne par un masque en forme de papillon[2]. Enfin, l'application itérative d'une différence de deux Gaussiennes sur la somme des sous bandes d'orientations différentes mais accordées à la même fréquence permet de favoriser les orientations les plus fortes au détriment des autres. La carte de saillance est calculée à partir d'une combinaison linéaire des différentes sous bandes fréquentielles. Basé sur de nombreux aspects psycho-visuels, on pourra regretter que ce modèle ne traite que des images achromatiques.

Le modèle de B. Bruce et E. Jernigan [Bruce 03] inspiré du modèle de L. Itti a la particularité d'intégrer dans une architecture classique de type C. Koch et S. Ullman un opérateur statistique basé sur le même principe que celui utilisé par A. Oliva et al [Oliva 03]. Ce principe, énoncé auparavant, lie la saillance d'un évènement à sa probabilité d'apparition. B. Bruce et E. Jernigan utilisent la mesure d'incertitude (c'est à dire l'espérance mathématique de l'entropie) pour définir la saillance d'un évènement.

[2]Nous aurons l'occasion dans la suite de présenter plus en détails ce type de filtre

Enfin, citons pour finir **le modèle proposé par R. Canosa** [Canosa 03]. Ce modèle est fondé sur la détermination de quatre cartes d'importances. Nous retrouvons les cartes d'importances classiques en intensité (directement obtenue via la composante achromatique), en couleur (obtenue via une simple combinaison des canaux chromatiques) et en orientation (obtenue via une décomposition hiérarchique de la carte d'importance en intensité et via des filtres orientés passe bandes de Gabor). L'intérêt de ce modèle réside dans la quatrième carte d'importance. Construite à partir de la carte d'importance en intensité, cette dernière est utilisée pour déterminer et localiser des objets sémantiques contenus dans l'image. La première étape consiste à partitionner l'image à analyser en régions de premier plan et d'arrière plan. Ensuite, des régions caractérisées par une faible variation de luminance et appartenant au premier plan sont identifiées. Cette carte d'importance, localisant les objets, est utilisée lors du calcul de la carte d'importance finale. Notons que la carte finale est obtenue via une simple combinaison linéaire des différentes cartes.

2.3.3 La dimension temporelle dans la modélisation

La modélisation de l'attention visuelle sur des scènes dynamiques n'a pas encore fait l'objet de nombreuses études. Il n'y a guère que les articles [Dhavale 03],[Yee 01] et la thèse de D. Parkhurst [Parkhurst 02] qui abordent l'intégration de cette dimension. Notons, tout de suite, que les modèles spatio-temporels proposés dans les articles [Dhavale 03],[Yee 01] sont une extension relativement simple du modèle de L. Itti. D'ailleurs, si on se reporte à la figure 2.9 extraite de [Itti 01a], on constate que L. Itti avait déjà suggéré une généralisation possible de son modèle à de nombreuses dimensions visuelles. Évidemment, cette extension au temporel souffre des mêmes défauts que ceux du modèle spatial donnés préalablement. La carte de saillance temporelle est, en effet, déterminée de la même façon que celle décrite au paragraphe 2.3.2.2.
Concernant le modèle proposé par D. Parkhurst dans le chapitre 7 de sa thèse[3], le traitement de l'information temporelle est intéressante car biologiquement plausible. Nous avons déjà abordé dans le paragraphe 2.2.7 la façon dont D. Parkhurst déterminait les caractéristiques temporelles pertinentes. L'utilisation de deux canaux, un canal dit *sustained* et l'autre dit *transient*, modélise respectivement les fréquences temporelles faibles et élevées. A partir du canal *transient*, une carte de saillance est déterminée en appliquant les mêmes procédés que ceux de L. Itti. On pourra regretter qu'une comparaison quantitative n'apparaisse pas dans ses travaux. Seule une comparaison qualitative succincte a été effectuée (page 166 de la thèse de D. Parkhurst).

2.3.4 Discussion

Les différents modèles d'attention visuelle sont pour la plupart basés sur l'architecture de C. Koch et S. Ullman. Le plus célèbre est celui de L. Itti qui offre des résultats pertinents. Au même moment, A. Chauvin développait son modèle. Son défaut majeur réside certainement sur le fait que la couleur ne soit pas traitée. Le fait de se contenter de

[3] D. Parkhurst a construit un modèle biologiquement plausible. Cette approche se base exactement sur l'architecture du modèle de L. Itti. La différence se situe dans la définition et la justification des étapes de traitement de l'information visuelle.

la luminance pour déterminer les régions visuellement importantes réduit les performances et contourne la difficile étape de fusion de cartes provenant d'origines différentes. L. Itti a proposé des méthodes de fusion dont une tout à fait intéressante. La méthode de fusion itérative, décrite précédemment, reproduit l'aspect temporel de la vision. Si on considère une seule itération (un temps d'observation faible), les zones saillantes n'apparaissent pas de façon évidente. La carte de saillance représente une information grossière de la scène et permet simplement de catégoriser la scène. Si on se réfère aux expériences de catégorisation menées par N. Guyader [Guyader 03], il suffit de quelques centaines de milisecondes à un observateur pour déterminer la famille d'appartenance d'une image. Par contre, lorsque le nombre d'itérations augmente (lorsque le temps de visualisation augmente), les zones saillances se détachent progressivement du fond. Plus le nombre d'itérations sera grand et plus le nombre de points saillants diminue.
En dépit de l'aspect intéressant de la méthode proposée par L. Itti, la fusion reste un problème paraissant mal abordé, particulièrement lorsqu'il y a plusieurs cartes de saillance (luminance, orientation, couleur...). Ce problème nuit aux performances actuelles des modèles psycho-visuels. Rappelons que, dans l'article [Oliva 03], les performances du modèle de L. Itti sont égalées par un modèle statistique. Notons également que dans le cadre de la reconnaissance d'objet, le modèle statistique couplé avec des informations contextuelles (liées à la nature de la scène), améliore significativement les performances [Oliva 01], [Torralba 03].
Seul R. Milanese [Milanese 93] a proposé une méthodologie de fusion qui nous semble très pertinente mais peu exploitée actuellement. Nous proposons donc dans ce travail une méthode de fusion originale s'inspirant des recherches de R. Milanese et des différents travaux (modèle psycho-visuel et statistique).
Très peu de travaux font référence à la dimension temporelle. Dans cette étude, une des contributions consiste à considérer la dimension temporelle dans la modélisation de l'attention visuelle.

Si on fait abstraction de la façon dont le modèle est construit, deux problèmes majeurs se posent : le problème de la référence avec laquelle la modélisation doit être comparée et le problème de l'évaluation des performances.

- le problème de la référence : aujourd'hui, il n'y aucun consensus sur la façon de créer une référence avec laquelle une comparaison serait possible. Le tableau 2.1 donne deux exemples de protocoles de tests oculométriques visant les mêmes objectifs. Des différences majeures sont notables : le nombre de participants, l'appareillage, le format des images... Par ailleurs, ce tableau ne donne que le protocole expérimental. Il reste encore à déterminer comment construire la référence (la vérité terrain) à partir des informations collectées lors de tests ;
- le problème de l'évaluation des performances : il n'y a guère de consensus s'agissant de l'évaluation des performances. On distingue deux types de mesures : les mesures globales et mesures locales basées sur la position spatiale des points de fixation. Pour illustrer ces deux types de mesures, nous nous basons sur les travaux de A. Chauvin et de D. Parkhurst :
  - les mesures globales : A. Chauvin utilise le coefficient de corrélation entre ses prédictions et les cartes de saillance provenant des observateurs. Il compare les valeurs de corrélations obtenues lorsque différentes méthodes de prédiction (le

TAB. 2.1: Protocole expérimentale d'études oculométrique.

	A. Chauvin [Chauvin 02]	D. Parkhurst [Parkhurst 02]
nombre de participants	60	4
nombre d'images	72	300
durée d'affichage	3s	5s
consigne	catégorisez la scène	explorez l'image
appareillage	casque (Eyelink)	ISCAN RK-416 (tête fixe)
précision	1°	1.7°
distance de visualisation	57cm	58cm
écran	21 pouces	17 pouces
angle de visualisation	21°	$30 \times 24^\circ$
échantillonnage	4ms	16ms

modèle développé, une carte uniforme, des détecteurs de contours...) des régions d'intérêt sont utilisées. La meilleure valeur de corrélation (0.33) est de très loin celle obtenue avec le modèle proposé. Lors de cette évaluation, A. Chauvin pointe un problème majeur inhérent au calcul du coefficient de corrélation. Ce dernier est, en effet, calculé entre deux distributions fondamentalement différentes : la carte de saillance prédite est relativement lisse alors que la carte de saillance provenant des observateurs est à bords francs. Mis à part ces travaux, il n'y a pas à notre connaissance d'autres évaluations globales de la similarité globale entre une vérité terrain et une prédiction ;

– les mesures locales basées sur la position spatiale des points de fixation : la méthode locale utilisée par A. Chauvin mesure les différences statistiques autour des fixations sélectionnées. Ces statistiques sont ensuite comparées aux résultats de P. Reinagel [Reinagel 99], préalablement évoqués dans le paragraphe 1.4.3. Les résultats obtenus sont conformes à ceux de P. Reinagel et A. Zador. D. Parkhurst, quant à lui, effectue la somme des saillances prédites extraites aux coordonnées spatiales des points de fixation des observateurs. Cette mesure est appelée fonction de probabilité cumulée. Nous aurons l'occasion de définir plus en détail cette mesure lors de l'évaluation quantitative des performances de notre modélisation.

Dans notre contexte d'étude (détection des zones saillantes pour des images projetées sur un moniteur), l'évaluation des performances est un problème majeur qui nécessite l'utilisation ou la définition d'une métrique globale de performances.

2.4 Conclusion

La vocation de ce chapitre était de présenter des modèles mathématiques simulant les propriétés du système visuel humain. Les propriétés, dite de bas niveau, peuvent être modélisées à partir d'expériences psychophysiques. L'exemple de la modélisation de la sensibilité aux contrastes (CSF) est certainement le plus pertinent. A partir d'expériences, basées sur des stimulations maîtrisées (des signaux simples la plupart du temps), il

est possible de déterminer le seuil de visibilité d'une composante visuelle parfaitement caractérisée (fréquence, orientation, ...). En faisant varier les paramètres de la stimulation, des fonctions de sensibilité aux contrastes sont obtenues. Néanmoins, ces courbes sont pertinentes que pour des signaux simples. Leurs applications à des scènes complexes, bien qu'elles soient largement usitées, introduisent un biais souvent négligé. Il est, en effet, clair que le contexte influe largement sur le seuil de visibilité. Cette influence est appelée masquage visuel. La bonne maîtrise de ces modèles bas niveau est essentielle pour tenter de reproduire le comportement du système visuel.
La modélisation de l'attention visuelle est un domaine en plein essor. L'intégration de propriétés de bas niveau dans la modélisation, accentuant le caractère mimétique, est la clé du succès. L. Itti a ouvert une porte sur la modélisation de l'attention visuelle, en proposant un modèle compétitif. La critique constructive de ce modèle et des autres permet d'entrevoir des pistes d'améliorations. Comme nous l'avons souligné, il n'existe pas de modèle intégrant d'une part une fusion cohérente et pertinente de cartes de saillance provenant de différentes dimensions et d'autre part la dimension temporelle. La seconde partie traite de ces problèmes.

Deuxième partie

Modèle cohérent de l'attention visuelle

Chapitre 1

Expérimentations oculométriques

1.1 Introduction

L'objet de ce chapitre est de présenter les expériences oculométriques réalisées à la fois sur des images fixes et sur des séquences d'images. Un oculomètre est un appareil mesurant le déplacement de l'oeil. Ce type d'expérimentations est fondamental pour le bon déroulement de nos travaux. Elles nous ont permis de construire pour chaque image fixe ou chaque séquence une référence, que nous appelons vérité terrain. Les données collectées peuvent être exploitées de différentes façons. Les zones attirant le regard des observateurs sont, tout d'abord, identifiables. La durée de fixation est un élément intéressant pour mesurer le degré de saillance de chacune des zones fixées. Une nuance doit être toutefois être apportée : la durée de fixation n'est pas proportionnelle à la saillance. Elle en dépend certes, mais dépend aussi de la complexité de la zone en terme d'indices visuels. Une zone complexe nécessite plus de temps pour être traitée qu'une zone simple. Néanmoins, leurs saillances peuvent être identiques.
Par ailleurs, la stratégie visuelle, c'est à dire le déplacement oculaire, peut faire l'objet d'études, même si cela semble a priori difficile à aborder du fait de l'idiosyncrasie de la stratégie visuelle ; la façon d'explorer une scène étant propre à chaque individu.
Nous décrivons, dans la première partie de ce chapitre, le dispositif oculométrique ainsi que ses spécificités. Les protocoles de mesure pour les tests sur images fixes et sur les séquences d'images sont ensuite abordés. La construction des cartes de saillance pour chaque observateur et la construction d'une carte de saillance qualifiant la stratégie d'un observateur moyen (c'est à dire représentatif d'un large panel d'observateurs) sont détaillées. Ce chapitre est conclu par une discussion concernant les avantages et inconvénients de telles expériences.

1.2 Le dispositif oculométrique

Comme introduit précédemment, un dispositif oculométrique est un système permettant d'étudier la stratégie visuelle d'un observateur. Excepté le relatif inconfort lié à l'expérimentation, le dispositif oculométrique ne produit aucune gêne. La vision des obser-

FIG. 1.1: Dispositif oculométrique utilisé.

vateurs n'est pas perturbée. L'appareil utilisé lors de nos expérimentations[1] est basé sur la capacité des yeux à réfléchir les infrarouges. En fait, deux types de reflets sont observés : des reflets fixes dûs à la réflexion des infrarouges sur la cornée (reflets de Purkinje), et des reflets mobiles dûs à la réflexion des infrarouges sur la pupille. La position relative de ces deux types de reflets permet de déterminer le positionnement de l'oeil.

Ce dispositif, présenté figure 1.1, est composé :

- d'un support horizontal réglable en hauteur, sur lequel l'observateur viendra poser son menton ;
- d'une sangle transversale sur laquelle l'observateur appuie son front. Ces précautions augmentent la précision des mesures en empêchant des mouvements de la tête de l'observateur ;
- d'un miroir infrarouge et de deux sources infrarouges ;
- d'une caméra infrarouge filmant l'oeil reflété sur le miroir.

Les spécifications techniques du dispositif oculométrique sont données dans le tableau 1.1.

TAB. 1.1: Spécifications techniques de l'oculomètre utilisé.

Spécifications techniques	
Technique de mesure	pupille et réflection cornéenne (Purkinje)
Fréquence d'échantillonnage	$50Hz$
Résolution	0.1°
Précision	$0.25 - 0.5^\circ$
Excursion horizontale, verticale	$\pm 40^\circ$, $\pm 20^\circ$
Mouvement de tête autorisé	$\pm 10mm$

[1]Dispositif de la société *Cambridge Research System*. L'adresse du site internet est la suivante *www.cvsltd.com*.

TAB. 1.2: Protocole des expérimentations pour l'acquisition de données oculométriques sur images fixes.

Protocole des expérimentations	
Distance d'observation	$4H$: H hauteur de l'écran de visualisation
Résolution de l'écran	800×600
Nombre d'images traitées	40
Type d'images	niveaux de gris et couleur
Nombre d'observateurs	40
Durée de l'observation	$15s$
Calibrage	20 points de calibrage

1.3 Expérimentations sur images fixes

1.3.1 Le protocole

Le protocole des expérimentations oculométriques que nous avons menées est donné dans le tableau 1.2.

L'expérimentation oculométrique mise en place consiste à enregistrer les mouvements oculaires d'un observateur naïf regardant un ensemble d'images pendant une durée d'observation de 15 secondes chacune[2]. Aucune indication n'est donnée à l'observateur, si ce n'est de regarder le moniteur d'affichage de la façon la plus naturelle possible ; nous sommes dans un contexte de *free viewing task*. L'objectif étant de modéliser l'attention visuelle de type *Bottom-Up*, le contenu des images à visualiser ne doit pas être connu des participants aux tests.

Le dispositif oculométrique enregistre la position de l'oeil tous les 20 ms (fréquence d'échantillonnage de $50Hz$). Ainsi, pour un observateur, on dispose de 750 points pour une durée d'observation maximale. Après chaque image, une phase de calibrage est effectuée afin de limiter d'éventuelles dérives (perte de précision, mauvais calibrage...).

La phase de calibrage de l'appareil est primordiale. Elle consiste à afficher séquentiellement un certain nombre de points positionnés aléatoirement sur l'écran (à un instant donné il y au maximum un point). L'observateur les fixe les uns après les autres. La correspondance entre les points à fixer et les zones fixées par l'observateur est donnée par le logiciel. Sur la figure 1.2, deux résultats de calibrage sont donnés. Au centre, le calibrage est bon. A droite, le calibrage est à refaire. La correspondance entre les points à fixer et les zones fixées par l'observateur est mauvaise pour les deux points en bas à droite. Par ailleurs, chaque enregistrement a été visualisé pour détecter des comportements oculaires atypiques, des problèmes de synchronisation ou encore des problèmes d'enregistrement dus à la morphologie de l'oeil (la paupière supérieure venant cacher une partie de la pupille). Les données oculométriques présentant ce type de problème ont été supprimées.

[2]Il est nécessaire de distinguer ici l'expérimentation de l'exploitation des données. Cette dernière se fera avec des durées d'observation variables, allant de une seconde à plusieurs secondes. L'étude de l'évolution temporelle des stratégies visuelles est envisageable.

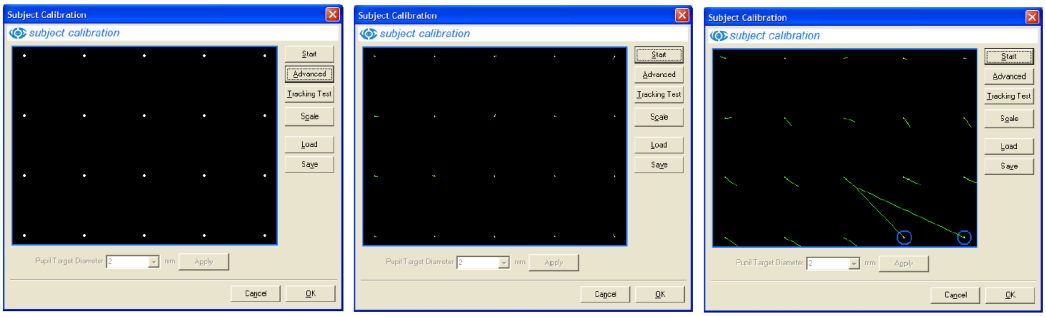

FIG. 1.2: Étape de calibrage de l'oculomètre. A gauche, tous les points à fixer sont affichés. Au milieu, le résultat d'un calibrage satisfaisante est donné. A droite, un exemple de mauvais calibrage est donné.

FIG. 1.3: Suppression des données relatives aux mouvements oculaires de saccade.

1.3.2 Des mesures à la saillance spatiale

1.3.2.1 Prétraitement des informations sources

Suite à ce type d'expérimentations, nous disposons pour chaque image de 40 enregistrements de mouvements oculaires indépendants provenant des 40 observateurs. Avant de fusionner ces données, il est nécessaire d'effectuer un prétraitement sur chaque enregistrement. En effet, les données relatives à des mouvements oculaires de saccade sont à supprimer. Ce prétraitement est obtenu via l'algorithme 1 : il consiste en fait à déterminer dans un voisinage donné de taille $NB\_VOISINS$ le nombre de points fixés. Si ce nombre est inférieur à un seuil NB_POINTS, la donnée courante est considérée comme une donnée fugitive probablement acquise lors d'une saccade. Un exemple est donné à la figure 1.3.

Les constantes $NB\_VOISINS$ et NB_POINTS ont été fixées afin de répondre à deux exigences. La première concerne la précision de l'oculomètre : 0.25 à 0.5 degré. Étant donné que la distance d'observation est de $4H$ et que la résolution de l'écran est de 800×600, un degré visuel correspond à environ une zone de 40 pixels. Par conséquent, si on considère une précision de l'ordre de 0.25 degré, la valeur 10 pour la constante $NB\_VOISINS$ est appropriée. Concernant la constante NB_POINTS prenant la valeur 5, on considère qu'une fixation a lieu dans une zone de 0.25 degré centrée sur le pixel courant si on

Algorithme 1 Prétraitement des données oculométriques : suppression des données liées aux saccades.

```
1:  // Initialisation
2:  NB_VOISINS = 10
3:  NB_POINTS = 5
4:  pour j ← 0, 1, 2, ..., K faire
5:     // K est le nombre de pixels à traiter
6:     nbFixation = 0
7:     // Le pixel est-il fixé (valeur = 1) ou non fixé (valeur = 0)
8:     si pixel[j] == 1 alors
9:        pour i parcourant une fenêtre carrée de NB_VOISINS · NB_VOISINS faire
10:          si pixel voisin de pixel[j] == 1 alors
11:             // Le nombre de fixations est incrémenté du nombre de fixations enregistrées
                sur le pixel j
12:             nbFixationVoisinage+ = nbreFixation[j]
13:          fin si
14:       fin pour
15:       si nbFixationVoisinage < NB_POINTS alors
16:          pixel[j] = 0
17:       fin si
18:    fin si
19: fin pour
```

dénombre au moins 5 points (soit une durée de fixation minimale de 100 ms sur la dite zone).

1.3.2.2 Création d'une carte de fixation pour l'observateur moyen

A partir de l'ensemble des données expérimentales (pour chaque image on dispose de 40 enregistrements oculométriques), le comportement moyen d'un observateur peut être identifié. Il suffit, en effet, de cumuler en chaque position spatiale de l'image les fixations provenant des différents enregistrements. En normalisant chaque point par le nombre d'observateurs, une carte de saillance, que nous noterons par la suite CS, est obtenue au sens de la définition donnée par C. Koch et S. Ullman [Koch 85]. En fait, la valeur de chaque pixel de la carte représente la capacité d'attraction ou encore l'attractivité perceptuelle du site spatial concerné ; plus cette valeur est élevée, plus l'attractivité du site est importante.
La carte de fixation pour un observateur k est donnée par la relation (1.1) :

$$CS^k(x, y) = \sum_{i=1}^{M} \Delta(x - x_i, y - y_i) \tag{1.1}$$

avec, M le nombre total de fixations par image et (x_i, y_i) les coordonnées de la fixation i. Δ est le symbole de Kronecker.

La carte CS est alors donnée par la relation (1.2) :

$$CS(x,y) = \frac{1}{N}\sum_{k=1}^{N} CS^k(x,y) \tag{1.2}$$

avec CS^k la carte de saillance de l'observateur k et N le nombre d'observateurs.

1.3.2.3 Création d'une densité de saillance pour l'observateur moyen

La carte de saillance est bien évidemment intéressante mais elle ne représente pas vraiment la réalité. Tout d'abord, il semble évident que l'oeil ne fixe pas un point sur une image mais une zone ayant une taille visuelle proche de celle de la fovéa. En outre, les cartes de saillance sont obtenues à partir d'expérimentations faisant intervenir un appareillage à la précision limitée. En effet, si on se réfère de nouveau à ses caractéristiques (cf. tableau 1.1), la précision de l'oculomètre est comprise entre 0.25 et 0.5 degré de champ visuel. A partir de ces deux considérations, les densités de saillance DS sont obtenues par convolution des cartes de saillance CS avec un filtre gaussien bi-dimensionnel :

$$DS(x,y) = CS(x,y) * g_{\sigma_x,\sigma_y}(x,y) \tag{1.3}$$

g_{σ_x,σ_y} une gaussienne bi-dimensionnelle donnée par

$$g_{\sigma_x,\sigma_y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2}\right\} \tag{1.4}$$

Les écarts-types σ_x et σ_y ont été pris à une valeur de 0.5 degré.

Remarques :

Deux autres approches de construction de cartes de saillance ont été envisagées. Comme ces méthodes introduisent un biais (elles n'utilisent pas toutes les données collectées), elles n'ont pas été retenues. En dépit de cela, il est intéressant de les présenter pour sensibiliser le lecteur à l'importance de la méthode de construction de la vérité terrain.
La première est une simplification de la construction de la carte CS précédemment décrite. La carte CS était issue d'une accumulation de toutes les données. La modification, proposée ici, consiste à considérer une saillance maximale, notée $Seuil_{max}$, limitant les valeurs de saillance. Cette nouvelle carte, notée CS', est alors obtenue par la relation suivante :

$$CS'(x,y) = min(CS(x,y), Seuil_{max}) \tag{1.5}$$

La carte de densité est ensuite obtenue de la même façon que précédemment. La qualité de la carte obtenue dépend fortement du seuil $Seuil_{max}$. Lorsque ce dernier est très élevé, les valeurs de la carte CS' tendent vers les valeurs de la carte CS. A l'inverse, leur dissimilarité augmente avec la diminution du seuil. Par ailleurs, lorsque ce seuil est faible, les mesures de saillance obtenues sont potentiellement perturbées par les données relatives aux fixations intermédiaires et transitoires. Malgré cet inconvénient, l'utilisation d'un seuil faible permet de se faire visuellement une idée du taux de couverture. Enfin, pour certaines applications, l'utilisation d'un seuil est utile pour classer les pixels en deux catégories (pixel de zone

TAB. 1.3: Estimation sur la durée du test des paramètres moyens suivants : durée de fixation (en ms), nombre de fixations par seconde et le nombre de saccades par seconde et ceux de A. Chauvin [Chauvin 02].

	Nos expérimentations	Chauvin [Chauvin 02]
Durée de fixations (en ms)	310	295
Nombre de fixations (par s)	2.7	3.4
Nombre de saccades (par s)	2.9	3

d'intérêt ou pixel de zone inintéressante).
La seconde méthode considère les données collectées par l'oculomètre dans l'ordre temporel d'arrivée. Ces données forment une suite temporelle, elle même constituée d'une succession de deux sous-suites. Elles représentent respectivement les phases de fixations et les phases de saccades. L'idée consiste alors à considérer chaque sous-suite de fixations comme une et une seule fixation. Cette méthode est donc insensible à la durée de fixation, ce qui dans notre étude, n'est pas été retenue. La durée de fixation n'est en effet pas directement proportionnelle, elle reste néanmoins l'un des meilleurs indicateurs du degré de saillance.

1.3.3 Résultats sur images fixes

1.3.3.1 Estimation des comportements oculaires moyens

Afin de vérifier la cohérence des données acquises, un certain nombre de paramètres sont estimés. Le tableau 1.3 donne la durée moyenne de fixation en milliseconde, le nombre moyen de fixations ainsi que le nombre moyen de saccades par seconde. Ces données sont comparées avec celles obtenues par A. Chauvin [Chauvin 02].
Les estimations moyennes obtenues sont cohérentes avec celles de A. Chauvin. Elles sont également en accord avec celles de la littérature [Henderson 99a].

1.3.3.2 Densité de saillance pour l'observateur moyen

La figure 1.4 présente sur la colonne de gauche trois densités de saillance obtenues pour un observateur moyen et pour trois durées d'observation (2, 8 et 14 secondes). Une seconde façon de représenter les zones attirant l'attention visuelle est également présentée sur la colonne de droite. Cela consiste à reproduire chaque pixel de l'image source en fonction de son degré de saillance. Une zone sombre représente une zone peu attractive alors qu'une zone éclairée est saillante. De plus amples résultats sont donnés dans l'annexe A.

1.3.3.3 Taux de couverture de l'image

Le taux de couverture de l'image a été défini par D. Wooding [Wooding 02] dans les termes suivants : *la couverture est une mesure de la quantité du stimulus original couvert par des phases de fixation.* Dans ce genre d'approche, un seuil, noté T, est nécessaire pour classer les zones de l'image comme ayant été soit fixées soit non fixées. Le rapport entre le nombre de pixels fixés et le nombre de pixels total donne alors le taux de couverture. Le tableau 1.4 présente le taux de couverture pour une durée d'observation donnée, pour un seuil donné et pour différents types d'images. A partir des données du tableau 1.4, deux

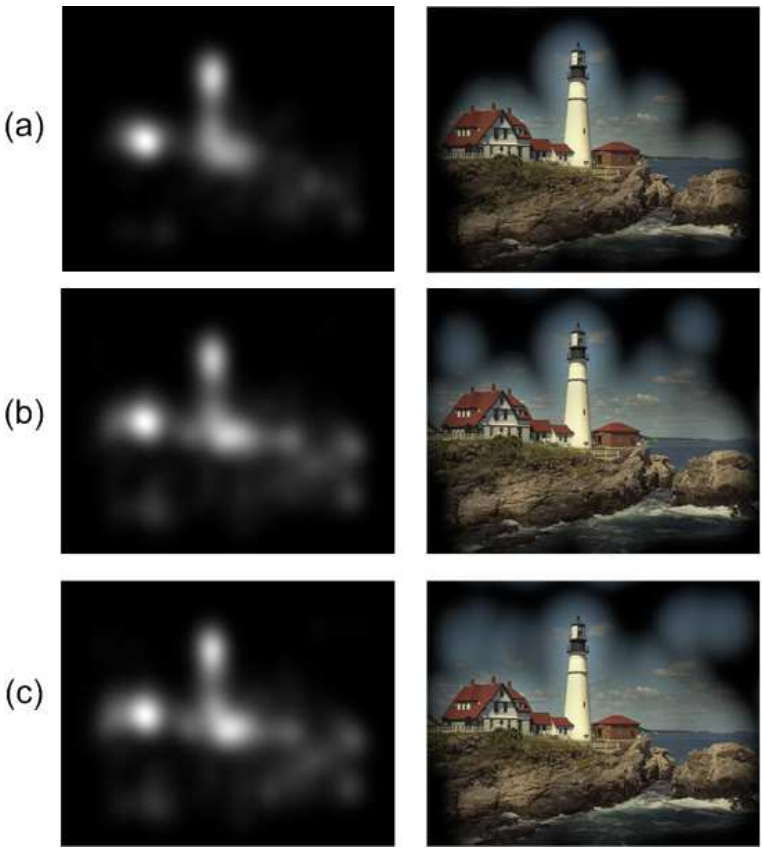

FIG. 1.4: Densités de saillance humaine (a), (b) et (c) obtenues respectivement pour les temps d'observation 2, 8 et 14 secondes. Exemple de l'image *Lighthouse*

TAB. 1.4: Évolution du taux de couverture en fonction de la durée d'observation.

Durée d'observation	t=2s	t=8s	t=14s
couverture (%),$T = 25$			
KayakCouleur0	7.4	32.7	46.3
Rapids	13.3	30.8	39.6
ChurchAndCapitol	17.8	46	57.8
couverture (%),$T = 50$			
KayakCouleur0	5.7	20.1	30.2
Rapids	9.2	20.9	28.7
ChurchAndCapitol	11.3	35.6	46.7
couverture (%),$T = 75$			
KayakCouleur0	4.6	13	21.8
Rapids	6.6	17.9	23.1
ChurchAndCapitol	7.2	29.7	40.2

constats sont faits :

- le premier constat est évident : plus le seuil T augmente et moins le taux de couverture est important ;
- par ailleurs, plus la durée d'observation augmente et plus le taux de couverture augmente. Une nuance est toutefois à apporter puisqu'on constate des différences notables : par exemple, pour un seuil $T = 75$ et une durée d'observation de $t = 14s$, le taux de couverture de l'image *KayakCouleur0* est de 21.8% alors que celui de l'image *ChurchAndCapitol* est de 40.2%, soit deux fois plus. Le taux de couverture caractérise ici la présence de zones perceptuellement importantes contenues dans l'image. En effet, pour un taux de couverture élevé, on peut considérer qu'aucune zone de l'image ne présente un intérêt perceptuel très nettement supérieur aux autres. Par contre, pour un taux de couverture faible, on peut être en présence d'une image présentant une seule zone ayant un intérêt perceptuel important. Une propriété de notre système visuel est donc une nouvelle fois constatée : l'observateur continue de porter son intérêt sur les zones perceptuellement importantes plutôt que de parcourir et de "découvrir" l'ensemble de l'image.

Finalement, le taux de couverture d'une image dépend fortement de son contenu.

1.4 Expérimentations sur séquences d'images

1.4.1 Le protocole

Le protocole des expérimentations oculométriques est donné tableau 1.5. Huit séquences de contenus variés (sport avec fort mouvement et mouvement moyen, publicité, séquence à forte activité,...) ont été utilisées. Le panel d'observateurs est composé d'étudiants de l'école Polytechnique de Nantes et de salariés de Thomson Rennes. Comme pour les données oculométriques collectées lors des tests sur images fixes, chaque enregistrement oculométrique a été visualisé afin de détecter des éventuels problèmes.

TAB. 1.5: Protocole des expérimentations pour l'acquisition de données oculométrique sur séquences d'images.

Protocole des expérimentations	
Distance d'observation	$5H$: H hauteur de l'écran de visualisation
Résolution de l'écran	800×600
Nombre de séquences	8
Nombre d'observateurs	30
Calibrage	12 points de calibrage

1.4.2 Des données à l'évolution temporelle de la saillance spatiale

Pour chaque observateur, nous disposons de deux données pour chaque image d'une séquence. La période d'échantillonnage est de 20 ms et les séquences vidéo sont progressives ; deux images consécutives sont donc séparées par une durée de 40 ms.

1.4.2.1 Création d'une séquence de fixation pour l'observateur moyen

Le prétraitement des données utilisé pour les images fixes et présenté précédemment n'est pas envisageable au regard du faible nombre de données par image. Les données collectées ont donc été utilisées sans modification, bien qu'un pré traitement spatio-temporel aurait pu être défini. L'obtention de la séquence de carte de fixations pour un observateur k se fait via la relation suivante :

$$CS^k(x, y; t) = \sum_{i=1}^{M} \Delta(x - x_i, y - y_i; t - t_i) \tag{1.6}$$

avec, M le nombre total de fixations pour la durée du test, $(x_i, y_i; t_i)$ les coordonnées de la fixation i et son instant temporel.
La séquence de fixation CS pour l'observateur moyen est donnée par la relation suivante :

$$CS(x, y; t) = \frac{1}{N} \sum_{k=1}^{N} CS^k(x, y; t) \tag{1.7}$$

avec CS^k la séquence de fixations de l'observateur k, N le nombre d'observateurs.

1.4.2.2 Création d'une séquence de densité pour l'observateur moyen

A partir des séquences de fixations obtenues pour chaque observateur, le comportement oculomoteur d'un observateur moyen est déterminé en accumulant chaque carte de fixation issue du même instant temporel. L'application d'une gaussienne bi-dimensionelle permet, comme précédemment, de construire une densité et de prendre en compte la précision limitée de l'oculomètre. La relation suivante définit cette séquence :

$$DS(x, y; t) = CS(x, y; t) * g_{\sigma_x, \sigma_y}(x, y) \tag{1.8}$$

avec g_{σ_x, σ_y} une gaussienne bi-dimensionnelle donnée par la relation (1.4). Les écarts-types σ_x et σ_y gardent la valeur de 0.5 degré utilisée sur image fixe.

1.4.3 Quelques résultats sur les séquences d'images

Deux résultats sont donnés figures 1.5 et 1.6 pour la séquence *Kayak*. La première figure montre l'évolution temporelle de la position des points de fixation. Pour des raisons évidentes, nous n'affichons que les images ayant une numérotation multiple de 5. En d'autres termes, les séquences d'images des figures 1.5 et 1.6 sont des versions temporellement sous échantillonnées d'un facteur cinq. Le nombre d'images par seconde de la séquence originale est de 25. La figure 1.6, quant à elle, montre la séquence de densité de fixation obtenue pour la séquence *Kayak* et l'observateur moyen. Il est intéressant de noter les points suivants : tout d'abord, la dépendance au centre de l'image apparaît clairement sur la première image de la séquence de densité. Par ailleurs, les densités de saillance présentent une distribution très concentrée et des bords francs. Enfin, des points atypiques apparaissent sur certaines images. Rappelons que pour une image, nous disposons au mieux d'un nombre de données égal au nombre d'observateurs multiplié par deux (temps d'échantillonnage de 20 ms ; période images de 40 ms). Étant donné que la

FIG. 1.5: Points de fixation des observateurs superposés à la séquence originale *Kayak*. L'ordre temporelle des images est de gauche à droite et de haut en bas. La séquence d'images illustrée ici est un sous échantillonnage temporel de l'originale (une image sur cinq est conservée).

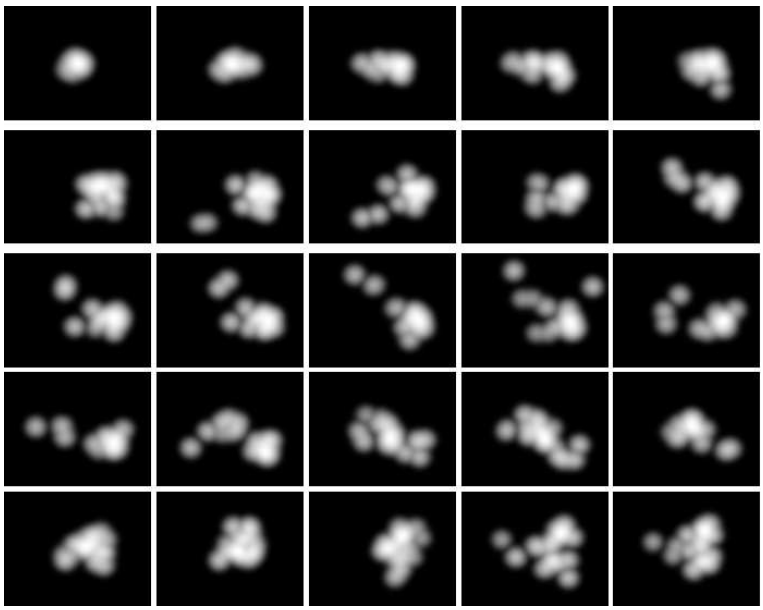

FIG. 1.6: Densités de fixation des observateurs superposés à la séquence originale *Kayak*. L'ordre temporelle des images est de gauche à droite et de haut en bas. La séquence d'images illustrée ici est un sous échantillonnage temporelle de l'originale (une image sur cinq est conservée).

dynamique des cartes est normalisée sur 255 niveau de gris, il n'est pas étonnant de voir apparaître ce type de situation.

Sur les sept autres séquences testées, les résultats sont tout à fait conformes à ceux que nous attendions. Deux phénomènes ont tout de même été constatés. Le premier concerne un effet d'anticipation. Sur la séquence de tennis *Stefan*, on constate lorsque le coureur de tennis se déplace rapidement vers le filet que les zones fixées par les observateurs se situent en avant du joueur. Les observateurs anticipent donc la visualisation de l'évènement qui a provoqué la course du joueur de tennis. Sur ce type de plan, l'aspect cognitif est important et donc difficilement modélisable. Le second point concerne l'aspect de surprise qui intervient à une rupture de plan. Sur les images suivant cette rupture, les observateurs ont tendance à fixer le centre de l'image, même lorsqu'une zone saillante apparaît.

1.5 Conclusion

La vocation de ce chapitre était de présenter les expérimentations de suivi du regard menées dans ces travaux te thèse. Des expérimentations oculométriques ont été réalisées à la fois sur images fixes et sur séquences d'images animées. Pour chaque image ou séquence d'images, les données collectées ont permis de construire une référence que nous appelons également vérité terrain. Cette dénomination est riche de sens mais nécessite tout de même quelques nuances. Tout d'abord, les tests oculométrique se sont déroulés dans un contexte dit de *free viewing*. Il est clair que nous souhaitions modéliser l'attention visuelle de type *Bottom-Up*. Ainsi, le contenu des scènes à visualiser n'était pas connu des participants aux tests et aucune instruction particulière n'était donnée aux participants, si ce n'est de regarder le moniteur d'affichage de la façon la plus naturelle possible. Malgré ces précautions, les données collectées ne sont pas exclusivement liées au mécanisme *Bottom-Up*. De nombreux aspects tels que l'aspect cognitif et l'environnement local agissent sur le comportement oculaire des participants. Les aspects précités constituent bien la première source de limitations des résultats et la première source de divergence avec une modélisation de l'attention visuelle utilisant uniquement des attributs de bas niveaux. Par ailleurs, le protocole expérimental est discutable. Nous avons utilisé un protocole qui affiche une image ou une séquence d'images centré sur un moniteur. Cette façon de procéder n'est certainement pas la meilleure. Le centre de l'image en lui-même est une zone particulière attirant l'attention visuelle. Cette attraction n'est pas exclusivement liée au contenu mais également à une donnée de haut niveaux. En effet, nous sommes très familiers aux contenus vidéo dans lesquels la région visuellement importante est régulièrement centrée. Nous aurons, dans les prochains chapitres, l'occasion de revenir sur ce point. Le fait d'afficher des images ou des séquences d'images centré sur un moniteur risque d'accentuer de façon involontaire l'importance du centre de l'image. Fort de cette expérience, des tests consistants à afficher une image ou une séquence d'images de façon aléatoire sur le moniteur aurait peut être été plus pertinents. Le protocole constitue donc la deuxième limitation majeure. Enfin, d'autres limitations peuvent être mentionnées d'ordres plus générales comme l'ori-

gine géographique des observateurs[3], l'aptitude des participants à pratiquer des jeux vidéo (les jeux vidéo améliorent significativement l'attention visuelle en terme de rapidité et de détection [Green 03])...
En dépit de ces limitations, les données oculométriques recueillies sont incontournables et primordiales pour évaluer les performances d'un modèle d'attention visuelle.

[3]Une étude de R. Nisbett [Nisbett 05] récemment parue consistait à enregistrer les mouvements oculaires de 25 étudiants américains et de 27 étudiants chinois. 36 photos, présentant une région fortement saillante au premier plan et un arrière plan complexe, ont chacune été présentées pendant 3 secondes. R. Nisbett et ses collègues ont constaté que les étudiants asiatiques portaient plus leur attention sur le fond que les étudiants américains. Ces derniers ont essentiellement concentré leur attention sur les objets saillants : *"the Americans fixated more on focal objects than did the Chinese, and the Americans tended to look at the focal object more quickly. In addition, the Chinese made more saccades to the background than did the Americans. Thus, it appears that differences in judgment and memory may have their origins in differences in what is actually attended as people view a scene".*

Chapitre 2

Modélisation de l'attention visuelle sur images fixes. Évaluation des performances

2.1 Introduction

L'objet de ce chapitre est de présenter notre contribution dans le domaine de la modélisation des propriétés fonctionnelles du système visuel. Nous nous intéressons, plus particulièrement, à la modélisation de la vision pré-attentive permettant de sélectionner les zones les plus saillantes d'une image couleur.
Le modèle que nous proposons est basé sur l'architecture de C. Koch et S. Ullman [Koch 85]. Pourtant, la conception de notre modèle se différencie fondamentalement des modèles de l'état de l'art basés sur ce type d'architecture. Le problème récurrent des modèles de l'état de l'art réside dans la façon de combiner et de mettre en compétition des données issues de différentes dimensions. Les modèles actuels font appel très tôt à des procédés de normalisation pour contrôler la dynamique des données. Ce type de procédés, consistant à prendre le maximum global et à diviser chaque site par cette valeur, est sensible à différents problèmes (le bruit impulsionnel par exemple). Pour pallier ce défaut, nous construisons un espace psycho-visuel dans lequel chaque composante est exprimée en fonction de son seuil différentiel de visibilité. A partir de cet espace, nous sommes en mesure de décliner différentes stratégies pour transformer les données exprimées en terme de visibilité en valeur de saillance. Les procédés mis en place doivent s'efforcer de déterminer parmi les zones visibles les zones qui "sautent aux yeux". Nous décrivons dans ce chapitre une façon de procéder. Elle consiste à construire trois densités de saillance associées aux trois dimensions principales exploitées par notre système visuel : la dimension achromatique et les deux dimensions chromatiques. La densité de saillance spatiale finale est ensuite obtenue en fusionnant ces trois cartes. Une méthode de fusion est déclinée sur la base d'outils de traitement d'images.
La seconde partie de chapitre concerne l'évaluation des performances du modèle proposé sur images fixes. Cette évaluation consiste à comparer de façon qualitative et quantitative les densités de saillance prédites avec la référence issue des tests oculométriques. Les évaluations qualitatives sont bien évidemment les plus simples à mener mais très criti-

quables. Elles permettent néanmoins d'apporter des indications concrètes très rapidement. Concernant l'évaluation quantitative, il n'existe actuellement pas de méthodes standards décrivant clairement le protocole à suivre. Trois méthodes complémentaires sont utilisées ici : le classique coefficient de corrélation, la divergence de Kullback-Leibler et une classification caractérisée par une matrice de confusion. Ces trois méthodes ainsi que leur complémentarité sont expliquées dans la suite du document. Une étude comparative est également menée avec le modèle de L. Itti.

2.2 Principe général de la modélisation proposée

Comme nous l'avons précédemment évoqué, la plupart des modèles psycho-visuels se basent sur l'architecture de C. Koch et de S. Ullman [Koch 85]. Cette architecture consiste à effectuer un premier traitement automatique massivement parallèle sur l'ensemble du champ visuel. Cette étape est suivie d'un traitement localisé, permettant d'extraire les caractéristiques visuellement saillantes.
Le premier traitement extrait des caractéristiques visuelles de la scène et les regroupe dans un certain nombre de cartes. Chaque carte fournit une représentation déformée et topologiquement exacte du champ visuel. Bien qu'il n'y ait pas de listes exhaustives des caractéristiques à extraire, les plus courantes sont l'orientation, la couleur, la courbure...
Le second traitement fait référence à la localisation des éléments qui "sautent aux yeux". C. Koch et S. Ullman définissent la saillance, c'est à dire le degré d'attractivité d'un site, comme une fonction dépendante de l'inhibition latérale entre cellules. En d'autres termes, dans une carte d'une caractéristique visuelle donnée, la saillance d'une zone dépend de son contexte local.
Le problème majeur de cette architecture concerne l'extraction des caractéristiques visuelles. Chaque caractéristique visuelle nécessite un procédé d'extraction propre, produisant des paramètres donnés. Il y a donc potentiellement autant de procédés d'extraction que de caractéristiques visuelles, pouvant générer des dynamiques fondamentalement différentes. La comparaison et la combinaison de cartes provenant de différentes caractéristiques visuelles s'avèrent donc très délicates.

La modélisation que nous proposons s'appuie également sur l'architecture de C. Koch et S. Ullman, mais la philosophie sous-jacente est différente. A partir d'une image incidente, un espace psycho-visuel est construit. Il est constitué des composantes naturelles de notre environnement, c'est à dire d'une composante achromatique et de deux composantes chromatiques. La différence fondamentale vis à vis de l'état de l'art se situe dans la normalisation de ces composantes. En effet, afin d'obtenir des caractéristiques visuelles homogènes et comparables, elles sont toutes normalisées par rapport à leur seuil différentiel de visibilité. Ainsi, que ce soit une valeur liée à une composante achromatique ou à une composante chromatique, elle s'exprime en fonction de leur seuil de visibilité. Comme évoqué précédemment, le premier intérêt est d'avoir des données homogènes, donc comparables. Le deuxième concerne la hiérarchisation des données. Les informations inférieures au seuil de visibilité sont négligées alors que d'autres sont mises en exergue. On ne s'intéresse donc qu'aux données perceptibles par le système visuel.
A partir de ce cadre conceptuel cohérent, la mesure de saillance de chaque site reste

à déterminer. Il s'agit de transformer des valeurs exprimées en terme de visibilité en valeurs de saillance. Pour cela, il est nécessaire de bien faire la différence entre visibilité et saillance.
La visibilité caractérise l'état d'un stimulus qui peut être visible ou invisible. Comme le passage d'un état à l'autre n'est pas instantané, la visibilité d'un stimulus est mesurée grâce à des expériences psychophysiques. Ces dernières cherchent à déterminer la probabilité de détection d'un stimulus lorsque les paramètres de ce dernier (amplitude, orientation...) varient. Concernant la saillance et bien qu'il n'y ait aucune définition formelle, le terme de saillance se réfère à l'intérêt, à l'attention ou encore la priorité qu'un site porte. La saillance qualifie également l'attractivité visuelle d'un site et par conséquent le rôle qu'il est susceptible de jouer dans la stratégie visuelle.
Dans cette étude, on propose de développer, à partir de l'espace psycho-visuel, une carte de saillance pour chaque composante extraite (une composant achromatique et deux composantes chromatiques). Ce choix est le plus intuitif. Les techniques mises en place ont pour but de déterminer les éléments de chaque composante qui "sautent aux yeux". Ce n'est qu'une utilisation particulière de cet espace. D'autres façons de combiner et de créer de la saillance sont possibles.

2.3 Conception de l'espace psycho-visuel

Le synoptique global de la modélisation spatiale de l'attention visuelle est présenté à la figure 2.1. Dans cette partie, nous nous intéressons à la construction de l'espace psycho-visuel. Les composantes de cet espace sont obtenues à partir d'un certain nombre de transformations appliquées sur une image numérique RVB. Ces transformations sont détaillées dans les paragraphes suivants.

2.3.1 Transformation non linéaire liée à l'écran

Pour simuler efficacement le système visuel, le modèle doit être alimenté par des données perceptuelles c'est à dire proches de la réalité. Une image numérique RVB, point d'entrée du modèle, a subi des distortions lors de l'acquisition mais également lors de la restitution. Le premier est difficilement modélisable car aucune connaissance a priori ne permet d'identifier clairement le procédé utilisé. Par contre, l'affichage peut être totalement maîtrisé et une correction peut être appliquée. Ainsi, les données de l'image sont converties en luminance couleur par les trois fonctions gamma de l'écran correspondant à chacune des trois composantes RVB. La fonction gamma de chaque composante est modélisée à l'aide des relations suivantes :

$$L_R = Offset_R + L_{R,max} \times (\frac{R}{R_{max}})^{\gamma_R} \quad (2.1)$$

$$L_V = Offset_V + L_{V,max} \times (\frac{V}{V_{max}})^{\gamma_V} \quad (2.2)$$

$$L_B = Offset_B + L_{B,max} \times (\frac{B}{B_{max}})^{\gamma_B} \quad (2.3)$$

avec,

- L_R, L_V et L_B intensité de la composante Rouge, Vert ou Bleu (respectivement) ;

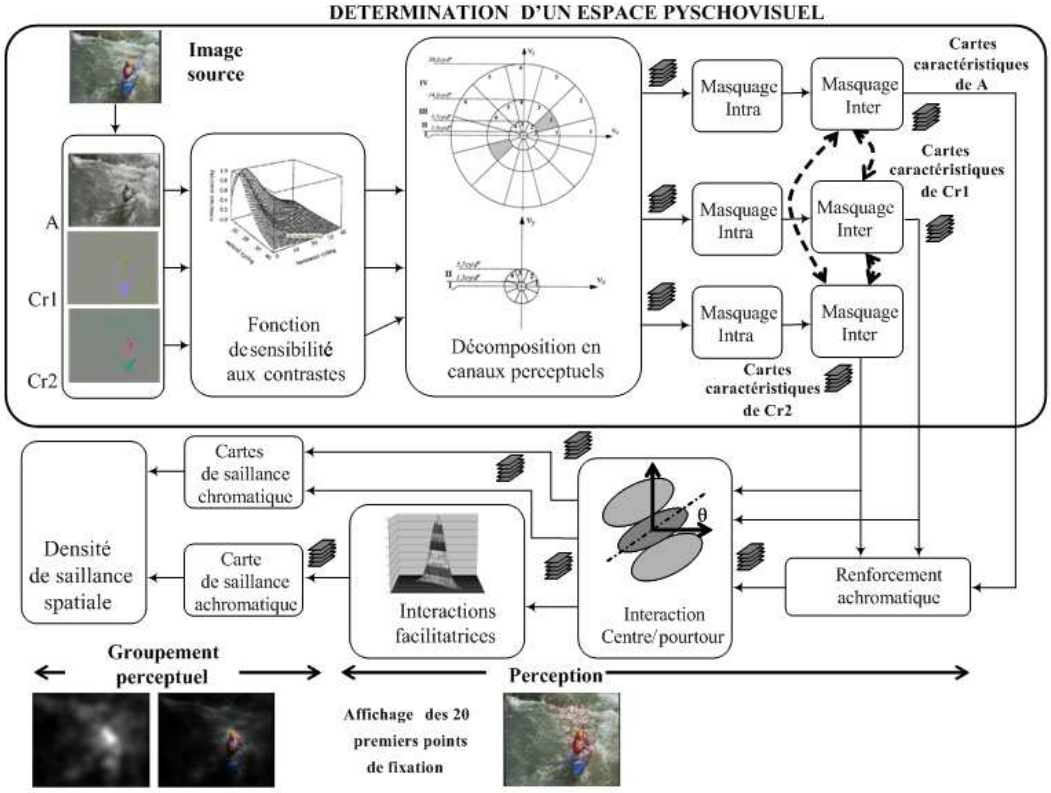

FIG. 2.1: Synoptique global de la modélisation spatiale de l'attention visuelle. A partir d'une image RVB, un espace psycho-visuel est déterminé. Une stratégie particulière est alors choisie pour extraire de cet espace la saillance spatiale. Un exemple de carte et de densité de saillance est donné. Sur la carte (en bas à gauche du synoptique), les zones claires correspondent aux zones saillantes de l'image.

- $Offset_R$, $Offset_V$ et $Offset_B$, valeur de la luminance pour un niveau nul de la composante Rouge, Vert ou Bleu (respectivement) ;
- $L_{R,max}(R)$, $L_{V,max}(V)$ et $L_{B,max}(B)$ valeur de la luminance pour un niveau maximum de la composante Rouge, Vert ou Bleu (respectivement) ;
- $R_{max} = V_{max} = B_{max} = 255$ pour un codage sur 8 bits ;
- γ_R, γ_V et γ_B, paramètres dépendants du dispositif d'affichage mesurés à l'aide d'un luxmètre (appelés communément les gammas de l'écran).

2.3.2 Projection dans un espace perceptuel de représentation couleur

Comme nous l'avons vu dans la première partie, le système visuel décompose l'information lumineuse incidente en trois composantes distinctes. De nombreux espaces sont décrits dans la littérature ; il n'est donc pas simple de choisir un espace de représentation et de justifier un tel choix. Dans notre étude, ce problème ne s'est pas posé puisque nous avons bénéficié des précédents travaux de thèse de L. Bedat [Bedat 98]. Ce dernier a validé l'espace de représentation défini par J. Krauskopf [Krauskopf 82] pour des images visualisées dans des conditions recommandées par l'I.T.U. Cet espace est caractérisé par

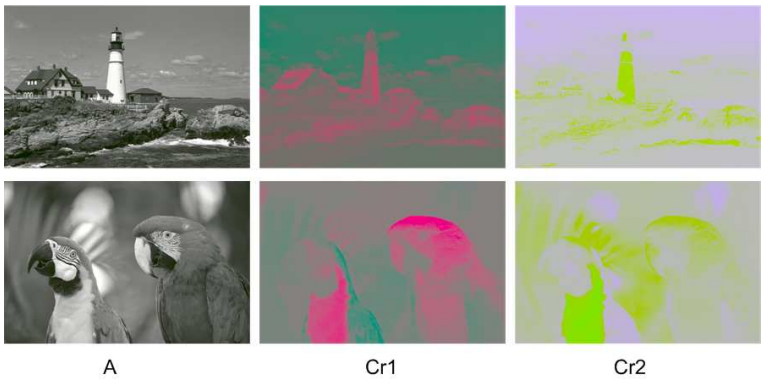

FIG. 2.2: Les trois composantes (A,Cr_1,Cr_2) de l'espace colorimétrique de J. Krauskopf pour les images *Lighthouse* et *Parrots*.

la transformation suivante :

$$\begin{pmatrix} A \\ Cr_1 \\ Cr_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ -0.5 & -0.5 & 1 \end{pmatrix} \begin{pmatrix} L \\ M \\ S \end{pmatrix} \tag{2.4}$$

où, A, Cr_1 et Cr_2 représentent les trois composantes perceptuelles et L, M et S les signaux issus des trois types de cônes. L'espace LMS est obtenu en transformant les signaux lumineux via l'espace CIE[1] x'y'.
La composante A représente la composante achromatique. La composante Cr_1 prend ses valeurs sur un axe rouge-vert. La composante Cr_2, quant à elle, prend ses valeurs sur un axe jaune-bleu. La relation de passage de (L_R, L_V, L_B) à (A, Cr_1, Cr_2) est donnée ci-dessous :

$$\begin{pmatrix} A \\ Cr_1 \\ Cr_2 \end{pmatrix} = L_{max} \begin{pmatrix} \frac{0.2244}{L_{R,max}} & \frac{0.6811}{L_{V,max}} & \frac{0.0942}{L_{B,max}} \\ \frac{0.0891}{L_{R,max}} & \frac{-0.0617}{L_{V,max}} & \frac{-0.0275}{L_{B,max}} \\ \frac{-0.1029}{L_{R,max}} & \frac{-0.2874}{L_{V,max}} & \frac{0.3903}{L_{B,max}} \end{pmatrix} \begin{pmatrix} L_R \\ L_V \\ L_B \end{pmatrix} \tag{2.5}$$

avec :
- L_R, L_V et L_B correspondent respectivement aux luminances des composantes Rouge, Vert et BLeu ;
- $L_{R,max}$, $L_{V,max}$ et $L_{B,max}$ correspondent respectivement aux luminances maximales des composantes Rouge, Vert et BLeu ;
- $L_{max} = L_R + L_V + L_B$ représente la luminance maximale.

La figure 2.2 présente les trois composantes (A, Cr_1, Cr_2) pour les images *Lighthouse* et *Parrots* (pour des besoins de visualisation les composantes Cr_1 et Cr_2 étant centrées sur 0, l'ajout d'un offset est effectué).

2.3.3 Application de fonctions de sensibilité aux contrastes

La sensibilité du système visuel à une stimulation visuelle dépend de nombreux paramètres, comme nous l'avons montré au paragraphe 2.2.3. Des CSFs (*Contrast Sensitivity*

[1]Comité International de l'Électricité

Functions) anisotropiques permettent de considérer deux de ces paramètres : l'orientation et la fréquence spatiale. Il est clair que plus la fréquence spatiale est élevée et plus la sensibilité du système visuel est faible. Concernant l'orientation, la sensibilité aux composantes horizontales et verticales est plus forte que la sensibilité aux composantes diagonales ; le système visuel humain, confronté à un environnement visuel majoritairement composé de composantes horizontales et verticales, a développé une hyper sensibilité à ces orientations. Chaque composante fréquentielle appartenant aux canaux (A, Cr_1, Cr_2) est donc pondérée par une CSF. Les composantes (A, Cr_1, Cr_2) sont dites normalisées vis à vis du seuil différentiel de visibilité ; les valeurs des différentes composantes sont alors proportionnelles à la valeur du seuil de visibilité.
Trois CSFs sont utilisées, une par composante. Elles s'expriment essentiellement en fonction de la pulsation radiale w en cy/deg et de l'orientation θ en degrés :

- sur la composante A, la CSF anisotropique de S. Daly est utilisée. Elle est donnée par la relation suivante, complètement explicitée dans le paragraphe 2.2.3 de la partie I :

$$S_A(w, \theta, l, s, d, e) = P \times min\left(S(\frac{w}{bw_a \times bw_e \times bw_\theta}, l, s), S(w, l, s)\right) \tag{2.6}$$

- sur la composante Cr_1, la CSF anisotropique proposée par P. Le Callet [Callet 01] est appliquée. Elle est donnée par la relation suivante, également détaillée dans le paragraphe 2.2.3 de la partie I :

$$S_{Cr_1}(w, \theta) = \frac{33}{1 + (\frac{w}{5.52}^{1.72})}(1 + 0.27sin(2\theta)) \tag{2.7}$$

- sur la composante Cr_2, la CSF anisotropique proposée par P. Le Callet [Callet 01] est appliquée. Elle est donnée par la relation suivante, également détaillée dans le paragraphe 2.2.3 de la partie I :

$$S_{Cr_2}(w, \theta) = \frac{5}{1 + (\frac{w}{4.12}^{1.64})}(1 - 0.24sin(2\theta)) \tag{2.8}$$

2.3.4 Décomposition en canaux perceptuels

Comme nous l'avons évoqué dans la première partie, les cellules du système visuel réagissent à une stimulation particulière. Il faut donc considérer le système visuel comme un procédé décomposant l'information visuelle en un ensemble de canaux, décrit en terme de sélectivité radiale et en terme de sélectivité angulaire.
La première transformation simulant les différentes populations de cellules fut proposée par A. Watson, avec la transformée Cortex. Des études approfondies, faisant appel à de nombreuses expérimentations psychophysiques menées lors de différents travaux [Sénane 96], [Bedat 98] et [Callet 01], ont permis de redéfinir les paramètres de cette transformée. Les sélectivités radiales, angulaires ainsi que les largeurs de bandes radiales et angulaires ont été déterminées dans des conditions d'observation normalisées. Les paramètres de cette nouvelle décomposition en sous bandes visuelles sont donnés par la figure 2.3 (a) pour la composante achromatique A et (b) pour les composantes Cr_1 et Cr_2. La composante A est décomposée en 17 sous bandes réparties sur 4 couronnes (un canal basses fréquences

non directionnel (noté I) et trois bandes de fréquences radiales directionnelles (notées de II à IV). Les couronnes II à IV sont décomposées en canaux angulaires dont le nombre varie avec la bande de fréquences radiales considérée.
La construction des différents canaux, que nous appelons également sous bandes visuelles, est obtenue à partir des filtres Cortex. Comme nous l'avons évoqué succinctement, les filtres Cortex sont construits à partir de deux types de filtres : les filtres DoM et les filtres Fan.

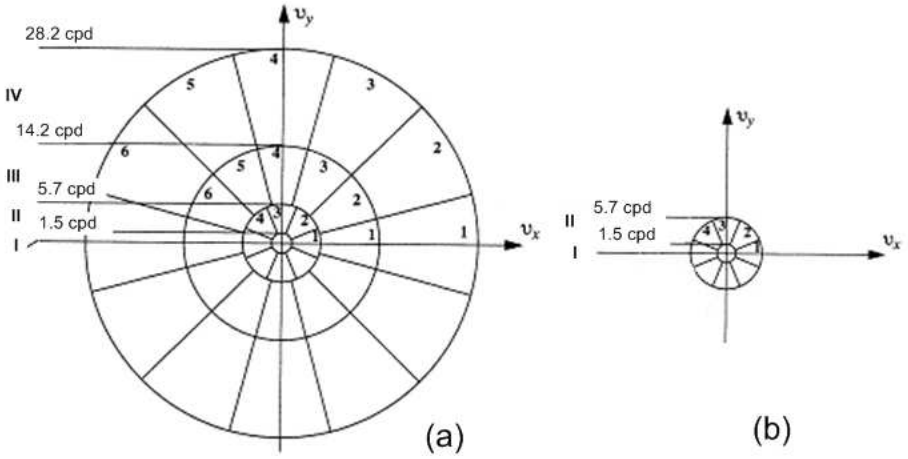

FIG. 2.3: Décomposition en canaux perceptuels : (a) décomposition de la composante achromatique en 17 sous bandes réparties sur les couronnes I à IV ; (b) décomposition des composantes chromatiques en 5 sous bandes réparties sur les couronnes I à II.

2.3.4.1 Filtres DoM à sélectivité radiale

Les filtres DoM (*Difference of Mesa*) sont des filtres passe bande en fréquences radiales. Ils sont utilisés pour construire les couronnes de la décomposition en canaux perceptuels. Comme son nom l'indique, les filtres DoM sont déterminés à partir de filtres $Mesa$. Ces derniers sont des filtres passe bas en fréquences radiales. La fonction de transfert d'un filtre $Mesa$ est obtenue grâce à la convolution d'un échelon de *Heaviside* et d'une gaussienne.

$$Mesa_{f_c}(f) = Echelon_{f_c}(f) * \left\{ \frac{1}{\sigma\sqrt{2\pi}} \times exp(-\frac{f^2}{2\sigma^2}) \right\} \tag{2.9}$$

avec :
- f fréquence radiale,
- $Echelon_{f_c}$, l'échelon de *Heaviside* ayant une valeur unité à l'intérieur d'un cercle de rayon f_c (fréquence de coupure) et une valeur nulle à l'extérieur.

Un filtre DoM est ensuite déterminé en faisant la différence de deux filtres $Mesa$ ayant des fréquences de coupures différentes :

$$DoM_{f_{c_1},f_{c_2}}(f) = Mesa_{f_{c_2}}(f) - Mesa_{f_{c_1}}(f) \tag{2.10}$$

avec, $f_{c2} > f_{c1}$.

2.3.4.2 Filtres Fan à sélectivité angulaire

La détermination d'un filtre Fan nécessite la construction d'un filtre $Step$. Ce dernier est obtenu par la convolution d'un échelon orienté avec une gaussienne. Un filtre $Step$ d'orientation θ est défini par la relation suivante :

$$Step_\theta(f) = Echelon_\theta(f) * \left\{ \frac{1}{\sigma\sqrt{2\pi}} \times exp(-\frac{f^2}{2\sigma^2}) \right\} \tag{2.11}$$

avec :
- f fréquence radiale,
- $Echelon_\theta$, un échelon orienté selon la direction θ.

Le filtre Fan s'obtient ensuite par deux filtres $Step$ d'orientations différentes :

$$Fan_{\theta_1,\theta_2}(f) = |Step_{\theta_2}(f) - Step_{\theta_1}(f)| \tag{2.12}$$

avec, $\theta_2 > \theta_1$.

2.3.4.3 Synthèse des filtres

Un filtre Cortex est ensuite déduit d'un filtre DoM et d'un filtre Fan par la relation suivante :

$$Cortex_{\rho,\theta}(f) = DoM_\rho(f) \times Fan_\theta(f) \tag{2.13}$$

avec :
- ρ correspond à la bande de fréquences isolée par le filtre passe bande DoM,
- θ correspond à la gamme d'orientations isolée par le filtre Fan.

La figure 2.4 présente les filtres nécessaires pour extraire les sous bandes de la couronne II d'une image de résolution 512×512. Chaque image de la figure a une résolution de 104×104, correspondant au rapport de décimation de la fréquence maximum (28.2 cpd) et de la fréquence maximum de la couronne II (5.7 cpd).

2.3.5 Masquage visuel

La décomposition en sous bandes perceptuelles permet de simuler le pavage fréquentiel du système visuel. Chaque sous bande peut être considérée comme une population donnée de cellules visuelles répondant à une stimulation bien particulière. Bien que ces cellules soient fortement dépendantes de la fréquence spatiale et de l'orientation de la stimulation, la réponse de la cellule dépend également de la présence d'autres stimuli, c'est à dire du contexte dans lequel elle se trouve. Ce phénomène est appelé effet de masquage. Comme indiqué préalablement dans le paragraphe 2.2.5 de la partie I, il existe deux types de masquage : un masquage intra composante et inter composantes.

2.3.5.1 Masquage intra composante

Ce type de masquage regroupe à la fois les influences entre signaux traités par le même canal et les signaux traités par des canaux différents portés par la même composante.

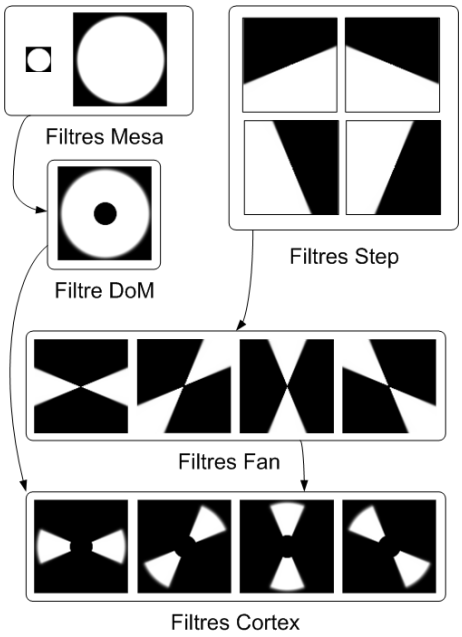

FIG. 2.4: Filtres *Mesa*, *Step*, *DoM*, *Fan* et *Cortex* nécessaires pour l'obtention des sous bandes de la couronne II.

Le modèle de masquage utilisé pour la composante A est celui de S. Daly présenté au paragraphe 2.2.5.3 de la partie I. Ce modèle n'est pas le plus performant de la littérature ; il ne prend en compte ni les effets de masquage inter sous bandes ni les effets de facilitation. Cependant, son intérêt réside dans l'optimisation des paramètres réalisée à partir d'un nombre important de données expérimentales.

Le modèle de masquage utilisé pour les composantes chromatiques est le modèle défini par P. Le Callet dans sa thèse [Callet 01]. A partir de données expérimentales collectées par L. Bedat [Bedat 98], un modèle de masquage modélisant uniquement les effets intra canal mais intégrant une zone de facilitation est calculé. La forme analytique de ce modèle, définissant l'élévation du seuil de visibilité de chaque site s d'une sous bande $R^{\alpha}_{\rho,\theta}(s)$ (α représente la composante Cr_1 ou Cr_2), est donnée par la relation suivante :

$$T^{\alpha}_{\rho,\theta}(s) = \frac{1 + a \times |R^{\alpha}_{\rho,\theta}(s)| + b \times |R^{\alpha}_{\rho,\theta}(s)|^2}{1 + c \times |R^{\alpha}_{\rho,\theta}(s)|} \tag{2.14}$$

Les valeurs des paramètres sont données dans le tableau B.1 de l'annexe B.

2.3.5.2 Masquage inter composantes

Ce type de masquage caractérise les interactions entre des signaux portés par des composantes différentes. Tout d'abord, il est important de mentionner que toutes les interactions (390 au total, issues des 17 canaux achromatiques et des 10 canaux chromatiques) ne sont pas considérées. D'après les travaux de L. Bedat, la plupart de ces interactions peuvent être négligées. Le tableau B.2 de l'annexe B présente les 14 interactions prises en

compte.
A partir des données expérimentales, il a été nécessaire de définir deux modèles. La fonction d'élévation du seuil de visibilité est notée dans les deux cas $T_{\rho,\theta}^{\alpha'\rightarrow\alpha}$. Le premier, appelé modèle A et exprimé par la relation (2.15), estime l'évolution du masquage intra intégrant une zone de facilitation. Le second modèle, modèle B, donné par la relation (2.16) exprime simplement un effet de masquage ayant une pente nulle à l'infini. Ces relations expriment l'élévation du seuil de visibilité du site s de la sous bande (ρ, θ) de la composante α en fonction de la sortie de la décomposition en canaux perceptuels au site s de la sous bande (ρ', θ') de la composante α' :

$$T_{\rho,\theta}^{\alpha'\rightarrow\alpha}(s) = \frac{1 + a \times |R_{\rho',\theta'}^{\alpha'}(s)| + b \times |R_{\rho',\theta'}^{\alpha'}(s)|^2}{1 + c \times |R_{\rho',\theta'}^{\alpha'}(s)|} \tag{2.15}$$

$$T_{\rho,\theta}^{\alpha'\rightarrow\alpha}(s) = a - b \times exp(-c \times |R_{\rho',\theta'}^{\alpha'}(s)|) \tag{2.16}$$

Les paramètres de ces deux relations sont donnés dans les tableaux B.3 et B.4 de l'annexe B.
La variation du seuil de visibilité final $T_{\rho,\theta,\alpha}(s)$ de chaque site s de chaque sous bande (ρ, θ) de chaque composante α est obtenue par un modèle multiplicatif, exprimé par la relation suivante :

$$T_{\rho,\theta,\alpha}(s) = T_{\rho,\theta}^{\alpha}(s) \times \prod_{\rho'} \prod_{\theta'} \prod_{\alpha'} T_{\rho,\theta}^{\alpha'\rightarrow\alpha}(s) \tag{2.17}$$

Chaque sous bande de la décomposition perceptuelle, normalisée par rapport au seuil de visibilité sans signal masquant (c'est à dire le seuil donné par une CSF), est alors divisée par la valeur $T_{\rho,\theta,\alpha}$ déterminée.

2.3.6 Illustration des différents mécanismes

La figure 2.5 présente le résultat de l'application d'une fonction de sensibilité aux contrastes, du masquage visuel intra composante de S. Daly et d'une décomposition en canaux perceptuels. Ces illustrations permettent d'appréhender le rôle des différents procédés utilisés.
Les images de la rangée (b) de la figure 2.5 correspondent à la différence entre l'image source et l'image pondérée par la fonction de sensibilité aux contrastes. Lorsque la différence est faible, représentée par les zones sombres des deux images de la rangée (b), la sensibilité de l'oeil est élevée. On retrouve les zones à faible activité spatiale tel que le ciel de l'image *Lighthouse*. A contrario, lorsque la différence est élevée, représenté par les zones claires, la sensibilité de l'oeil est faible. On retrouve les zones texturées des deux images.
Les images de la rangée (c) de la figure 2.5 correspondent à l'élévation du seuil de visibilité, précédemment noté T, déterminée par le masquage visuel intra de S. Daly. Lorsque la valeur de T tend vers l'unité (la facilitation n'est pas prise en compte dans le masquage de Daly), le potentiel de masquage est faible. Sur l'image *Lighthouse*, le masquage est quasi nul sur le ciel principalement constitué de zones uniformes. Par contre, lorsque la valeur de T augmente, représentée par les zones claires des images (c), il y a un phénomène de masquage. On constate bien évidemment que l'élévation du seuil de visibilité dépend fortement de la valeur de la composante traitée.

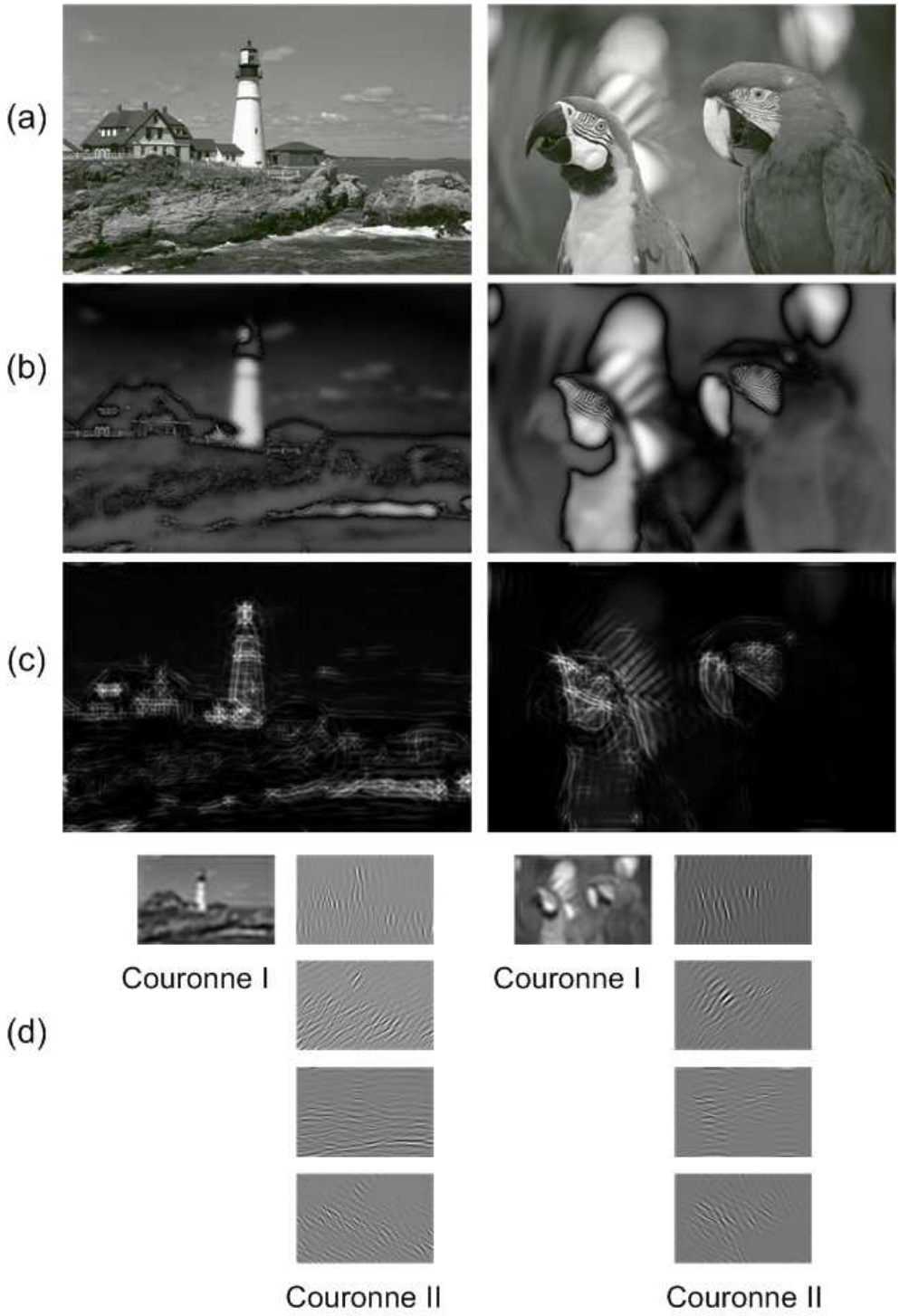

FIG. 2.5: Exemples d'applications sur la composante A des images *Lighthouse* et *Parrots* (a), d'une CSF (b), du masquage intra de S. Daly (c) et de la décomposition en canaux perceptuels (d). Pour la décomposition, seules les sous bandes de la couronne I et II sont données.

Les images de la rangée (d) présentent les sous bandes des couronnes I et II issues de la décomposition en canaux perceptuels de la composante achromatique de *Lighthouse* et *Parrots*. L'effet de masquage, précédemment illustré, dépend de cette décomposition. Pour des raisons de place, seules ces sous bandes sont illustrées. Il va de soit que la décomposition génère 17 sous bandes au total. Notons, également que les sous bandes de la figure 2.5 n'ont pas été sous échantillonnées. Enfin, les niveaux de gris $ng_{\rho,\theta}(s)$ sont :

- pour la couronne I : $ng_{\rho,\theta}(s) = 2 \times R^A_{\rho,\theta}(s) + 128$
- pour la couronne II : $ng_{\rho,\theta}(s) = 4 \times R^A_{\rho,\theta}(s) + 128$

avec, $R^A_{\rho,\theta}(s)$ la valeur au site s de la sous bande (ρ, θ) issue de la décomposition de la composante A.

2.4 Construction d'une saillance spatiale à partir de l'espace psycho-visuel

2.4.1 Génération de saillance achromatique

2.4.1.1 Objectif et remarques

L'objectif est de transformer les valeurs de visibilité des sous bandes achromatiques issues de l'espace psycho-visuel en valeurs de saillance. Comme définie précédemment, la saillance est une mesure d'attractivité. La détermination des zones qui "sautent aux yeux" s'articule autour de deux aspects. Le premier consiste à renforcer les structures achromatiques en fonction de leur environnement local couleur. Le second point consiste à supprimer la redondance des composantes visuelles.
Les étapes, évoquées précédemment et décrites dans les paragraphes suivants, s'appliquent uniquement sur la couronne II de la composante achromatique. La couronne II concerne la bande de fréquences radiales s'étendant de 1.5 à 5.7 *cpd*. Plusieurs explications sont à l'origine de ce choix. La première concerne l'importance de la bande de fréquences intermédiaires dans la perception visuelle au sens large. Rappelons que la sensibilité de l'oeil est maximale dans cette gamme de fréquences. Par ailleurs, les sous bandes de la couronne II présentent un compromis acceptable entre quantité d'informations transmise et précision de leur localisation spatiale. En basses fréquences, la quantité d'informations est grande mais la précision de leur localisation spatiale est faible. En hautes fréquences, la précision de la localisation spatiale est excellente mais la quantité d'informations faible. A ces explications d'ordre général, soulignons le fait que les sélectivités angulaires des sous bandes de la couronne II des composantes achromatique et chromatiques sont identiques. Des interactions pourront donc être envisagées plus facilement. Enfin, des tests ont montré que l'introduction de la couronne III et IV pour la détermination de la carte de saillance achromatique n'améliore pas les résultats.
Suite à ce choix, les sous bandes d'une composante α sont notées par défaut R^α.

2.4.1.2 Renforcement des structures achromatiques

Les structures achromatiques d'une image sont à la base de la perception visuelle. Elles n'ont cependant pas toutes la même capacité à attirer l'attention. Certaines configurations présentent un intérêt particulier : c'est le cas des structures achromatiques séparant des

zones couleurs fortement contrastées. La première étape de la génération de la saillance

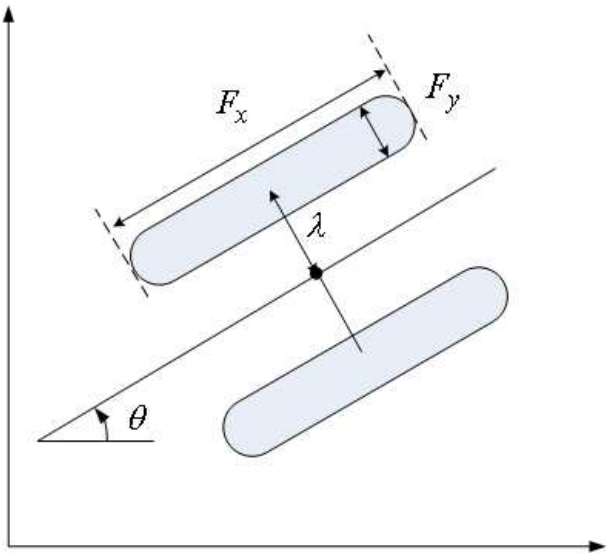

FIG. 2.6: Synoptique du procédé utilisé pour calculer le coefficient de renforcement.

achromatique consiste donc à renforcer les structures achromatiques en fonction de leur environnement couleur. Par conséquent, à chaque site s d'une sous bande achromatique appartenant à la couronne II, un coefficient de renforcement, noté η_{renf} est déterminé. Le coefficient η_{renf} est égal à la somme des coefficients η_{Cr_1} et η_{Cr_2}, traduisant respectivement le renforcement engendré par la composante Cr_1 et Cr_2. Ainsi, pour chaque site s d'une sous bande achromatique R^A, un gradient de couleur est calculé en effectuant la différence point à point entre deux ensembles de sites appartenant aux basses fréquences des composantes couleurs. Les ensembles de points sont orientés selon la direction privilégiée de la sous bande achromatique considérée. La figure 2.6 illustre le procédé. Les centres des deux sous-ensembles de tailles (F_x, F_y) sont situés de chaque côté du point courant à une distance λ, dans la direction perpendiculaire à l'orientation θ de la sous bande achromatique considérée. La distance λ est fonction de la fréquence spatiale centrale de la sous bande considérée. Elle est égale à une demi-période spatiale de la fréquence centrale $f_{min} + \frac{f_{max}-f_{min}}{2}$ de la sous bande considérée.
En pratique et pour la couronne II, λ est égal à 5 pixels. Les valeurs (F_x, F_y) exprimées en pixels sont choisies arbitrairement à $(5, 1)$.
La figure 2.7 présente les cartes de coefficients de renforcement obtenues pour les deux images *Lighthouse* et *Parrots* et pour les deux composantes Cr_1 et Cr_2. Les contrastes de couleur sont correctement détectés.
Finalement, les sous bandes achromatiques de la couronne II sont pondérées par le coefficient de renforcement η_{renf} :

$$R^A_{(1)}(s) = R^A(s) \times (1 + \eta_{renf}(s)) \tag{2.18}$$

avec, $\eta_{renf}(s) = \eta_{Cr_1}(s) + \eta_{Cr_2}(s)$. L'indice $_{(1)}$ du terme $R^A_{(1)}$ signifie que c'est la première étape de modification des sous bandes achromatiques.

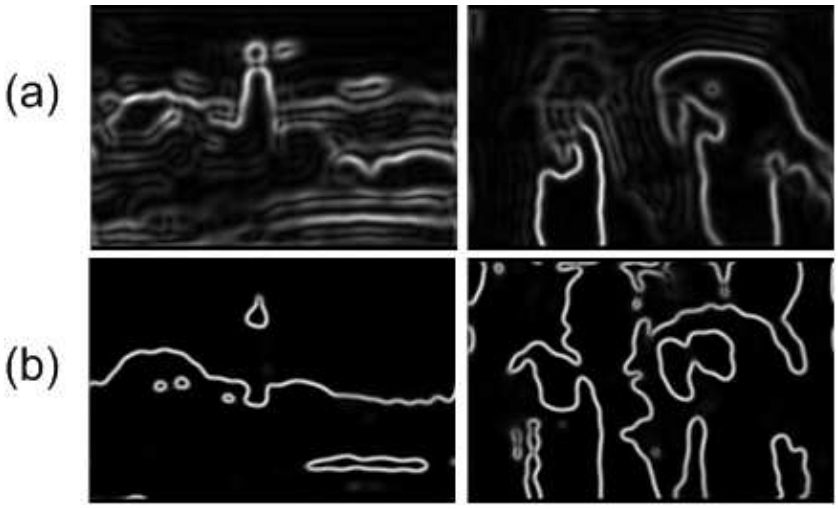

FIG. 2.7: Carte des coefficients de renforcement pour les images *Lighthouse* et *Parrots*. Les coefficients de renforcement issus de la composante Cr_1 et Cr_2 sont respectivement donnés en (a) et (b).

2.4.1.3 Suppression des données achromatiques redondantes

Pour pouvoir traiter rapidement les informations de notre champ visuel, le SVH supprime la redondance spatialement localisée dans un environnement de taille donnée. Différents types de cellules, et plus particulièrement les cellules corticales, participent à ce mécanisme. Ces cellules répondent fortement à des singularités locales telles que les contours. Dans la stratégie que nous avons mis en place, la suppression des données achromatiques redondantes s'effectue en reproduisant le comportement ces cellules corticales.
La modélisation de ce comportement se fait de façon relativement classique via une différence de gaussiennes (abrégé *DoG* en anglais) anisotropes, orientées suivant la direction préférée de la sous bande considérée. Une *DoG* est constituée d'une gaussienne dite excitatrice et d'une gaussienne dite inhibitrice, présentant une étendue spatiale bien plus importante (2 à 5 fois plus grande). Ce type de modélisation permet d'appréhender convenablement et simplement les principales propriétés des cellules corticales qui sont juste rappelées ici :

- les cellules corticales présentent une sélectivité à la fréquence spatiale et à l'orientation de la stimulation ;
- les cellules corticales répondent fortement sur des zones contrastées et ne répondent pas ou peu sur des zones uniformes.

La contribution inhibitrice de la réponse de la cellule corticale est obtenue en convoluant le signal de chaque sous bande R^A avec la partie inhibitrice de la *DoG* modifiée dont le profil est présenté à la figure 2.8. Ce résultat est ensuite soustrait à la valeur courante de la sous bande R^A. Seuls les résultats positifs sont considérés. Ainsi, les valeurs des sites s du signal des sous bandes R^A sont obtenues via la relation (2.19) :

$$R^A_{(2)}(s) = H(R^A_{(1)}(s) - R^A_{(1)}(s) * w_{\sigma_x^{ex},\sigma_y^{ex},\sigma_x^{inh},\sigma_y^{inh}}(s)) \tag{2.19}$$

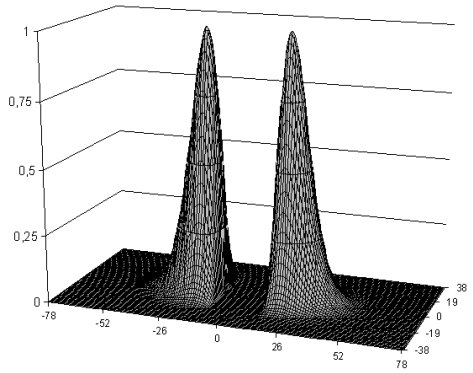

FIG. 2.8: Profil de la fonction $w_{\sigma_x^{ex},\sigma_y^{ex},\sigma_x^{inh},\sigma_y^{inh}}$ modélisant la contribution inhibitrice d'une cellule corticale.

avec,

$$w_{\sigma_x^{ex},\sigma_y^{ex},\sigma_x^{inh},\sigma_y^{inh}}(s) = \frac{1}{\left\| H(DoG_{\sigma_x^{ex},\sigma_y^{ex},\sigma_x^{inh},\sigma_y^{inh}}) \right\|} H(DoG_{\sigma_x^{ex},\sigma_y^{ex},\sigma_x^{inh},\sigma_y^{inh}}(s')) \tag{2.20}$$

avec, $s = (x, y)$ et $s' = (x', y')$ le site ayant subi une rotation d'angle θ donné par la relation (2.21).

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} cos\theta_{\rho,\theta} & sin\theta_{\rho,\theta} \\ -sin\theta_{\rho,\theta} & cos\theta_{\rho,\theta} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{2.21}$$

La fonction $H(\cdot)$ permet de conserver uniquement les valeurs positives. Ici, cette fonction est utile lorsque l'inhibition dépasse la valeur d'excitation.

$$H(z) = \left\{ \begin{array}{ll} 0 & \text{si } z < 0 \\ z & sinon \end{array} \right\} \tag{2.22}$$

La fonction $DoG_{\sigma_x^{ex},\sigma_y^{ex},\sigma_x^{inh},\sigma_y^{inh}}$ est caractérisée par les écarts types suivants :

- le couple $(\sigma_x^{ex}, \sigma_y^{ex})$ caractérise la gaussienne dite excitatrice, avec $\sigma_x^{ex} > \sigma_y^{ex}$ (les indices $_x$ et $_y$ représentent respectivement l'axe principal et l'axe secondaire de la gaussienne) ;
- le couple $(\sigma_x^{inh}, \sigma_y^{inh})$ caractérise la gaussienne dite inhibitrice, avec $\sigma_x^{inh} > \sigma_y^{inh}$.

La détermination des écarts types de ces gaussiennes est réalisée en fonction de la fréquence spatiale maximale des sous bandes de la couronne II. Le profil de la fonction de pondération $w_{\sigma_x^{ex},\sigma_y^{ex},\sigma_x^{inh},\sigma_y^{inh}}$ est donné figure 2.8 pour une orientation θ nulle. La figure 2.9 illustre le résultat de l'application de la relation (2.19) sur les images *Lighthouse* et *Parrots*.

2.4.1.4 Modélisation des interactions facilitatrices de type iso-orienté et co-linéaire

Des études psychophysiques, notamment celles de U. Polat et D. Sagi [Polat 93, Polat 94], ont montré que la réponse des cellules visuelles à un signal cible pouvait être

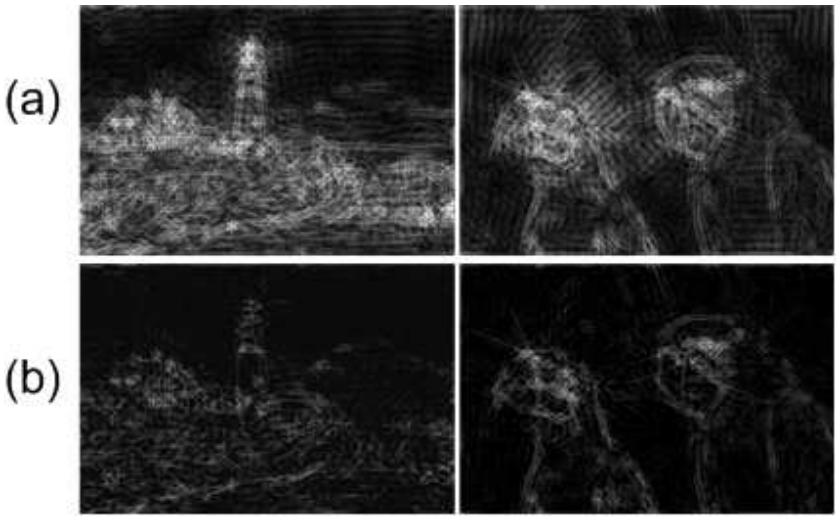

FIG. 2.9: Résultats d'application de l'opérateur de modélisation des cellules corticales sur les images *Lighthouse* et *Parrots*. Sur la rangée (a), les images avant application de l'opérateur sont illustrées. Sur la rangée (b), le résultat du filtrage est donné. Notons, que les images après l'application de l'opérateur de modélisation des cellules corticales ont été normalisées avec le maximum de la dynamique de niveau gris avant modification. Les images sont donc tout à fait comparables.

modifiée par la présence de deux stimuli périphériques, de même orientation et de même fréquence spatiale. Ils ont observé que la variation de cette réponse était dépendante de nombreux paramètres. Une liste non exhaustive des paramètres influençant le seuil est donnée ci-dessous :

- l'orientation et l'alignement relatif des stimuli périphériques : l'effet est maximal lorsque les 3 stimuli présentent les mêmes caractéristiques et sont alignés ;
- la distance séparant les différents stimuli : un phénomène de masquage est observé pour des courtes distances ($< 2\lambda$, où λ est la période du stimulus). Au delà de cette distance, des effets de facilitation sont observés, décroissant progressivement avec la distance ;
- le degré de colinéarité : lorsque les stimuli périphériques sont déplacés dans la direction orthogonale à l'axe de l'orientation préférée du stimulus cible, l'effet facilitateur décroît très rapidement.

La modélisation de ces mécanismes est un sujet d'études à part entière. Notre objectif, bien plus modeste sur ce point, est de proposer un modèle simple intégrant les propriétés majeures listées ci-dessous. Pour cela, nos travaux s'appuient sur les travaux de K. Mizobe [Mizobe 01] et de T. Hansen [Hansen 02], basés sur le principe gestalien de bonne continuité.

Les interactions facilitatrices sont simulées par deux filtres, notés $B^0_{\rho,\theta}$ et $B^1_{\rho,\theta}$, déduits d'un filtre papillon $B_{\rho,\theta}$. Ce dernier est obtenu par un terme directionnel $D_{\rho,\theta}(s)$ et un terme radial C_r lissé par une gaussienne $G(s)$. Le rayon r du cercle C_r permet de définir l'étendue spatiale à examiner. Le filtre en forme de papillon a une réponse donnée par la

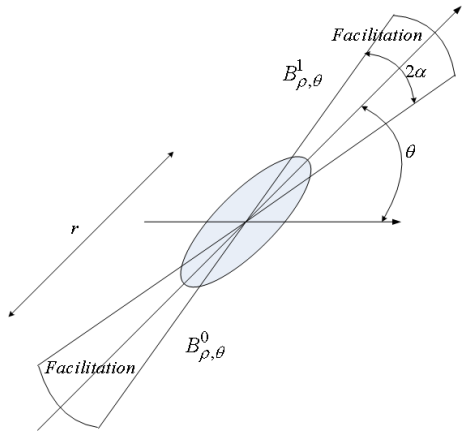

FIG. 2.10: Profil du filtre papillon $B_{\rho,\theta}$ et des filtres $B^0_{\rho,\theta}$ et $B^1_{\rho,\theta}$ pour une orientation préférée θ.

relation :

$$B_{\rho,\theta}(s) = D_{\rho,\theta}(s) \cdot C_r(s) * G(s) \tag{2.23}$$

avec,

$$D_{\rho,\theta}(s) = \begin{cases} cos(\frac{\pi/2}{\alpha}\varphi) & \text{if } \varphi < \alpha \\ 0 & \text{sinon} \end{cases} \tag{2.24}$$

and $\varphi = arctan(\frac{y'}{x'})$ ou $(x', y')^T$ sont les coordonnées après la rotation définie par la relation (2.21) du site s. Le paramètre α définit l'angle d'ouverture du filtre papillon. Il dépend de la sélectivité angulaire de la sous bande considérée.
A partir du filtre $B_{\rho,\theta}$, deux filtres, scindant le profil de $B_{\rho,\theta}$ en deux parties complémentaires, sont déduits : $B_{\rho,\theta}(s) = B^0_{\rho,\theta}(s) + B^1_{\rho,\theta}(s)$. La figure 2.10 illustre le procédé utilisé. Un coefficient de facilitation η^{iso} est alors déterminé pour chaque site s de chaque sous bande (ρ, θ) :

$$\eta^{iso}_{\rho,\theta}(s) = \frac{L^0_{\rho,\theta}(s) + L^1_{\rho,\theta}(s)}{|L^0_{\rho,\theta}(s) - L^1_{\rho,\theta}(s))|} \tag{2.25}$$

avec,
$L^0_{\rho,\theta}(s) = R^{(2)}_{\rho,\theta}(s) * B^0_{\rho,\theta}(s)$ et $L^1_{\rho,\theta}(s) = R^{(2)}_{\rho,\theta}(s) * B^1_{\rho,\theta}(s)$.

La relation (2.25) est basée sur deux déclinaisons du principe de bonne continuation. La première, relative au numérateur de cette relation, exprime la bonne continuation directionnelle. La seconde, exprimée par le dénominateur, s'interprète comme un gain. Des réponses $L^0_{\rho,\theta}$ et $L^1_{\rho,\theta}$ proches expriment la bonne continuation en terme d'amplitude. Le gain, qui est l'inverse de la différence tendant alors vers zéro, augmente significativement. La sous bande renforcée, notée $R_{\rho,\theta,(3)}$, est obtenue via la relation suivante :

$$R_{\rho,\theta,(3)}(s) = R_{\rho,\theta,(2)}(s)(1 + \kappa^{iso}(s)\eta^{iso}_{\rho,\theta}(s)) \tag{2.26}$$

avec,

$$\kappa^{iso}(s) = \frac{\max_s(\eta^{iso}_{\rho,\theta}(s))}{\max_\theta(\max_s(\eta^{iso}_{\rho,\theta}(s)))} \tag{2.27}$$

La valeur κ^{iso} est obtenue en faisant le rapport entre le coefficient de renforcement local maximum de la sous bande considérée et le coefficient de renforcement global maximum. Un exemple de résultat est donné sur la figure 2.11 pour une image de test (image (a)).

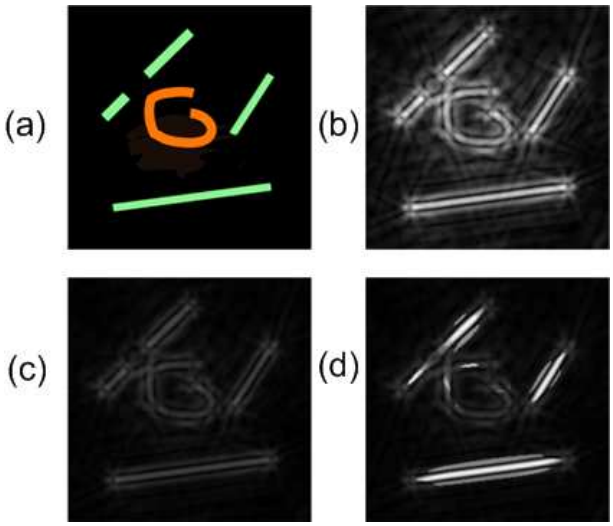

FIG. 2.11: Interactions facilitatrices sur une image test.

Cette image est utilisée afin de mettre clairement l'effet du procédé développé en exergue. L'image (b) de la figure 2.11 représente l'image normalisée avant l'application des interactions centre/pourtours. Les images (c) et (d), normalisées avec le maximum de l'image (d) représentent respectivement le résultat sans et avec renforcement des structures linéaires.

2.4.1.5 Construction de la carte et de la densité de saillance achromatique

L'espace psycho-visuel, dans lequel toutes les données sont homogènes en terme de visibilité, permet de déterminer facilement la carte de saillance CS^A. Une simple sommation des différentes sous bandes de la couronne II est effectuée :

$$CS^A(s) = \sum_{s,\rho,\theta} \left(R^A_{\rho,\theta,(3)}(s) \right) \tag{2.28}$$

Rappelons que les sous bandes achromatiques ont subi trois traitements, indiqués par l'indice $_{(3)}$: le renforcement achromatique, un procédé de modélisation des réponses des cellules corticales et un renforcement des contours co-linéaire alignés.
La carte CS^A ainsi obtenue donne pour chaque site son degré de saillance. Bien que cette carte soit intéressante, elle n'est pas en bonne adéquation avec notre système visuel. La carte de saillance donne localement la saillance d'un site alors que l'oeil ne traite pas l'information d'un point particulier mais plutôt celle d'une zone de notre environnement visuel ; cette zone ayant une taille proche de celle de la fovéa. Ainsi, la saillance d'une

zone dépend du nombre de sites saillant qu'elle possède et de leur valeur de saillance. Un filtrage par une gaussienne bi-dimensionnelle est donc appliqué pour déterminer la densité de saillance achromatique, notée DS^A. Elle s'obtient par la relation suivante :

$$DS^A(s) = CS^A(s) * g_\sigma(s) \tag{2.29}$$

avec g_{σ_x,σ_y} une gaussienne bi-dimensionnelle. L'écart type σ a une valeur de un demi degré visuel.

2.4.2 Génération de saillance chromatique Cr_1 et Cr_2

La génération de la saillance chromatique pour Cr_1 et Cr_2 est plus directe que celle que nous venons de voir. Les deux cartes de saillance chromatique sont déduites après avoir supprimé leur redondance spatiale.

2.4.2.1 Suppression des données chromatiques redondantes

Comme pour les sous bandes achromatiques, le comportement des cellules corticales est modélisé. Ainsi, les valeurs des sites s des sous bandes R^{Cr_1} et R^{Cr_2} sont respectivement obtenues via les relations (2.30) et (2.31) pour la composante Cr_1 et pour la composante Cr_2 :

$$R^{Cr_1}_{(1)}(s) = H(R^{Cr_1}(s) - R^{Cr_1}(s) * w_{\sigma_x^{ex},\sigma_y^{ex},\sigma_x^{inh},\sigma_y^{inh}}(s)) \tag{2.30}$$

$$R^{Cr_2}_{(1)}(s) = H(R^{Cr_2}(s) - R^{Cr_2}(s) * w_{\sigma_x^{ex},\sigma_y^{ex},\sigma_x^{inh},\sigma_y^{inh}}(s)) \tag{2.31}$$

avec,

$$w_{\sigma_x^{ex},\sigma_y^{ex},\sigma_x^{inh},\sigma_y^{inh}}(s) = \frac{1}{\left\| H(DoG_{\sigma_x^{ex},\sigma_y^{ex},\sigma_x^{inh},\sigma_y^{inh}}) \right\|} H(DoG_{\sigma_x^{ex},\sigma_y^{ex},\sigma_x^{inh},\sigma_y^{inh}}(s')) \tag{2.32}$$

Les différentes composantes de la relation (2.32) sont les mêmes que celles décrites dans le paragraphe 2.4.1.3.

2.4.2.2 Construction des cartes et des densités de saillance chromatique

Comme précédemment, les différentes sous bandes chromatiques sont issues de l'espace psycho-visuel. Elles sont homogènes en terme de visibilité. Les cartes de saillance chromatique CS^{Cr_1} et CS^{Cr_2} sont donc obtenues de la même façon que la carte de saillance achromatique :

$$CS^{Cr_1}(s) = \sum_{s,\rho,\theta} \left(R^{Cr_1}_{\rho,\theta,(1)}(s) \right) \tag{2.33}$$

$$CS^{Cr_2}(s) = \sum_{s,\rho,\theta} \left(R^{Cr_2}_{\rho,\theta,(1)}(s) \right) \tag{2.34}$$

Les sous bandes de la couronne II des composantes couleurs ont subi un seul traitement (procédé de modélisation des réponses des cellules corticales).

La carte de saillance représente le degré de saillance de chaque site. Le passage à la densité de saillance permet de prendre en compte la population de sites saillants dans un voisinage donné. Comme pour la densité de saillance achromatique, les densités de saillance, notées DS^{Cr_1} et DS^{Cr_2} respectivement pour les composantes Cr_1 et Cr_2, sont obtenues via la relation (2.29) précédemment décrite.

2.4.3 Création de la densité de saillance spatiale finale

L'espace psycho-visuel a permis de créer un ensemble homogène en normalisant par le seuil de visibilité chaque valeur de chaque composante extraite de l'image incidente. A partir de ce cadre conceptuel cohérent, les valeurs de visibilité de chaque composante ont été transformées par plus ou moins de mécanismes pour aboutir à une valeur de saillance. A partir des trois densités de saillance DS^A, DS^{Cr_1} et DS^{Cr_2}, une densité de saillance finale doit être déterminée. Bien que la normalisation des différentes composantes soit cohérente, il reste à traiter un problème de fusion. En effet, par quel mécanisme, peut-on fusionner les données de haut niveaux que sont les densités de saillance. Comment peut-on construire une densité de saillance finale à partir de données ayant des dynamiques différentes de saillance ? De façon plus pragmatique, quelle valeur de saillance peut-on attendre d'un triplet de données $(12, 4, 9)$, représentant la saillance locale des composantes A, Cr_1 et Cr_2 d'un site donné ? Quel est le poids de chaque valeur de saillance dans la constitution de la saillance finale ?
Il existe très peu de littératures sur ce sujet. Pour tenter de résoudre ce problème, nous nous basons sur les éléments de réponses apportés par les travaux de R. Milanese [Milanese 93] et ceux de L. Itti [Itti 01b] qui ont été respectivement évoqués aux paragraphes 2.3.2.3 et 2.3.2.2 de la partie I. On se place donc dans une approche empirique, c'est à dire une approche faisant appel à des algorithmes de traitement d'images dont l'unique objectif est de produire les meilleurs résultats possibles. Les méthodes utilisées sont donc susceptibles de diverger sensiblement du fonctionnement du système visuel. En dépit de cela, nous essaierons dans la mesure du possible de garder un maximum de cohérence avec le fonctionnement du système visuel, en justifiant chacune des étapes utilisées. Il est évident que le traitement rigoureux de ce problème est un sujet à part entière nécessitant des expérimentations psychophysiques poussées.

Dans la suite du document, deux approches de fusion sont présentées. La première est la méthode la plus simple qui puisse être envisagée. Elle est basée sur une sommation après l'utilisation d'une simple normalisation. La deuxième est bien plus innovante et bien plus cohérente avec notre système visuel. Cette méthode s'articule sur deux notions fondamentales exposées par R. Milanese : la complémentarité et la redondance des cartes de saillance à fusionner. En d'autres termes, une fusion cohérente doit prendre en compte un procédé de compétition intra carte ainsi qu'un procédé de compétition inter cartes.

2.4.3.1 Fusion naïve

La fusion naïve consiste à normaliser les trois densités de saillance DS^A, DS^{Cr_1} et DS^{Cr_2} de façon à obtenir la même dynamique. La normalisation d'une carte C, notée $\mathcal{N}(C)$, utilise le maximum global déterminé sur la carte C. La densité de saillance finale

DS est alors simplement obtenue par une simple addition :

$$DS(s) = \mathcal{N}(\mathcal{N}(DS^{A}(s)) + \mathcal{N}(DS^{Cr1}(s)) + \mathcal{N}(DS^{Cr2}(s))) \tag{2.35}$$

Le seul avantage de cette méthode est sa simplicité. Par contre, les inconvénients sont nombreux :

- cette méthode ne fait pas la distinction entre une densité de saillance présentant une distribution quasi uniforme et une densité présentant un ou plusieurs pics de saillance ;
- lorsque plusieurs pics de saillance sont présents dans une densité de saillance, ce type de fusion favorise clairement le pic de saillance le plus élevé ;
- cette méthode est très sensible au bruit impulsionnel ;
- il n'y a aucune interaction inter cartes.

2.4.3.2 Fusion cohérente

La méthode de fusion présentée dans ce paragraphe s'inspire des deux concepts évoqués par R. Milanese. Pour qu'une fusion soit performante, elle doit mettre en place deux procédés de compétition :

- une compétition intra permettant d'identifier les zones les plus pertinentes de la densité ;
- une compétition inter cartes tirant profit de la redondance et de la complémentarité des différentes densité. L'utilisation de la redondance inter cartes permet de renforcer la saillance de certaines zones lorsque celles-ci génèrent de la saillance dans plusieurs dimensions. Par contre, lorsqu'une zone ne génère de la saillance que dans une seule dimension visuelle, il est nécessaire d'utiliser la complémentarité inter cartes.

La fusion cohérente est présentée pour deux cartes, notées DS^{C_1} et DS^{C_2} issues d'une composante C_1 et C_2. La généralisation à n densités est facilement envisageable. La densité finale, notée DS, est obtenue par la fusion des cartes DS^{C_1} et DS^{C_2}, notée $\mathcal{F}(DS^{C_1}, DS^{C_2})$:

$$DS(s) = \mathcal{F}(DS^{C_1}(s), DS^{C_2}(s)) \tag{2.36}$$

L'opérateur de fusion $\mathcal{F}(\cdot)$ est composé d'une série de trois transformations que nous allons décrire. Ces trois transformations s'utilisent successivement.

Une étape de normalisation

Tout d'abord, un procédé de fusion ne peut se faire sans une étape préalable de normalisation de dynamique. Contrairement aux procédés de fusion proposés par L. Itti qui utilisaient une normalisation à partir du maximum global de chaque carte, la normalisation que nous utilisons se base sur le maximum empirique de chaque dimension visuelle. Ces maximums sont déterminés expérimentalement en utilisant des tests particuliers. Par exemple, pour la composante Cr_1, une image à luminance uniforme mais présentant un motif rouge saturé va générer une dynamique proche de la dynamique maximale de l'axe visuel Cr_1. La répétition de ce type d'expérimentation a permis de définir les maximum empiriques des composantes A, Cr_1 et Cr_2.

FIG. 2.12: Exemple de recherche des maximum locaux sur la composante achromatique A originale de l'image *Lighthouse*.

Les deux cartes de densité DS^{C_1} et DS^{C_2} sont donc normalisées pour être sur la même dynamique. Ensuite, afin de construire un histogramme, ces données sont quantifiées linéairement sur L niveaux. Elles sont respectivement notées $DS^{C_1}_{NQ}$ et $DS^{C_2}_{NQ}$.

Compétition intra carte

La compétition intra carte modifie la valeur de chaque site s des cartes $DS^{C_1}_{NQ}$ et $DS^{C_2}_{NQ}$ en fonction de la valeur du maximum local le plus proche. Ce type de compétition est donné par la relation suivante :

$$intraMap^{C_1}(s) = \frac{DS^{C_1}_{NQ}(s)}{PlusProcheMax_{C_1}(s)} \tag{2.37}$$

$$intraMap^{C_2}(s) = \frac{DS^{C_2}_{NQ}(s)}{PlusProcheMax_{C_2}(s)} \tag{2.38}$$

La fonction $PlusProcheMax_{C_1}$ (respectivement $PlusProcheMax_{C_2}$) retourne la valeur du maximum local de la composante C_1 (respectivement C_2) la plus proche de la valeur du site s. Cette valeur est extraite de la liste $\mathcal{L}_1$ (respectivement $\mathcal{L}_2$) de taille k_1 (respectivement k_2) valeurs. La taille des listes est déterminée de façon à avoir un rapport entre le maximum local n et le maximum local $n+1$ supérieur à un seuil, fixé arbitrairement à 1.3. Cet artifice permet de prendre uniquement en compte les principales zones de saillance. Par ailleurs, le maximum local $n+1$ est déterminé en inhibant une zone circulaire centrée autour du maximum local n et d'un rayon de un degré visuel, reproduisant une sélection de type *Winner-Take-All*. La figure 2.12 présente le procédé de recherche des maximum locaux sur la composante A originale de l'image *Lighthouse*.

Compétition inter cartes

La compétition inter cartes tire profit de la redondance et de la complémentarité des différentes cartes. Le terme $interMap$, lié à la compétition inter cartes, est donné par la relation suivante :

$$interMap(s) = complementarite(s) + redondance(s) \tag{2.39}$$

La complémentarité, notée *complementarite* dans la relation (2.39) s'obtient en sommant les résultats de la compétition intra carte :

$$complementarite(s) = intraMap^{C_1}(s) + intraMap^{C_2}(s) \tag{2.40}$$

La redondance inter cartes est traitée à partir d'une analyse conjointe des distributions des cartes à fusionner. Elle est notée *redondance* et donnée par la relation (2.41) :

$$redondance(s) = intraMap^{C_1}(s) \times intraMap^{C_2}(s) \frac{Log \frac{N}{H(DS_{NQ}^{C_1}(s), DS_{NQ}^{C_2}(s))}}{3Log(L)} \tag{2.41}$$

avec, N le nombre de sites des cartes considérées.

Le facteur $\frac{1}{3Log(L)} Log \frac{N}{H(DS_{NQ}^{C_1}(s), DS_{NQ}^{C_2}(s))}$ déduit de l'histogramme conjoint des cartes $DS_{NQ}^{C_1}$ et $DS_{NQ}^{C_2}$ modifie la valeur du site s considéré en fonction de sa probabilité d'apparition. Cette approche statistique est inspirée des travaux de A. Oliva [Oliva 01] et de ceux de B. Bruce [Bruce 03] que nous avons évoqué au paragraphe 2.3 de la partie I. Ces travaux utilisent le fait que la quantité d'informations portée par un site s est inversement proportionnelle à sa probabilité d'apparition. Par conséquent, le facteur déduit de l'analyse conjointe augmente la valeur d'un site s lorsque sa probabilité d'apparition est faible. Réciproquement, la valeur du site s est diminuée lorsque sa probabilité d'apparition est forte.
L'opérateur de fusion $\mathcal{F}$ est donc équivalent au terme $interMap$. Ce dernier intègre à la fois la compétition intra carte et la compétition inter cartes.

Application à la fusion des cartes DS^A, DS^{Cr_1} et DS^{Cr_2}
La fusion des cartes DS^A, DS^{Cr_1} et DS^{Cr_2} est réalisée via une approche hiérarchique. Une densité de saillance chromatique est d'abord déterminée. Ensuite, la densité de saillance finale DS^{SP}, représentant l'intérêt spatial de chaque pixel, est obtenue en fusionnant la carte de saillance achromatique et chromatique :

$$DS^{SP}(s) = \mathcal{F}(DS^A, \mathcal{F}(DS^{Cr_1}, DS^{Cr_2})) \tag{2.42}$$

L'utilisation de l'approche hiérarchique reproduit l'un des principaux aspects de notre système visuel. En effet, comme nous avons pu le voir dans différents paragraphes, le système visuel extrait les informations à partir d'une analyse hiérarchique du signal lumineux.

2.5 Performance de la modélisation sur images fixes

2.5.1 Évaluation qualitative

L'évaluation qualitative consiste en une simple comparaison subjective entre les données oculométriques et les cartes de saillance prédites. Le résultat de ce type d'évaluation doit être considéré comme une simple indication. Aucune conclusion sérieuse ne peut émaner de ce genre de méthode pour diverses raisons. La raison la plus évidente est la façon de présenter les résultats et de les afficher. Sur la figure 2.13, des exemples de cartes de saillance réelles et prédites sont données. La colonne (a) présente les images originales. Les colonnes (b) et (c) présentent respectivement les densités de saillance oculométrique non modifiées et modifiées par une loi Gamma. La loi Gamma augmente la dynamique des zones sombres et diminue celle des zones claires. La similarité entre les

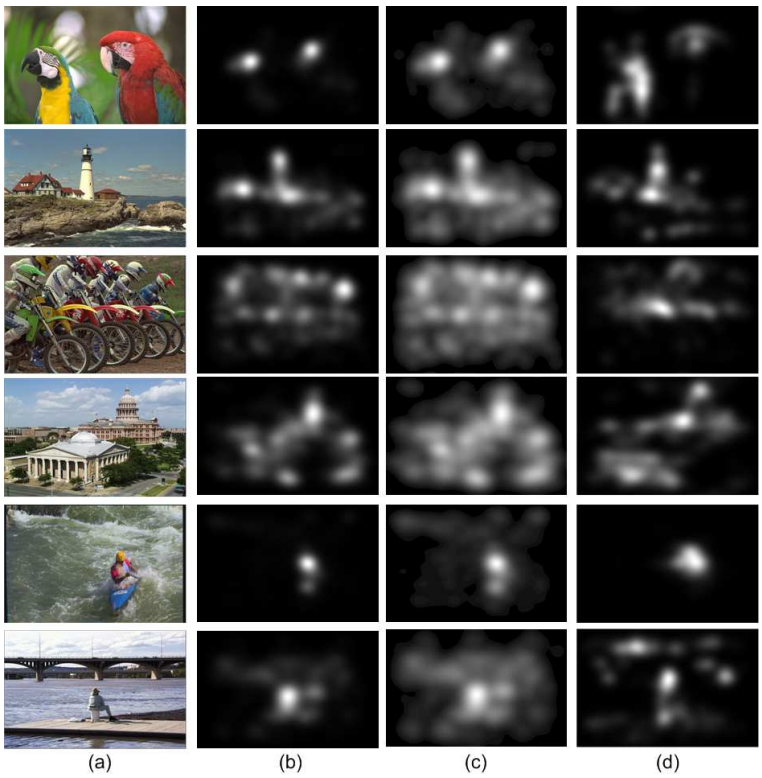

FIG. 2.13: Comparaison qualitative des densités de saillance issues des données oculométriques non modifiées (b), modifiées par une loi Gamma (c) et les densités de saillance de la modélisation (d). La durée d'observation est de 14 secondes.

cartes (b) et (d) est faible. Par contre lorsque la loi Gamma est appliquée, la similarité est nettement améliorée. La représentation sur 255 niveaux de gris pose ici un véritable problème de normalisation de dynamique. La façon de convertir la dynamique de saillance observée sur 255 niveaux de gris conditionne complètement l'appréciation qualitative des résultats. En dépit de cela, et au regard de la figure 2.13, il semble que le modèle détecte relativement correctement les zones les plus saillantes. Cela est particulièrement appréciable sur la seconde, la quatrième, la cinquième et la sixième image de la colonne (a).

2.5.2 Évaluation quantitative

2.5.2.1 Coefficient de corrélation

Le coefficient de corrélation linéaire donne des informations sur l'existence d'une relation linéaire entre les deux grandeurs considérées. Il varie entre -1 et 1. Un coefficient de corrélation nul signifie l'absence de toute relation linéaire entre les deux grandeurs. Il peut néanmoins exister une relation non linéaire entre elles. Le signe du coefficient indique le sens de la relation linéaire liant les grandeurs. Sa valeur absolue indique, quant à elle, l'intensité de la relation. Dans la suite, nous nous intéressons uniquement à la valeur absolue de ce coefficient.
Le coefficient de corrélation, noté cc, entre les données expérimentales oculométriques,

notées h, et les prédictions, notées p, est obtenu par la relation suivante :

$$cc(p,h) = \frac{cov(p,h)}{\sigma_p \sigma_h} \tag{2.43}$$

avec,
σ_h et σ_p représentent respectivement l'écart-type de la carte de densité de saillance de l'observateur moyen et de la modélisation. $cov(p,h)$ représente la covariance des valeurs h et p.
Les résultats quantitatifs de la modélisation sont commentés suivant trois axes :

- le premier concerne à la fois l'évaluation des performances intrinsèques du modèle et la comparaison avec un modèle gaussien (détaillé dans la suite) et le modèle de L. Itti. Les résultats sont donnés pour une durée d'observation de quatorze secondes. Ce choix est critiquable mais justifié dans la suite du paragraphe. En effet, l'objectif de ces travaux est de simuler le mécanisme *Bottom-Up* de l'attention visuelle. Ce mécanisme est considéré comme extrêmement rapide et éphémère disparaissant au profit du mécanisme *Top-Down*. Par conséquent, la logique aurait voulu que l'évaluation du modèle se fasse avec une durée d'observation très courte. Néanmoins, nous avons choisi de prendre une durée d'observation relativement importante pour les raisons suivantes :
  1. aujourd'hui, aucune étude n'a permis d'établir concrètement la contribution des mécanismes *Bottom-Up* et *Top-Down* en fonction du temps d'observation. Nous savons simplement que la contribution du mécanisme *Bottom-Up* est maximale lors des premières secondes de visualisation. Mais, comment évolue cette contribution ensuite ? Les travaux de D. Parkhurst [Parkhurst 02] ont tenté de répondre à cette question. Bien qu'aucune généralisation quantitative ne soit donnée concernant la contribution de tel ou tel mécanisme, il apparaît clairement que le mécanisme *Bottom-Up* ne se dissipe pas après les premières fixations. Sa contribution reste élevée. Nous verrons que nos travaux viennent parfaitement confirmer ce résultat ;
  2. par ailleurs, il est maintenant connu que l'observateur porte son attention de façon privilégiée et de façon répétitive sur les zones visuellement importantes, au détriment du reste. Cette propriété, que nous avons constatée et commentée dans le paragraphe 1.3.3 relatif à nos expérimentations oculométriques, se traduit par un taux de couverture qui n'augmente pas linéairement avec le temps d'observation. Par conséquent, afin de discriminer plus facilement les zones visuellement importantes du reste, il est intéressant de considérer un temps d'observation relativement important ;
  3. de plus, les premières fixations des observateurs sont fortement biaisées du fait du protocole expérimental qui a tendance à favoriser le centre de l'écran. C'est d'ailleurs le défaut de l'étude de D. Parkhurst. Néanmoins, nous confirmons ses conclusions en travaillant sur des durées de visualisation élevées, minimisant ainsi l'influence du centre de l'image ;
  4. bien que nous ayons privilégié cette durée d'observation, des tests complémentaires, détaillés dans la dernière partie de ce paragraphe, permettent d'évaluer l'influence de la durée d'observation sur les performances du modèle.

- le second axe d'évaluation utilise les conclusions de la précédente évaluation : le modèle est en effet couplé avec une information cognitive permettant d'attribuer plus d'importance au centre de l'image qu'à sa périphérie. Le second volet de ce paragraphe concerne donc l'évaluation d'un modèle modifié par une information de haut niveau ;
- finalement, nous avons étudié l'influence de la durée d'observation sur les performances de la modélisation.

TAB. 2.1: Corrélation linéaire entre les données réelles obtenues pour une durée d'observation de 14s et différents modèles : le modèle gaussien, le modèle proposé avec une fusion naïve, avec une fusion cohérente et le modèle de L. Itti.

14s	gaussien	naïf	cohérent	Itti
vautour692Couleur	0.44	0.18	0.32	xx
vautour825Couleur	0.40	0.25	0.34	xx
bikes	0.42	0.52	0.66	xx
ocean	0.70	0.22	0.38	xx
paintedhouse	0.61	0.40	0.10	xx
stream	0.60	0.02	0.34	xx
churchandcapitol2	0.55	0.50	0.61	0.35
vautour538Couleur	0.68	0.41	0.51	0.28
patinCouleur	0.82	0.53	0.56	0.29
lighthouse2	0.61	0.68	0.70	0.57
dancers2	0.70	0.66	0.75	0.28
sailing1	0.64	0.24	0.31	0.45
kayakCouleur	0.43	0.44	0.40	0.38
manfishing	0.77	0.49	0.58	0.26
zebres797Couleur	0.70	0.05	0.40	0.38
parrots	0.28	0.49	0.56	xx
plane	0.56	0.52	0.62	xx
rapids	0.53	0.53	0.52	0.45
Moyenne	0.58	0.40	0.48	xx
Moyenne sur le sous-ensemble Itti	0.64	0.45	0.54	0.37

Performances intrinsèques du modèle

Le tableau 2.1 présente les performances de plusieurs modèles en terme de coefficient de corrélation. Deux déclinaisons de notre modélisation et deux autres modèles sont comparés. Dans l'ordre d'apparition du tableau 2.1, les modèles testés sont :

- le modèle gaussien obtenu via une gaussienne bi-dimensionnelle centrée sur l'image. Son étendue spatiale a été optimisée sur une base de 18 images. Sa valeur est de 2.5 degrés visuel ;
- le modèle dit naïf correspondant à notre modélisation intégrant la fusion naïve ;
- le modèle dit cohérent correspondant à notre modélisation intégrant la fusion cohérente ;

– le modèle de L. Itti[2].

Plusieurs remarques émanent du tableau 2.1 :

– *comparaison des techniques de fusion, dites naïve et cohérente* : la fusion cohérente présente des performances moyennes sensiblement supérieures. Le coefficient de corrélation augmente de 0.08 (soit 20%). Deux régressions sont à noter sur les images *KayakCouleur* et *Paintedhouse*. La diminution du coefficient de corrélation sur la première image s'explique aisément. Cette image ne contient qu'une région d'intérêt fortement colorée. Le fait de normaliser chaque carte de saillance classiquement (c'est à dire en prenant le maximum local de chaque carte) permet d'accroître significativement la saillance de cette zone. Par contre, l'utilisation d'un maximum théorique pour chacune des composantes ne permet pas de retrouver ce niveau de saillance. L'image *Paintedhouse* est à considérer comme un cas particulier (l'image source est présentée à l'annexe A). Cette image ne présente pas véritablement de régions d'intérêt. L'aspect sémantique domine dans cette image et conditionne les résultats expérimentaux. Notre modèle, quant à lui, détecte les fortes transitions (bords de la toiture). Cet effet est particulièrement accentué lorsque la fusion cohérente est utilisée ;
– *comparaison entre les deux déclinaisons de notre modèle et celui de L. Itti* : pour effectuer cette comparaison, les coefficients de corrélation moyens sont calculés sur le sous-ensemble d'images de L. Itti. On constate que le modèle proposé améliore le coefficient de corrélation moyen de 0.08 et 0.17 respectivement pour la fusion naïve et pour la fusion cohérente. L'amélioration relative au modèle naïf signifie que le modèle proposé est intrinsèquement meilleur que celui de L. Itti. Par ailleurs, un test statistique de Student (t-test) effectué entre les données du modèle cohérent et celle de L. Itti confirme que les deux ensembles sont statistiquement différents. La valeur de Student est de 2.923, ce qui donne une probabilité de 0.016 d'avoir obtenu ces résultats par chance ;
– *comparaison des modèles psycho-visuels avec un modèle Gaussien* : les résultats sont tout à fait intéressants. L'utilisation d'un modèle gaussien peut être légitimée pour différentes raisons. Tout d'abord, il existe une forte dépendance vis à vis du centre de l'écran. Cette dépendance est liée à un aspect cognitif. Nous sommes, en effet, très familiers aux contenus vidéo ou la région d'intérêt est centrée. Par ailleurs, le protocole expérimental des tests oculométriques à tendance à renforcer ce phénomène. Ces deux remarques expliquent en grande partie les bons résultats obtenus par ce modèle. Il améliore le coefficient de corrélation moyen notablement. Néanmoins, il est intéressant de noter que notre modèle présente de meilleurs résultats pour un certain nombre d'images : *bikes*, *churchandcapitol2*, *lighthouse2*, *dancers2*, *parrots*, *plane*. L'explication est simple : ces images présentent soient des régions d'intérêt non centrées soient des régions d'intérêt de grandes tailles. Afin de coupler cet aspect de dépendance au centre et la modélisation proprement dite, les cartes de saillance prédites sont pondérées par le modèle gaussien. L'interprétation des résultats fait l'objet du paragraphe suivant.

[2]On ne dispose pas des cartes de saillance pour toutes les images de tests. Pour ces dernières, la valeur de corrélation est remplacée par xx. Ce sous-ensemble est appelé sous-ensemble de L. Itti.

TAB. 2.2: Corrélation linéaire entre les données réelles obtenues pour une durée d'observation de 14s et différents modèles : le modèle gaussien, le modèle proposé avec une fusion naïve, avec une fusion cohérente et le modèle de L. Itti, tous couplés avec le modèle gaussien.

14s	gaussien	naïf	cohérent	Itti
vautour692Couleur	0.44	0.54	0.60	xx
vautour825Couleur	0.40	0.54	0.54	xx
bikes	0.42	0.42	0.47	xx
ocean	0.70	0.85	0.84	xx
paintedhouse	0.61	0.55	0.56	xx
stream	0.60	0.58	0.64	xx
churchandcapitol2	0.55	0.62	0.61	0.54
vautour538Couleur	0.68	0.78	0.77	0.66
patinCouleur	0.82	0.87	0.82	0.83
lighthouse2	0.61	0.68	0.71	0.64
dancers2	0.70	0.75	0.74	0.77
sailing1	0.64	0.69	0.68	0.60
kayakCouleur	0.43	0.68	0.62	0.45
manfishing	0.77	0.85	0.82	0.79
zebres797Couleur	0.70	0.65	0.66	0.70
parrots	0.28	0.42	0.46	xx
plane	0.56	0.79	0.84	xx
rapids	0.53	0.69	0.55	0.58
Moyenne	0.58	0.66	0.66	xx
Moyenne sur le sous-ensemble Itti	0.64	0.73	0.70	0.66

Performances du modèle couplé au modèle gaussien

Le tableau 2.2 présente les coefficients de corrélation moyens lorsqu'une pondération gaussienne centrée sur l'image est couplée aux modèles psycho-visuels. Les résultats de tous les modèles sont considérablement améliorés. L'application de cette fonction profite particulièrement au modèle de L. Itti. Ses performances en terme de corrélation passent de 0.37 à 0.66, soit un gain de 78%. Concernant notre modèle, le gain est de 69% et de 29% respectivement pour le modèle naïf et le modèle cohérent. Les performances des trois modèles précités présentent maintenant des coefficients de corrélation moyens supérieurs à celui du modèle gaussien. Le modèle naïf et le modèle cohérent présentent les meilleures performances. Leurs performances sont similaires sur toute la base. Sur le sous-ensemble de L. Itti, le modèle naïf présente les meilleures performances. Le test de student, effectué entre les valeurs du modèle cohérent et celui de L. Itti donne cette fois une probablité de 0.08 que les moyennes obtenues ne soient pas significativement différentes (Valeur de la variable de Student de 1.9).

En conclusion, l'apport de la fonction de pondération est dans ce contexte (visualisation d'images sur un écran et évaluation du coefficient de corrélation) un atout indéniable. Elle n'est heureusement pas suffisante. Notre modélisation couplée avec cette fonction permet

en effet d'améliorer nettement ces résultats. Néanmoins, il est frustrant de constater qu'une simple gaussienne puisse rivaliser avec une modélisation réfléchie et travaillée. Pour autant, soulignons une nouvelle fois que ces résultats sont intimement liés à la façon dont les cartes de saillance réelles sont déterminées. Le problème de l'obtention de la vérité terrain se pose une nouvelle fois : comment l'obtenir ? Quelles images utiliser ? Quel protocole ?...

Variation du coefficient de corrélation moyen en fonction de la durée d'observation

Le coefficient de corrélation moyen est donné dans le tableau 2.3 pour différentes durées d'observation (4, 10 et 14 secondes) et pour les trois modèles testés. Les résultats sont donnés avec et sans l'application de la fonction gaussienne.
L'augmentation du temps d'observation conduit à l'augmentation du coefficient de corrélation moyen de notre modèle et de celui de L. Itti. C'est tout à fait logique, puisque ces modèles n'intègrent pas le temps d'observation. La similarité est alors plus forte lorsque les données expérimentales sont obtenues avec une durée d'observation importante. Cette amélioration est moins remarquable lorsque la fonction d'excentricité est couplée avec ces modèles. Ceci est le résultat de la dépendance au centre de l'image.
Dans le tableau 2.3, le gain apporté par notre modélisation en fonction du temps d'observation, en prenant comme référence le modèle de L. Itti, apparaît dans la dernière colonne. Comme précédemment, lorsque la fonction d'excentricité n'est pas appliquée, le gain en performance se situe autour de 44%. La conception intrinsèque de notre modéle est donc plus performante que celle du modèle de L. Itti. L'application de la fonction d'excentricité réduit clairement ce gain.

Tab. 2.3: Évolution du coefficient de corrélation moyen en fonction du temps d'observation pour les différents types de modélisation. Le modèle proposé et le modèle de L. Itti sont évalués sans et avec le modèle gaussien.

Images	gaussien	cohérent		L. Itti		Gain de notre modélisation	
		sans	avec	sans	avec	sans	avec
$4s$	0.64	0.46	0.68	0.32	0.65	+43%	+4.62%
$10s$	0.65	0.51	0.70	0.35	0.66	+45%	+6.06%
$14s$	0.64	0.54	0.70	0.37	0.66	+45,9%	+6.06%

2.5.2.2 Divergence de Kullback-Leibler

Alors que le coefficient de corrélation estime le degré de similarité de deux ensembles, il peut être intéressant d'estimer la dissimilarité entre deux ensembles. A partir de divergence généralisée (ou ϕ-divergence) [Csiszar 67], l'expression qualifiant la divergence entre deux lois de probabilité p et h est définie par

$$D_\phi(p|h) = \sum_x h(x)\phi\left(\frac{p(x)}{h(x)}\right) \tag{2.44}$$

avec, ϕ une fonction convexe de $\mathcal{R}^+ - \{0\}$ dans $\mathcal{R}$. $D_\phi(p|h)$ est supérieur ou égal à $\phi(1)$ et l'égalité est constatée si et seulement si les lois de probabilité sont strictement égales. La divergence de Kullback et Leibler[3] est obtenue pour $\phi(x) = xLog(x)$. La relation précédente devient alors :

$$D(p|h) = \sum_x p(x)Log(\frac{p(x)}{h(x)}) \tag{2.45}$$

Dans la suite de ce paragraphe, h et p représentent respectivement la loi de probabilité de saillance provenant des expériences oculométriques et celle provenant de la modélisation. La pertinence des résultats du modèle sera d'autant plus grande vis à vis des données expérimentales que la distance qui les sépare est petite. Notons que cette expression est dissymétrique, et ne représente pas une distance au sens strict du terme puisque l'inégalité triangulaire n'est pas respectée.

Tab. 2.4: Divergence de Kullback-Leibler entre les données réelles obtenues et différents modèles : le modèle gaussien, le modèle proposé avec une fusion naïve, avec une fusion cohérente et le modèle de L. Itti. La durée d'observation est de 14s.

14s	gaussien	naif	cohérent	Itti
vautour692Couleur	0.63	1.93	1.44	xx
vautour825Couleur	1.30	1.97	1.46	xx
bikes	0.30	1.17	1.41	xx
ocean	0.63	2.40	1.60	xx
paintedhouse	0.44	1.92	1.40	xx
stream	0.50	1.88	1.25	xx
churchandcapitol2	0.53	1.10	0.65	0.88
vautour538Couleur	0.59	3.12	2.18	2.73
patinCouleur0	0.29	2.47	1.53	2.11
lighthouse2	0.80	1.53	0.95	1.01
dancers2	0.42	0.72	0.49	0.89
sailing1	0.51	2.33	1.55	1.30
kayakCouleur0	1.00	1.80	1.41	1.66
manfishing	0.32	1.34	0.79	1.24
zebres797Couleur	0.35	1.25	0.82	0.87
parrots	1.34	2.38	1.50	xx
plane	1.41	2.31	0.99	xx
rapids	0.91	2.13	1.64	1.87
Moyenne	0.68	1.87	1.28	xx
Moyenne sur le sous-ensemble Itti	0.57	1.78	1.20	1.46

Performances intrinsèques du modèle

Le tableau 2.4 présente les résultats de l'évaluation en terme de divergence de Kullback-

[3]D'autres exemples de fonctions ϕ sont possibles. Citons la divergence de Hellinger $\phi(x) = (\sqrt(x) - 1)^2$, de Wald $\phi(x) = 1 - min(x, 1)$, en variation $\phi(x) = |x - 1|$

Leibler. Les conclusions sont les mêmes que celles issues de l'évaluation en terme de coefficient de corrélation. Nous les rappelons brièvement ici :

- le modèle cohérent est plus performant que les autres modèles psycho-visuels (naïf et L. Itti) ;
- le modèle gaussien est le modèle le plus performant.

Performances du modèle couplé au modèle gaussien

Comme précédemment, les modèles psycho-visuels sont couplés au modèle gaussien. Les résultats apparaissent dans le tableau 2.5. Les mêmes tendances que celles précédemment évoquées sont constatées :

- le couplage des modèles psycho-visuels avec la fonction gaussienne permet d'accroître significativement les résultats ;
- les plus mauvais résultats sont obtenus pour le modèle gaussien ;
- le modèle intégrant la fusion cohérente est le modèle le plus pertinent.

TAB. 2.5: Divergence de Kullback-Leibler entre les données réelles obtenues et différents modèles : le modèle gaussien, le modèle proposé avec une fusion naïve, avec une fusion cohérente et le modèle de L. Itti, tous couplés avec le modèle gaussien. La durée d'observation est de $14s$.

14s	gaussien	naif	cohérent	Itti
vautour692Couleur	0.63	0.54	0.48	xx
vautour825Couleur	1.30	0.90	0.88	xx
bikes	0.30	0.71	0.54	xx
ocean	0.63	0.29	0.30	xx
paintedhouse	0.44	0.62	0.53	xx
stream	0.50	0.58	0.49	xx
churchandcapitol2	0.53	0.43	0.43	0.53
vautour538Couleur	0.59	0.45	0.43	0.64
patinCouleur0	0.29	0.22	0.27	0.23
lighthouse2	0.80	0.51	0.45	0.62
dancers2	0.42	0.33	0.34	0.31
sailing1	0.51	0.49	0.47	0.60
kayakCouleur0	1.00	0.71	0.71	1.07
manfishing	0.32	0.21	0.24	0.27
zebres797Couleur	0.35	0.46	0.39	0.40
parrots	1.34	1.00	0.93	xx
plane	1.41	0.48	0.32	xx
rapids	0.91	0.69	0.78	0.70
Moyenne	0.68	0.53	0.50	xx
Moyenne sur le sous-ensemble Itti	0.57	0.45	0.45	0.54

Évolution de la divergence de Kullback-Leibler moyenne en fonction de la durée d'observation

L'évolution de la valeur de la divergence de Kullbak-Leibler en fonction de la durée d'observation est donnée tableau 2.6. Comme précédemment, trois durées d'observation (4, 10 et 14 secondes) sont testées. Le modèle gaussien, le modèle de L. Itti et notre modèle associés ou non à la pondération gaussienne sont comparés. Tout d'abord, les meilleures performances sont données par notre modèle couplé au modèle gaussien excepté pour la durée d'observation la plus faible, pour laquelle le centre des images présente en moyenne la majorité de la saillance. Lorsque la durée d'observation augmente, la divergence entre les données réelles et prédites diminue logiquement (le temps d'observation n'est ni pris en compte dans notre modélisation ni dans celle de L. Itti). Enfin, le gain de notre modélisation en considérant le modèle de L. Itti comme la référence, est donné dans les deux dernières colonnes du tableau. Excepté pour un cas (pour une durée d'observation de 4 secondes), la diminution de la valeur de divergence, et donc le gain, se situe autour de 13%.

Tab. 2.6: Évolution du coefficient de Kullback-Leibler moyen en fonction du temps d'observation pour les différents types de modélisation. Le modèle proposé et le modèle de L. Itti sont évalués sans et avec le modèle gaussien.

Images	gaussien	cohérent		Itti		Gain de notre modélisation	
		sans	avec	sans	avec	sans	avec
$4s$	0.57	2.48	0.75	2.8	0.66	-11.4%	+13.64%
$10s$	0.55	1.37	0.49	1.63	0.56	-15.9%	-12.05%
$14s$	0.59	1.28	0.45	1.46	0.54	-12.32%	-16.67%

Comparaison avec la dispersion inter observateurs

Il est intéressant de qualifier la dispersion des densités de saillance des différents observateurs avec la carte de densité de saillance reproduisant le comportement d'un observateur moyen. Cela consiste à calculer la divergence de Kullback-Leibler de chaque observateur avec le comportement moyen constaté :

$$KL_{avg} = \frac{1}{N} \sum_i KL(h_i|h) \tag{2.46}$$

où,
h est la densité de probabilité obtenue en considérant toutes les données de tous les observateurs, h_i est la densité de probabilité associée à un observateur donné et N est le nombre d'observateurs considérés.

L'interprétation de la valeur de KL_{avg} s'effectue de la façon suivante :

- une forte valeur caractérise une hétérogénéité dans la stratégie visuelle des observateurs ;
- une faible valeur caractérise une certaine homogénéité dans la stratégie visuelle de tous les observateurs ;

– une valeur nulle, peu probable, est obtenue lorsque tous les observateurs ont regardé les mêmes zones de l'image pendant la même durée.

La valeur KL_{avg} qualifie le degré avec lequel l'attention visuelle peut être approchée. Dans le cas extrême ou la valeur de KL_{avg} est très forte, un modèle uniforme peut présenter les meilleures performances. Les valeurs KL_{avg} pour chaque image et pour deux durées d'observation sont données tableau 2.7. La première remarque concerne l'évolution temporelle de la moyenne des valeurs KL_{avg}. Cette valeur diminue lorsque la durée d'observation augmente. Si on considère à la fois l'augmentation du taux de couverture en fonction de la durée d'observation et le fait que les observateurs reviennent continuellement sur les régions saillantes, ce résultat est tout à fait cohérent. Par ailleurs, si on considère la durée d'observation de $14s$, les plus fortes valeurs de KL_{avg} correspondent aux images présentant soit de nombreuses régions d'intérêt (*vautour692Couleur*, *Bikes*) soit des régions d'intérêt peu évidentes (*Stream*, *Ocean*). En effectuant la comparaison avec les valeurs de divergence obtenues avec notre modélisation, on constate que la plupart des valeurs de divergence calculées à partir de nos prédictions n'excèdent pas cette valeur moyenne. Cela signifie qu'en moyenne, nous reproduisons correctement le comportement visuel moyen des observateurs et que le modèle retrouve correctement les zones les plus saillantes.

Tab. 2.7: Divergence de Kullback-Leibler entre les données réelles obtenues et celles du modèle proposé avec une fusion cohérente et la dispersion inter observateurs. Les durées d'observation sont de $4s$ et $14s$.

	$4s$		$14s$	
	KL_{avg}	modèle cohérent	KL_{avg}	modèle cohérent
vautour692Couleur	1.03	0.62	0.7	0.48
vautour825Couleur	0.73	1.19	0.54	0.88
bikes	1.09	0.54	0.71	0.54
ocean	1.08	0.24	0.69	0.30
paintedhouse	0.82	0.64	0.6	0.53
stream	1.2	0.48	0.77	0.49
churchandcapitol2	1.01	0.39	0.54	0.43
vautour538Couleur	0.74	0.84	0.51	0.43
patinCouleur0	0.82	0.40	0.6	0.27
lighthouse2	0.84	0.51	0.58	0.45
sailing1	0.89	0.82	0.57	0.47
kayakCouleur0	0.75	1.17	0.61	0.71
manfishing	0.7	0.56	0.54	0.24
parrots	0.66	1.05	0.48	0.93
plane	0.67	0.43	0.47	0.32
rapids	0.74	1.59	0.42	0.78
Moyenne	0.86	0.72	0.58	0.52

2.5.2.3 Matrice de confusion

Les deux précédentes méthodes d'évaluation sont nécessaires pour estimer le degré de pertinence des données brutes du modèle. La complémentarité de ces deux méthodes est également tout à fait intéressante. Néanmoins, elles sont dépendantes de plusieurs facteurs :

- pour le coefficient de corrélation, la façon dont évolue la saillance issue des tests oculométriques est un facteur important. Est-il en effet correct de dire qu'une zone ayant une valeur de saillance de 10 attire 10 fois plus l'attention qu'une zone ayant une saillance de 1 ? Rien ne nous permet de l'affirmer. Pour appréhender ce problème, l'estimation des performances du modèle a été également réalisée avec des données expérimentales modifiées par différents opérateurs non linéaires. Ces tests n'ont pas permis d'obtenir de meilleurs résultats ;
- pour la divergence de Kullback-Leibler, la dynamique des signaux comparés joue considérablement sur les résultats. Une étape de normalisation est en effet réalisée pour transformer la carte de saillance en densité de probabilité ;
- le protocole expérimental influe fortement sur les résultats donnés par les deux métriques (dépendance au centre de l'image).

Un autre type d'évaluation minimisant la dépendance vis à vis de ces facteurs s'avère donc nécessaire. Une méthode de classification, étiquetant chaque pixel de l'image comme étant d'intérêt ou non, est choisie. L'étiquette de chaque pixel est bien évidemment liée à la détermination d'un seuil. Nous avons choisi de considérer que la population des pixels d'intérêt ne devait pas excéder trente pour cent de la population totale. Plusieurs paramètres tels que le taux de couverture (sa définition est donnée au paragraphe 1.3.3 de la partie II) et notre propre expérience (visualisation des cartes de saillance expérimentales) ont contribué à la détermination empirique de ce seuil. La façon dont la méthode de classification est définie permet d'être totalement insensible aux deux premiers facteurs précités. L'évaluation des performances se fait ensuite classiquement via une matrice de confusion, détaillée tableau 2.8.

Observateurs-Prédiction	Saillant	Non saillant
Saillant	VP (Vrais Positifs)	FP (Faux Positifs)
Non saillant	FN (Faux Négatifs)	VN (Vrais Négatifs)

Tab. 2.8: Matrices de confusion associées à la classification

Deux paramètres sont extraits des matrices de confusion. Le premier est la précision du système, ou taux de recouvrement global, obtenue en effectuant le rapport des pixels bien classés sur la population totale (c'est en fait le rapport de la population portée par la diagonale sur la population globale). Ce premier paramètre est noté $\mathcal{P}$ et est obtenu par la relation suivante :

$$\mathcal{P} = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.47}$$

Le second paramètre est relatif à la capacité du système à détecter correctement les zones d'intérêt. Le taux de vrais positifs prédits effectivement bien classé, noté $\mathcal{V}$, est donné par

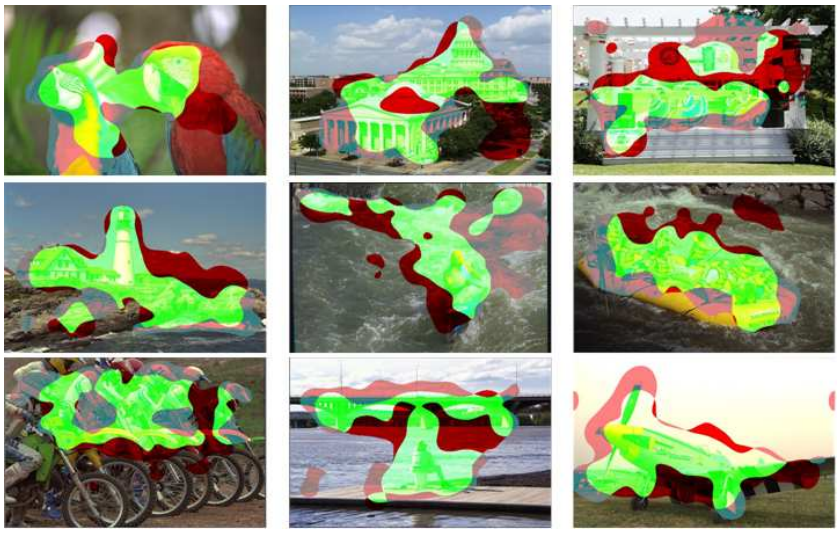

FIG. 2.14: Résultats de la classification semi-supervisée. Les zones vertes sont les zones saillantes bien classées (Vrais Positifs). Les zones rouges vifs sont des zones de saillance non détectées par le modèle (Faux Positifs). Les zones rouges pâles sont des zones d'intérêt détectées uniquement par le modèle (Faux Négatifs). Enfin, les zones non colorées sont des zones de non intérêt correctement détectées par le modèle (Vrais Négatifs).

la relation suivante :

$$\mathcal{V} = \frac{VP}{VP + FN} \tag{2.48}$$

Le tableau 2.9 liste l'évaluation de ces paramètres pour un ensemble d'images. Les résultats sont commentés dans les paragraphes suivants. La figure 2.14 illustre le résultat de la classification sur différentes images.

Précision globale $\mathcal{P}$ du modèle

En moyenne, la précision globale $\mathcal{P}$ du modèle sur les 18 images testées est de 77%. Cela signifie que 77% des prédictions de notre modèle sont cohérentes avec celles issues des tests oculométriques. Ce résultat est très encourageant puisqu'il est uniquement basé sur des attributs visuels de bas niveau.

Capacité de détection des vrais positifs $\mathcal{V}$ du modèle

La capacité du modèle à détecter les zones d'intérêt est correcte. La valeur moyenne $\mathcal{V}$ est proche de 63%. Cela signifie que 63% des zones de saillance prédites sont effectivement d'intérêt. Notons qu'un calcul complémentaire n'apparaissant pas dans cette évaluation (donnant quasiment la même valeur que la précédente) consiste à prendre non plus comme référence la population des vrais positifs ($VP + FN$) mais la population des vrais positifs de la référence ($VP + FP$).

Comparaison avec le modèle de L. Itti

Les valeurs de précision globale et de capacité de détection des zones saillantes données par la modélisation proposée sont meilleures que celles données par le modèle de L. Itti.

TAB. 2.9: Résultats de la classification pour le modèle proposé et le modèle de L. Itti. La précision et la capacité de détection des vrais positifs sont données. Les résultats sont donnés pour une durée d'observation de $14s$.

	Modèle cohérent		Itti	
	$\mathcal{P}$	$\mathcal{V}$	$\mathcal{P}$	$\mathcal{V}$
vautour692Couleur	0.66	0.45	xx	xx
vautour825Couleur	0.78	0.64	xx	xx
bikes	0.79	0.71	xx	xx
ocean	0.83	0.72	xx	xx
paintedhouse	0.75	0.58	xx	xx
stream	0.53	0.21	xx	xx
churchandcapitol2	0.69	0.49	0.72	0.57
vautour538Couleur	0.81	0.68	0.65	0.59
patinCouleur0	0.81	0.69	0.71	0.42
lighthouse2	0.79	0.65	0.72	0.52
dancers2	0.85	0.75	0.77	0.53
sailing1	0.84	0.73	0.65	0.63
kayakCouleur0	0.72	0.54	0.74	0.42
manfishing	0.85	0.75	0.69	0.57
zebres797Couleur	0.69	0.49	0.73	0.49
parrots	0.85	0.75	xx	xx
plane	0.85	0.75	xx	xx
rapids	0.83	0.72	0.74	0.56
Moyenne	0.77	0.63	xx	xx
Moyennne sous-ensemble Itti	0.79	0.65	0.71	0.53

C'est le deuxième paramètre qui est le plus remarquable : le modèle de L. Itti commet 47% $(100 - 53)$ d'erreurs sur la détection des vrais positifs, alors que nous commettons 35% $(100 - 65)$ d'erreurs. Cela confirme donc les précédents résultats.

Cette méthode d'évaluation confirme les résultats obtenus avec le coefficient de corrélation linéaire et la divergence de Kullback-Leibler. Le modèle d'attention visuelle proposé peut donc être considéré comme plus performant que celui de L. Itti.

Remarque :

Cette méthode d'évaluation est insensible à la dynamique des signaux comparés et à la façon dont elle varie. Elle reste néanmoins sensible à l'application d'une fonction gaussienne ou d'excentricité. En effet, la valeur $\mathcal{P}$ passe de 0.77 à 0.82. Concernant la valeur $\mathcal{V}$, elle est augmentée de 0.08 (passage de 0.63 à 0.71). Les résultats détaillées de l'application de la fonction gaussienne sur nos prédictions n'apparaissent pas dans ce mémoire.

2.6 Conclusion

L'objectif de ce chapitre était de présenter notre contribution dans le domaine de la modélisation de l'attention visuelle sur images couleurs fixes. Bien que basé comme la majorité des modèles d'attention visuelle sur l'architecture proposée par C. Koch et S. Ullman, notre contribution se différencie de l'existant par la construction d'un espace psycho-visuel. Ce dernier contient des informations relatives à la composante achromatique et aux deux composantes chromatiques de l'image à analyser, toutes exprimées en fonction de leur seuil de visibilité. Le modèle proposé est qualifié par le terme cohérent puisque, avant même de chercher à déterminer les zones saillantes d'une image, la visibilité de chaque site de chaque composante est calculée. Ce calcul est fondé sur des modèles mathématiques déduits d'expériences psychophysiques.
Après cette étape, l'ensemble des données de l'espace psycho-visuel, exprimé en terme de visibilité, peut être utilisé pour déterminer les zones saillantes. Concernant la détermination de la saillance spatiale, on propose de créer trois cartes de saillance : une carte de saillance pour la composante achromatique et une carte pour chacune des composantes chromatiques. Les mécanismes utilisés ont tous pour objectif de hiérarchiser les données afin de faire apparaître les zones qui "sautent aux yeux", présentant un fort contraste local. La création de la saillance spatiale est un problème délicat car de nombreuses dimensions visuelles sont susceptibles d'intervenir. Nous avons considéré les deux principales, la luminance et la couleur ; d'autres types de mécanismes basés sur les textures par exemple sont certainement complémentaires de ceux mis en oeuvre.
Enfin, les performances du modèle proposé vis à vis d'une vérité terrain acquise lors d'expérimentations oculométriques ont été évaluées. Des comparaisons qualitatives indiquant une certaine ressemblance ont été confirmées par des méthodes objectives de comparaison. Trois méthodes ont été utilisées, chacune présentant des caractéristiques intéressantes. La première, le coefficient de corrélation linéaire, permet de mesurer le degré de linéarité existant entre deux ensembles de données. La seconde, la divergence de Kullback-Leibler, mesure le degré de dissimilarité de deux densités de probabilité. En d'autres termes, ce sont les erreurs de modélisation qui contribuent à déterminer la note finale. Enfin, pour pallier certains défauts de ces deux approches (dépendance à la dynamique des signaux, par exemple), la capacité du modèle à prédire correctement les zones d'intérêt et de non intérêt est évaluée à partir d'une matrice de confusion. Toutes ces méthodes ont permis de confirmer le sentiment subjectif issu des évaluations qualitatives. Par ailleurs, une comparaison avec le modèle de L. Itti a été effectuée permettant de positionner avantageusement notre modèle en terme de performances.

Chapitre 3

Extension du modèle à la dimension temporelle. Évaluation des performances

3.1 Introduction

L'objet de ce chapitre est de présenter un exemple d'extension à la dimension temporelle du modèle spatial de l'attention visuel précédemment décrit. Alors que la saillance spatiale fait intervenir de nombreuses dimensions (achromatique, chromatiques...), la détermination de la saillance temporelle s'avère plus directe. En effet, cette saillance est directement liée au contraste ou aux singularités de mouvement. La saillance spatio-temporelle est ensuite obtenue en fusionnant la densité de saillance spatiale et la densité de saillance temporelle. L'algorithme de fusion a été détaillé dans le chapitre précédent.
Enfin, les performances du modèle spatio-temporel sont évaluées sur plusieurs séquences et à partir de la vérité terrain obtenue via les tests oculométriques. Pour cela, deux métriques sont utilisées : une fonction de probabilité cumulée et une matrice de confusion. La première est issue des travaux de D. Parkhurst [Parkhurst 04] alors que la seconde a déjà été présentée dans le chapitre précédemment. Grâce à ces différentes évaluations effectuées, nous apportons également des éléments de réponses concernant la contribution du mécanisme *Bottom-Up* dans le déploiement de l'attention visuelle. Le mécanisme *Bottom-Up* s'efface-t-il au profit du mécanisme *Top-Down* ou sa contribution reste-t-elle importante dans le temps ?

3.2 Construction d'une saillance temporelle à partir de l'espace psycho-visuel

3.2.1 Objectif

L'aspect temporel est primordial dans la modélisation de l'attention visuelle. Dans un contexte de recherche visuelle, J. Wolfe [Wolfe 89] a clairement identifié le mouvement comme un attracteur visuel. Une cible en mouvement enfouie dans un ensemble de distracteurs fixes attire l'attention. En outre, une cible fixe enfouie dans un ensemble

de distracteurs en mouvement attire l'attention mais dans une moindre mesure. Dans ce contexte d'études, le contraste en mouvement est l'élément déterminant qui attire notre attention visuelle. La cible en contraste de mouvement saute littéralement aux yeux.
De plus, pour la détection de zones saillantes d'une séquence d'images projetées sur un écran, il est intéressant d'avoir à l'esprit les règles en vigueur dans la façon de filmer [Zettl 90]. Les mouvements de caméra influencent clairement la stratégie visuelle de l'observateur. La présence ou de non de mouvement permet de hiérarchiser les différents évènements. Par ailleurs, la prise de vue est significative du message que le metteur en scène souhaite faire passer. Elle incite inconsciemment le téléspectateur à regarder quelque chose à un endroit particulier.
En conclusion, l'objectif est de déterminer les zones présentant un contraste de mouvement. Un estimateur de mouvement local, travaillant sur toutes les sous bandes de la composante achromatique, est nécessaire ainsi qu'un estimateur du mouvement dominant. A partir de ces deux procédés, il est possible de déterminer en chaque site le contraste de mouvement. Ce dernier à la base de la construction d'une saillance temporelle.

3.2.2 Synoptique de l'extension à la dimension temporelle

Le synoptique de l'extension du modèle à la dimension temporelle est présenté à la figure 3.1. Les éléments constitutifs de ce synoptique sont les suivants :

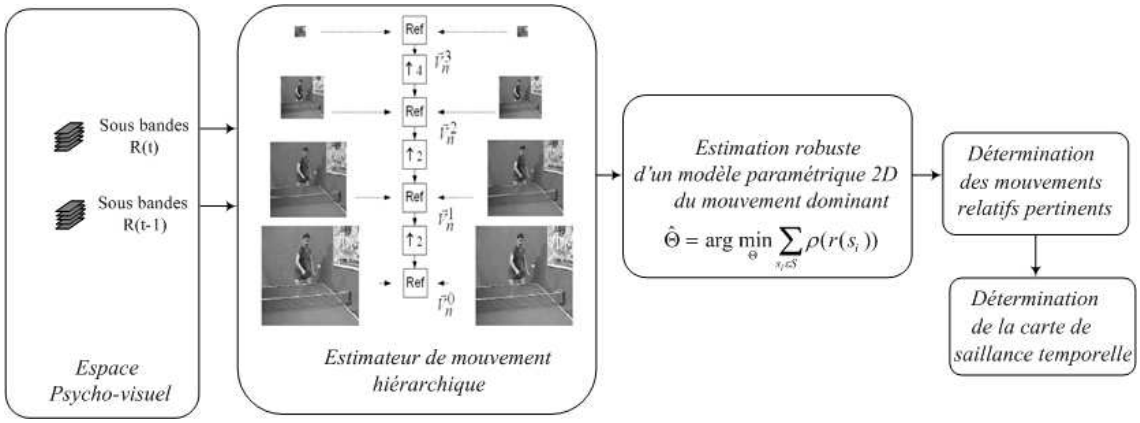

FIG. 3.1: Synoptique de l'extension du modèle à la dimension temporelle.

- un estimateur hiérarchique du mouvement local est utilisé pour déterminer le déplacement subi par chaque site s entre deux images ;
- une étape de détermination d'une représentation paramétrique du mouvement dominant via un modèle 2D paramétrique polynomial. Cela permet d'exprimer le déplacement en chaque site de l'image comme une fonction polynomiale de la position du point ;
- détermination du mouvement relatif en chaque site s de l'image ;
- détermination de la carte de saillance temporelle.

Ces différentes étapes sont décrites dans les paragraphes ci-après.

3.2.3 Estimation hiérarchique du mouvement local

A partir des sous bandes des couronnes I, II, III et IV issues de l'espace psycho-visuel, une pyramide multirésolution pour un couple d'images pour lequel on cherche à estimer le mouvement local est construite. Étant donné que la décomposition en sous bandes perceptuelle n'est pas une décomposition dyadique et que les facteurs de sous échantillonnage ne sont pas simples (matrice de sous échantillonnage non diagonale), le problème est simplifié en estimant des rapports de sous échantillonnages faciles à utiliser. Cette simplification a peu de conséquence sur le résultat de l'estimation de mouvement.
On rappelle que :

- le domaine I correspond aux fréquences spatiales radiales comprises entre 0 et 1.5 cpd (cycles par degré) ;
- le domaine II correspond aux fréquences spatiales radiales comprises entre 1.5 et 5.7 cpd ;
- le domaine III correspond aux fréquences spatiales radiales comprises entre 5.7 et 14.2 cpd ;
- le domaine IV correspond aux fréquences spatiales radiales comprises entre 14.2 et 28.2 cpd.

Les facteurs de sous échantillonnage que nous avons considérés sont les suivants :

- le passage de la pleine résolution aux basses fréquences se fait avec un facteur de sous échantillonnage de 16 (facteur inférieur aux rapports des fréquences spatiales des deux sous bandes considérées) ;
- le passage de la pleine résolution à la couronne II se fait avec un facteur de sous échantillonnage de 4 ;
- le passage de la pleine résolution à la couronne III se fait avec un facteur de sous échantillonnage de 2 ;
- la couronne IV est à la pleine résolution.

Les grandes lignes de l'algorithme adapté de la méthode très classique d'estimation de mouvement hiérarchique sont décrites ci-après. La figure 3.2 illustre le mécanisme d'estimation sur l'exemple de la séquence *Stefan*.

1. génération de la pyramide multirésolution pour chacune des images, notées I_t et I_{t-1}. Pour chacune des images, on obtient 4 résolutions I_t^l avec $l = 0, 1, 2, 3$. La résolution indicée 3 représente l'image de plus faible taille ;
2. estimation du mouvement à la résolution spatiale la plus grossière. L'estimation est réalisée par bloc (4×4) dans une fenêtre de recherche définie par l'excursion en x et en y. Par défaut, l'excursion en x (respectivement en y) est égale à 4 (respectivement 2). L'utilisation de prédicteurs spatiaux et hiérarchiques (pour les niveaux 0 et 1) permet d'accroître les performances. La figure 3.3 illustre ces différents prédicteurs :
   - les prédicteurs spatiaux sont le vecteur nul et les vecteurs causaux directement adjacents au bloc courant (soit 5 prédicteurs) ;
   - les prédicteurs hiérarchiques proviennent d'un niveau de résolution plus faible. Ils sont remis à l'échelle en utilisant le facteur approprié (facteur de sous échantillonnage permettant de passer de la résolution courante à la résolution inférieure).

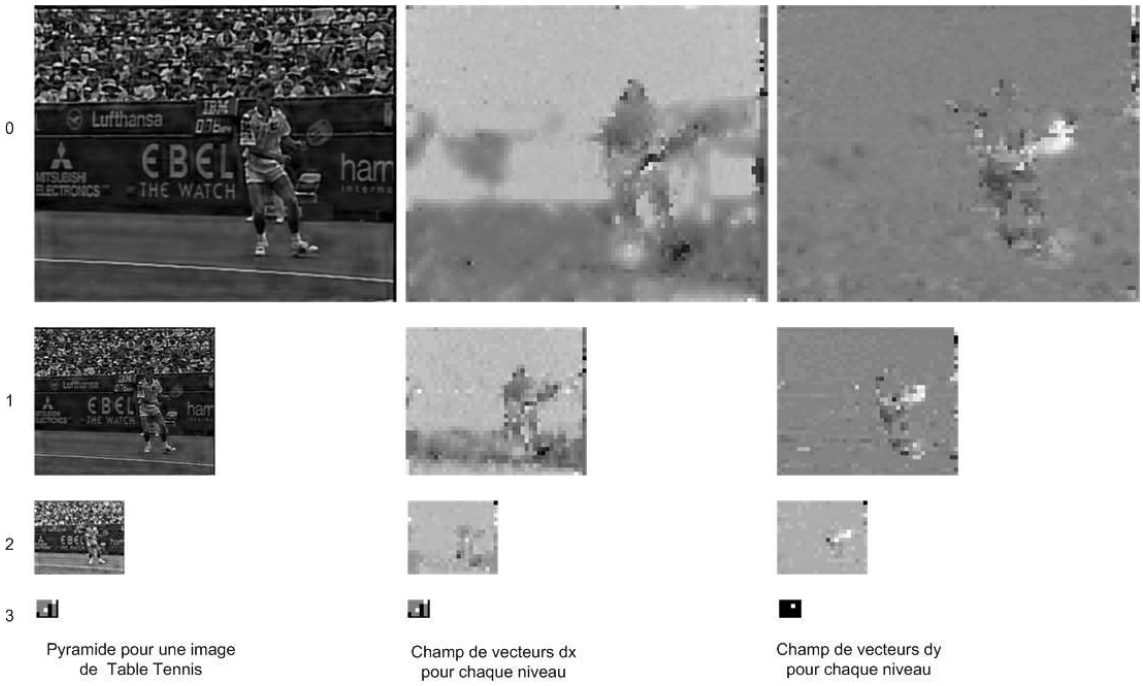

FIG. 3.2: Illustrations sur la séquence *Stefan* de l'estimation de mouvement. Exemples de champs de vecteurs obtenus pour chaque niveau de la pyramide.

Le critère de mise en correspondance entre blocs est la somme des différences absolues (SAD, *Sum of Absolutes Differences*). Nous aurions pu tout aussi bien prendre la somme des erreurs quadratiques comme critère à minimiser. A noter que le calcul de la position spatiale du pixel compensée s'effectue à partir d'une interpolation bilinéaire. Pour chaque prédicteur, la SAD est calculée. Un terme de régularisation dépendant des vecteurs précédemment trouvés est appliqué pour favoriser une solution homogène. Le prédicteur donnant la plus petite erreur est conservé pour initialiser une recherche exhaustive (*Full Search*). Alors que la détermination du meilleur prédicteur s'effectue au quart de pixel, la recherche exhaustive se fait avec une précision pixelique. A la fin de cette étape, l'ensemble des vecteurs $\overrightarrow{V} = (dx^l, dy^l)^T$ pour la résolution l est déterminé ;

3. le procédé d'estimation se poursuit en changeant de résolution, en prenant soin de remettre à l'échelle les prédicteurs. Tant que la résolution la plus élevée n'est pas atteinte, l'algorithme d'estimation réitère les étapes 2 et 3.

Les vecteurs de mouvement provenant de l'estimation sont par la suite appelés $\overrightarrow{V}_{local}$.

3.2.4 Détermination d'une représentation paramétrique du mouvement dominant via un modèle 2D paramétrique polynomial

Il existe dans la littérature différentes méthodes permettant d'analyser le mouvement apparent d'une séquence d'images. On propose ici de décrire une méthode identifiant une représentation paramétrique du mouvement dominant via un modèle 2D paramétrique

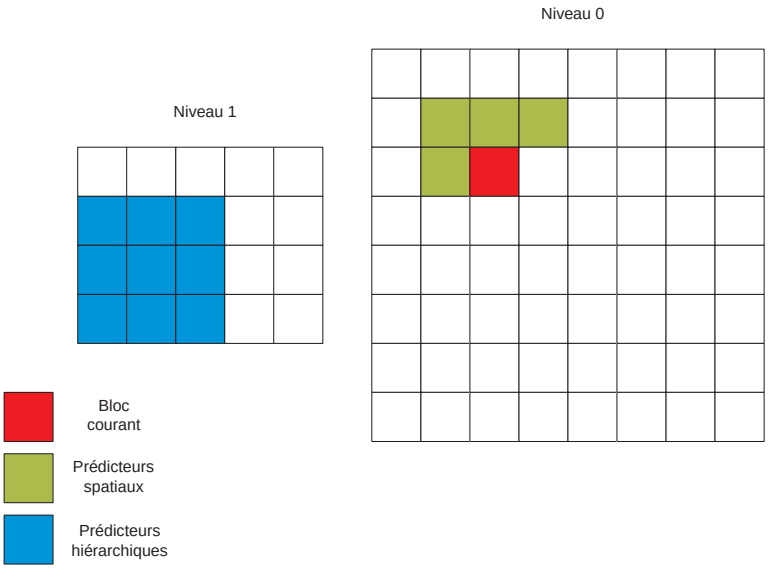

FIG. 3.3: Les prédicteurs spatiaux et hiérarchiques utilisés lors de l'estimation de mouvement.

polynomial.
Ce type de modèle concerne des modèles polynomiaux exprimant le déplacement d'un site de l'image en fonction de ses coordonnées. En utilisant des notations matricielles, le vecteur de vitesse $\overrightarrow{v}_\Theta(s)$ au site s de l'image S relativement au modèle paramétrique de paramètres Θ est donné par la relation suivante :

$$\overrightarrow{v}_\Theta(s) = \begin{pmatrix} u_\Theta(s) \\ v_\Theta(s) \end{pmatrix} = B(s)\Theta \tag{3.1}$$

où $s = (x, y)$ est un site de l'image et $B(s)$ une matrice dont la forme dépend de l'ordre du modèle. En règle générale, les modèles considérés se limitent aux polynômes de degrés 1 et 2 (modèles affines et quadratiques, respectivement). Pour un modèle affine complet à six paramètres, la relation (3.1) devient :

$$\overrightarrow{v}_\Theta(s) = \begin{pmatrix} a_1 + a_2 x + a_3 y \\ a_4 + a_5 x + a_6 y \end{pmatrix} \tag{3.2}$$

où $\Theta = [a_1, a_2, a_3, a_4, a_5, a_6]$ contient les six paramètres du modèle de mouvement affine.

Cette représentation paramétrique, utilisant un modèle affine complet, permet d'appréhender de nombreuses situations comme les translations ($a_3 = a_4 = a_5 = a_6 = 0$), les zooms ($a_1 = a_3 = a_4 = a_5 = 0$), les rotations ($a_1 = a_2 = a_4 = a_6 = 0$)... Par ailleurs, l'estimation des paramètres se fait sur un support étendu (toute l'image) qui permet une estimation fiable du mouvement dominant global.
Pour minimiser la complexité, l'estimation du mouvement apparent dominant est réalisée sur un champ de vecteurs préalablement estimé. L'algorithme d'estimation retenu est issu de travaux de J.M. Odobez [Odobez 95] sur les estimateurs robustes. Avant de présenter

les étapes de cet algorithme, le principe des estimateurs robustes est rappelé.
Le défaut majeur de l'estimation de paramètres via les moindres carrés classiques est la grande sensibilité de la méthode aux données non conformes au modèle, communément appelées *outliers*. Cette sensibilité rend cette estimation instable. On rappelle qu'un estimateur aux moindres carrés minimise la somme des erreurs résiduelles (notée $r(s_i)$) au carré :

$$\widehat{\Theta} = arg \min_{\Theta} \sum_{s_i \in S} r_{\Theta}(s_i)^2 \tag{3.3}$$

où $r_{\Theta}(s_i) = I(s_i + \overrightarrow{v}_{\Theta}(s_i), (t+1) - I(s_i, t)$ représente la différence de l'image déplacée, souvent appelée DFD ou erreur résiduelle.
L'estimation robuste, utilisant un M-estimateur, pondère l'importance allouée à chaque donnée en fonction de sa conformité au modèle. La réduction des effets des données aberrantes sur la détermination du modèle paramétrique s'effectue en remplaçant dans la relation (3.3) l'erreur résiduelle au carré par une fonction de l'erreur résiduelle qui est symétrique, positive et ayant un minimum unique en 0. L'erreur résiduelle est ainsi pondérée en fonction de son importance via cette fonction notée $\rho()$. Les fonctions les plus utilisées sont :

- la fonction de Geman - Mc Clure que nous avons choisie : $\rho(x) = \frac{x^2/2}{1+x^2}$
- la fonction de Welsh : $\rho(x) = \frac{c^2}{2}\left[1 - exp(-(\frac{x}{c})^2)\right]$
- ou encore les fonctions de Tukey, Huber, Cauchy...

L'estimateur $\widehat{\Theta}$ doit alors minimiser la fonction suivante :

$$\widehat{\Theta} = arg \min_{\Theta} \sum_{s_i \in S} \rho(r_{\Theta}(s_i)) \tag{3.4}$$

Pour des raisons de lisibilité, $r_{\Theta}(s_i)$ sera noté r_i.
Une condition nécessaire pour minimiser (3.4) requiert que la dérivée de l'erreur résiduelle par rapport à chaque composante Θ_i du vecteur de paramètres Θ soit nulle :

$$\widehat{\Theta} = arg \min_{\Theta} \sum_{s_i \in S} \rho(r_i) \Leftrightarrow \sum_{s_i \in S} \varphi(r_i) \frac{\partial(r_i)}{\partial(\Theta_j)} = 0, j = 1, ..., m \tag{3.5}$$

où,
$\varphi(x) = \frac{d\rho(x)}{dx}$ est appelé la fonction d'influence et m le nombre de paramètres considérés.
En introduisant le terme de pondération $w(r_i) = \frac{\varphi(r_i)}{r_i}$, la formule (3.5) devient :

$$\sum_{s_i \in S} \varphi(r_i) \frac{\partial(r_i)}{\partial(\Theta_j))} = \sum_{s_i \in S} w(r_i) r_i \frac{\partial(r_i)}{\partial(\Theta_j)} = 0, j = 1, ..., m \tag{3.6}$$

Cela revient donc à minimiser l'équation (3.7) via un moindre carré itératif pondéré :

$$min \sum_{s_i \in S} w(r_i^{(k-1)}) r_i^2 \tag{3.7}$$

où,
k est l'indice de l'itération. Les pondérations $w(r_i^{(k-1)})$ doivent être recalculées après

chaque itération.
Le nombre d'itérations nécessaire pour converger vers une solution acceptable est de l'ordre de 5 à 7. Généralement, les paramètres estimés par un moindre carré classique servent d'initialisation à la première d'itération.

3.2.5 Mouvement relatif et saillance temporelle

A partir de la connaissance du mouvement apparent dominant $\overrightarrow{V}_{\Theta}$ et du déplacement local $\overrightarrow{V}_{local}$ pour chaque site s, le mouvement relatif $\overrightarrow{V}_{relatif}$, exprimé dans le référentiel rétinien est obtenu simplement par la relation suivante :

$$\overrightarrow{V}_{relatif}(s) = \overrightarrow{V}_{\Theta}(s) - \overrightarrow{V}_{local}(s) \tag{3.8}$$

Le mouvement relatif est nécessaire pour estimer le contraste de mouvement inhérent à un site particulier. Mais, ce n'est pas suffisant de le considérer de cette façon. En effet, l'oeil est capable de poursuivre des objets en déplacement. Cette faculté, liée au mouvement oculaire de poursuite, permet de conserver l'objet suivi dans la fovéa, partie de la rétine présentant la sensibilité spatiale la plus élevée. Par conséquent, considérer directement le mouvement relatif donné par la relation (3.8) serait réducteur. Il n'est pas correct de dire que plus le mouvement relatif est important, plus la saillance est forte. Il faut prendre en compte la capacité maximale de poursuite de l'oeil. S. Daly a montré que la vitesse de poursuite maximale de l'oeil pouvait aller jusqu'à 80 deg/sec. Par ailleurs, sachant que le contexte de l'étude s'effectue dans un cadre multi-observateurs, toutes les zones en déplacement sont susceptibles d'être suivies. La relation (3.8) est donc modifiée. Plus la vélocité du mouvement relatif est supérieure à la vélocité maximale de poursuite et plus la saillance est atténuée :

$$\overrightarrow{V}_{relMod}(s) = \overrightarrow{V}_{relatif}(s) \cdot \left\{ \frac{\overrightarrow{v}_{max}}{\|\overrightarrow{V}_{relatif}(s)\|} \right\}^{\gamma} \text{ si } \|\overrightarrow{V}_{relatif}(s)\| > \overrightarrow{v}_{max} \tag{3.9}$$

avec, $\overrightarrow{v}_{max}$ la vélocité maximale de poursuite de l'oeil. Le paramètre γ contrôle la modification de la pente. En pratique, nous avons γ égal à 3.
Une autre propriété est à prendre en compte. Il est relativement connu qu'un objet en mouvement sur un fond fixe attire plus facilement l'attention qu'un objet fixe sur un fond en mouvement [Wolfe 89]. L'amplitude du mouvement dominant est donc un facteur important pour déterminer la saillance finale. Une façon pertinente pour l'évaluer consiste à prendre la valeur médiane de l'histogramme des mouvements relatifs quantifiées, noté par $Med(\|\overrightarrow{V}_{relMod}(s)\|_Q)$. La saillance temporelle S^T est alors déduite en pondérant le mouvement relatif quantifié par $Med(\|\overrightarrow{V}_{relMod}(s)\|_Q)$:

$$S^T(s) = \frac{\|\overrightarrow{V}_{relMod}(s)\|_Q}{Med(\|\overrightarrow{V}_{relMod}(s)\|_Q)} \tag{3.10}$$

3.3 Détermination d'une densité de saillance spatio-temporelle

A partir de la densité spatiale de saillance DS^{SP} et de la densité temporelle de saillance DS^T, la densité spatio-temporelle est à déterminer.

A partir de l'opérateur de fusion $\mathcal{F}$ défini préalablement dans le paragraphe 2.4.3.2, une carte de densité de saillance finale, notée DS est obtenue via la relation suivante :

$$DS(s) = \mathcal{F}(DS^{SP}, DS^{T}) \tag{3.11}$$

3.4 Performance de la modélisation sur séquences d'images

3.4.1 Évaluation qualitative

Les figures 3.4 et 3.5 présentent les points de fixations réels et les points de fixations prédits pour respectivement les séquences *Kayak* et *Stefan*. Ces images ont été déterminées de la façon suivante : les 20 premiers points de fixation (un point de fixation est représenté par un cercle de diamètre de 1 degré visuel centré sur un maximum local) sont déterminés à partir de la densité de saillance associée à l'image courante. Chaque point de fixation est remplacé par le contenu de l'image source. Concernant les résultats de la figure 3.4, il y a indéniablement un fort degré de similarité entre les données réelles et les données prédites. La zone d'intérêt principale est correctement détectée par le modèle proposé. Concernant les résultats de la figure 3.5 relatif à la séquence *Stefan*, les résultats sont moins bons que les précédents mais restent tout à fait corrects. Le taux de couverture obtenu sur cette séquence semble être plus important que le précédent. Cette séquence est en effet plus riche en informations visuelles qu'elles soient de bas ou de haut niveau (texte en arrière plan). Cette richesse d'information explique en partie la relative dispersion des points de fixation des observateurs.
Comme précédemment, il est difficile de conclure définitivement sur la qualité de la modélisation proposée. Elle semble toutefois permettre de détecter relativement fiablement les zones les plus saillantes d'une séquence d'images.

3.4.2 Évaluation quantitative

3.4.2.1 Fonction de probabilité cumulée

D. Parkhurst et E. Niebur [Parkhurst 04] ont défini une méthode, appelée fonction de probabilité cumulée, permettant d'évaluer la pertinence d'un modèle d'attention visuelle. Cette mesure consiste, tout d'abord, à normaliser les distributions des valeurs de saillance prédite afin d'obtenir des densités de probabilité. Ensuite, les coordonnées des N premiers points de fixations sont déterminées à partir des cartes de saillance provenant des expérimentations oculométriques. Un point de fixation est une zone circulaire d'un degré visuel de rayon centrée sur un maximum local. La détermination du point de fixation n s'effectue en recherchant le maximum de la carte de saillance, dans laquelle la zone circulaire centrée sur les coordonnées du point de fixation $n-1$ a été inhibée. A partir de ces N points de fixations, les auteurs calculent la probabilité cumulée définies par ces N zones :

$$C_P = \sum_{k=1}^{N} P(x_k, y_k) \tag{3.12}$$

avec, $P()$, la densité de probabilité issue de la carte de saillance prédite. (x_k, y_k) représentent les coordonnées du point de fixation k.

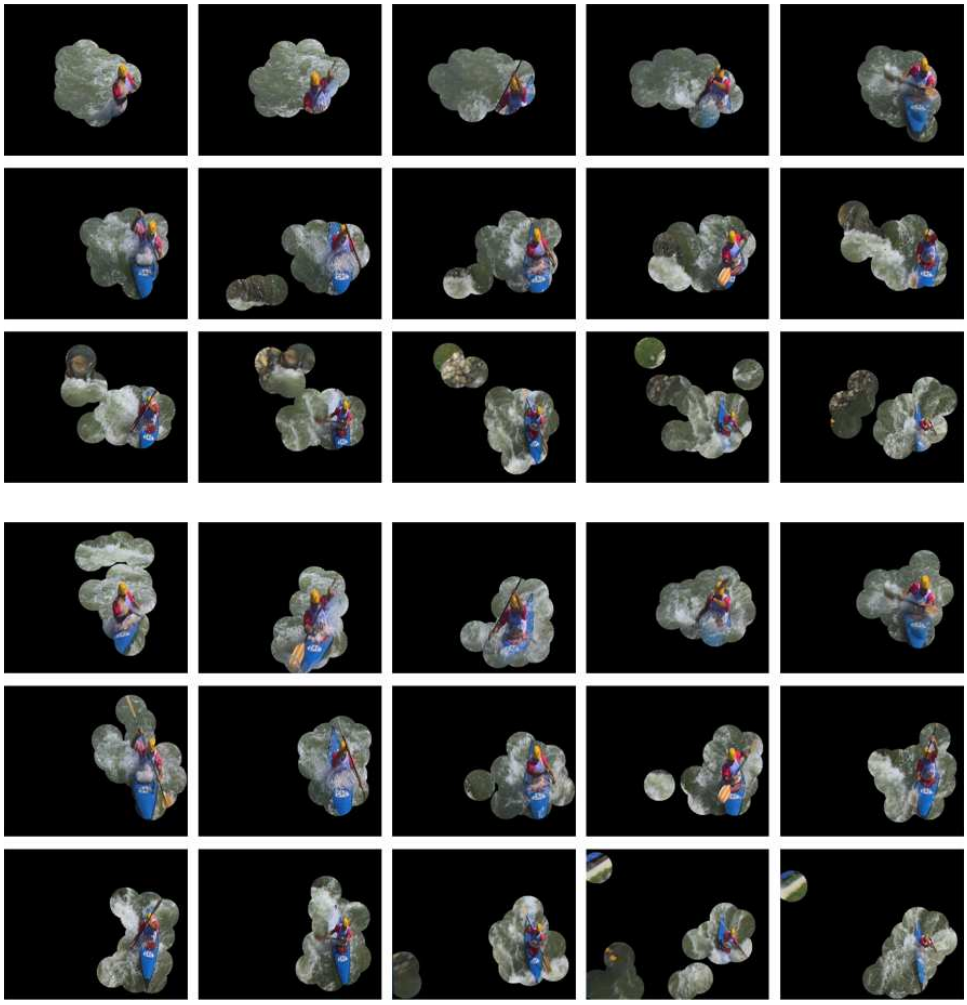

FIG. 3.4: Évaluation qualitative entre les données réelles (15 premières images) et les données prédites (15 dernières images) de la séquence *Kayak*. La première image du premier groupe d'images est à comparer avec la première image du deuxième groupe. La deuxième image (à droite de la première) du premier groupe est à comparer avec la deuxième du deuxième groupe. Et, ainsi de suite...

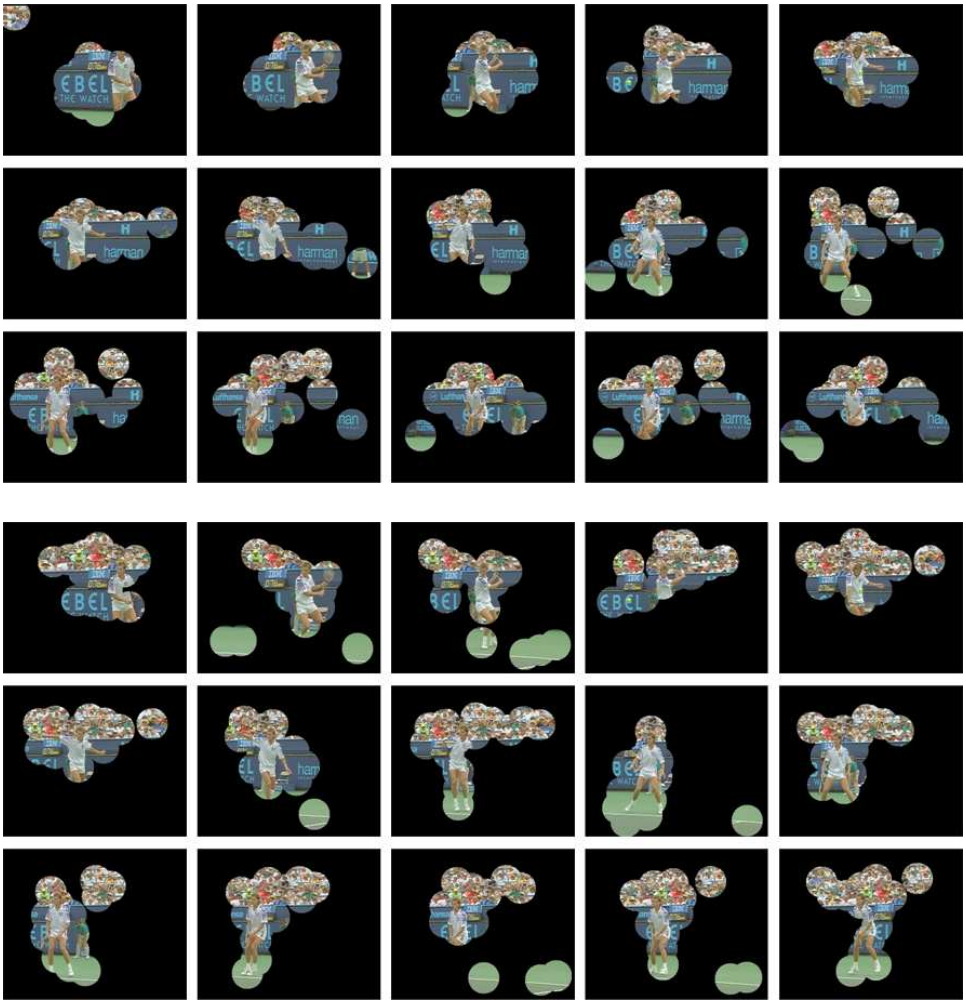

FIG. 3.5: Évaluation qualitative entre les données réelles (15 premières images) et les données prédites (15 dernières images) de la séquence *Stefan*. La première image du premier groupe d'images est à comparer avec la première image du deuxième groupe. La deuxième image (à droite de la première) du premier groupe est à comparer avec la deuxième du deuxième groupe. Et, ainsi de suite...

Ce procédé est étendu simplement à la dimension temporelle par la relation (3.13) :

$$C_P^i = \sum_{k=1}^{N} P^i(x_k, y_k) \tag{3.13}$$

L'indice i représente l'index de l'image dans la séquence. Pour une séquence d'images comprenant M images, la valeur finale est ensuite obtenue par :

$$\overline{C_P} = \frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{N} P^i(x_k, y_k) \tag{3.14}$$

Probabilité cumulée moyenne, calculée sur un ensemble d'images

La figure 3.6 donne la probabilité cumulée moyenne calculée sur 1040 images de différentes séquences lorsque les 20 premiers points de fixation sont considérés. Les résultats $\overline{C_P}$ sont donnés pour différents types de modélisation :

- le modèle spatial, le modèle temporel et le modèle spatio-temporel ont été décrits dans les sections précédentes ;
- le modèle aléatoire correspond au tirage aléatoire des coordonnées (x_k, y_k) du point de fixation k ;
- le modèle uniforme fournit une distribution de saillance uniforme.

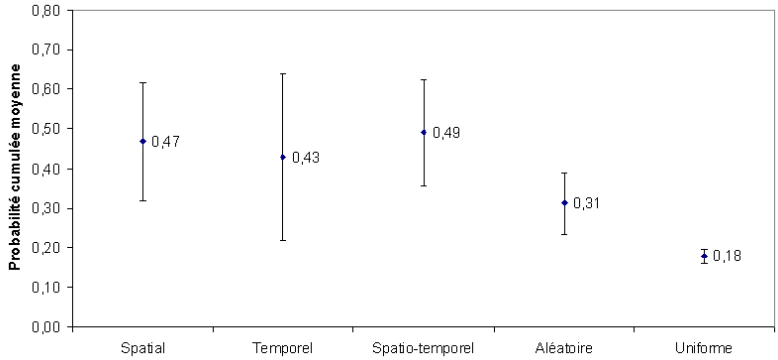

FIG. 3.6: Probabilité cumulée moyenne $\overline{C_P}$ calculée sur 1040 images de différentes séquences. Les valeurs moyennes sont données avec plus ou moins l'écart type de l'erreur.

Le meilleur prédicteur des points de fixation réels est le modèle spatio-temporel, c'est à dire le modèle intégrant la dimension spatiale (luminance et couleur) et la dimension temporelle. Ce premier résultat est logique et tout à fait cohérent avec les résultats de D. Parkhurst. Le modèle spatial et le modèle temporel présentent des résultats justes inférieurs. Il est intéressant de noter la forte dispersion autour de la valeur moyenne obtenue par le modèle temporel. Ce résultat est également logique puisque les performances de ce modèle dépendent de la présence ou non de mouvement dans les séquences évaluées. Les deux derniers modèles testés sont le modèle dit aléatoire et uniforme. Le résultat donné par le modèle aléatoire, significativement inférieur aux résultats issus des modèles biologiquement plausibles, montre clairement que l'attention visuelle est attirée par des caractéristiques visuelles de bas niveaux. Enfin, le plus mauvais résultat est donné, sans

TAB. 3.1: Probabilité cumulée moyenne calculée sur les 90 premières images des séquences *Kayak*, *Table* et *Stefan*. Les résultats sont donnés pour les 10 et 20 premiers points de fixation et pour différents modèles. La colonne intitulée saillance à prédire donne la probabilité cumulée $\overline{C_H}$ calculée à partir de la densité de probabilité réelle ($\overline{C_H} = \frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{N} H^i(x_k, y_k)$).

		Spatial	Temporel	Spatio-temporel	Uniforme	Saillance à prédire
Kayak	N=20	0.58	0.46	0.56	0.16	0.94
	N=10	0.4	0.31	0.39	0.09	0.73
Table	N=20	0.40	0.64	0.55	0.17	0.91
	N=10	0.25	0.45	0.39	0.11	0.68
Stefan	N=20	0.44	0.41	0.51	0.18	0.89
	N=10	0.29	0.31	0.35	0.09	0.68

surprise, par le modèle uniforme.

Le tableau 3.1 donne la probabilité cumulée moyenne calculée sur les 90 premières images de trois séquences pour différentes modélisations. Remarquons, tout de suite, que les 20 premiers points de fixation permettent de considérer 90% de la saillance réelle (dernière colonne du tableau), que nous notons $\overline{C_H}$. Cela s'explique par le fait que la densité de probabilité réelle est relativement concentrée. Si on considère la séquence *Kayak*, le modèle spatial est le meilleur prédicteur. Sur cette séquence se caratérisant par une région d'intérêt majeure contrastant fortement avec le fond, la dimension spatiale est suffisante. Sur la séquence *Table*, c'est la dimension temporelle qui fournit les meilleurs résultats. Le mouvement est dans ce cas un facteur discriminant. Enfin, pour la séquence *Stefan*, la dimension spatiale et la dimension temporelle présentent des résultats comparables. L'algorithme de fusion tire ensuite profit de la redondance et de la complémentarité des deux résultats de modélisation. Néanmoins, si on considère le rapport $\frac{\overline{C_P}}{\overline{C_H}}$, c'est à dire la quantité de saillance que nous avons réussi à prédire, le résultat est décevant. En effet, ce rapport se situe autour de 0.6. Il existe donc une différence fondamentale entre les distributions de probabilité réelles et prédites. Alors que la densité de probabilité de saillance réelle est à bords francs, la densité de probabilité prédite est plus lisse. Il est donc difficile de conclure formellement. Une autre méthode d'évaluation, décrite dans la section suivante, est nécessaire pour éclaircir cette situation.

Évolution temporelle de la probabilité cumulée

La figure 3.7 présente l'évolution temporelle de la quantité C_P^i (i indiquant l'indice de l'image) sur la séquence *Stefan*. Il est intéressant de noter que le modèle spatio-temporel réussit à tirer profit de la dimension spatiale et temporelle. Par exemple, sur des plans relativement fixes (images de 1 à 15, images 48 à 74), la dimension temporelle n'est pas un bon prédicteur des fixations des observateurs. L'opérateur de fusion, tel que nous l'avons défini précédemment, permet de pallier ce problème en favorisant clairement la dimension spatiale pour construire le résultat final. Par contre, lorsque la dimension temporelle est pertinente pour la construction de la prédiction finale (images de 16 à 32, images de 41 à 49), l'opérateur de fusion l'utilise clairement.

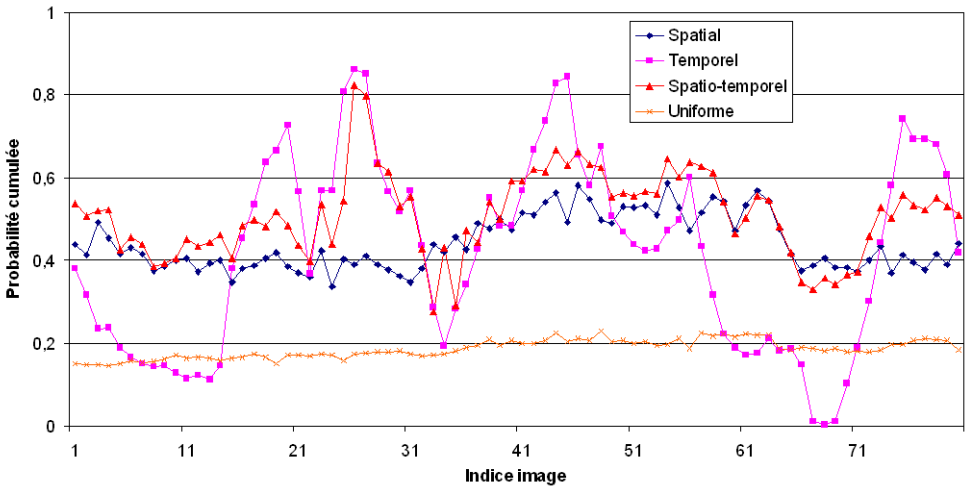

FIG. 3.7: Évolution temporelle de la probabilité cumulée C_P^i sur la séquence *Stefan* en fonction du modèle (spatial, temporelle, spatio-temporelle, uniforme).

La figure 3.8 présente l'évolution temporelle de la quantité moyenne $\overline{C_P^i}$ calculée à partir de données relatives à 14 séquences. Ces résultats confirment les données de la figure 3.6. Le modèle spatio-temporel est le meilleur prédicteur des points de fixations réels.

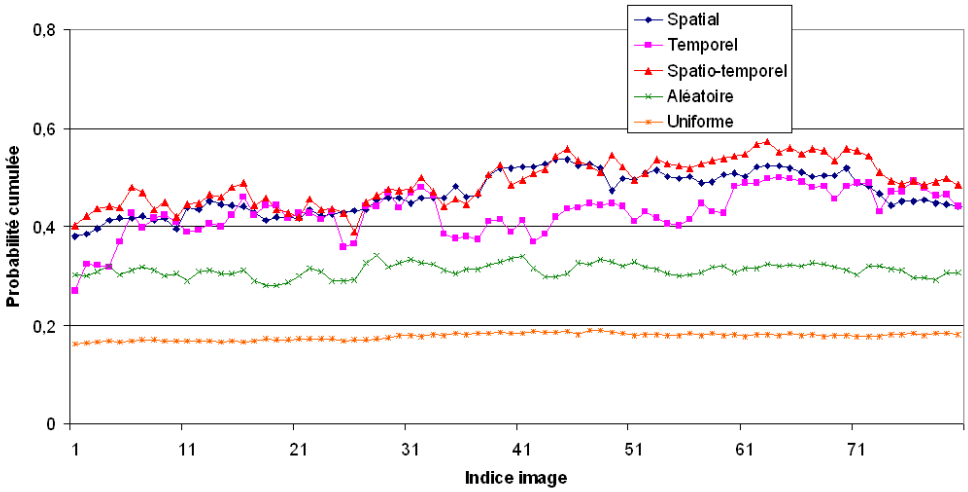

FIG. 3.8: Évolution temporelle de la probabilité cumulée calculée sur les 80 premières images de 14 séquences.

3.4.2.2 Matrice de confusion

Comme pour l'évaluation du modèle de saillance spatiale, une matrice de confusion est utilisée pour évaluer le taux de précision du modèle. Cette méthode permet d'être insensible à la dynamique de la saillance. Le tableau 3.2 présente pour quatre séquences la précision moyenne de la classification. La précision est particulière bonne pour les

TAB. 3.2: Précision moyenne de la modélisation pour quatre séquences.

	Kayak	Table	Stefan	Titleist	Moyenne
$\mathcal{P}$	0,8	0,78	0,75	0,75	0,77
σ	0,05	0,05	0,04	0,09	0,05

séquences ayant une région d'intérêt contrastant avec le fond (séquences *Kayak* et *Stefan*). Pour les autres, la précision, un peu plus faible, reste satisfaisante. Ces derniers résultats confirment les précédents.

3.4.3 Quelle est l'influence du mécanisme *Bottom-Up* pour l'attention visuelle ?

L'influence du mécanisme *Bottom-Up* est évidente dans le domaine de la recherche visuelle (*Visual Search*). Comme présenté précédemment, des expériences de recherche visuelle mesurent le temps de réaction nécessaire pour trouver un objet cible enfoui parmi d'autres objets agissant comme des distracteurs. Dans un cas disjonctif, le temps de réaction est constant et cela quel que soit le nombre de distracteurs. Les informations de bas niveau, extraites par un mécanisme *Bottom-Up*, attirent clairement et indiscutablement l'attention visuelle.
Concernant les scènes naturelles, il y a peu de résultats démontrant l'importance de ce mécanisme. D. Parkhurst [Parkhurst 02], dans sa thèse, a montré que le mécanisme *Bottom-Up* ne se dissipait pas au profit du mécanisme *Top-Down* après les premières fixations. Son travail portait uniquement sur la dimension spatiale et il se contentait des toutes premières fixations (15 premières fixations) pour réaliser ses différentes expérimentations. Bien que ses travaux soient critiquables sur plusieurs points (dépendance des premières fixations avec le centre de l'image), il est le premier à avoir réalisé ce genre d'études. Pour qualifier l'importance de la vision précoce, il a utilisé le modèle de L. Itti.
Lors de l'évaluation de la pertinence de notre modèle, nous contribuons indirectement à répondre cette question. Tout d'abord, si on considère le modèle spatial appliqué sur les séquences d'images, les résultats obtenus montrent que le mécanisme *Bottom-Up* est omniprésent durant toute la durée de l'expérience. Il suffit pour s'en convaincre de faire l'hypothèse suivante : si le mécanisme *Bottom-Up* est un mécanisme d'une durée brève, annihilé par le mécanisme *Top-Down*, les performances de la modélisation devraient s'effriter avec le temps. Au vu de nos résultats, ce n'est le cas ni pour les résultats de l'évaluation du modèle spatial ni pour ceux du modèle spatio-temporel. Pour le modèle de saillance spatiale, le choix de la durée d'observation, c'est à dire une durée d'observation importante pour effectuer l'évaluation des performances, est donc cohérente.
Finalement, au vu de ces résultats, la conclusion serait de dire que le mécanisme *Bottom-Up* ne disparaît pas au profit du mécanisme *Top-Down*. Citons, pour finir, une expérience intéressante : cette dernière consistait à mesurer les positions des points de fixations d'un panel d'observateurs regardant une image. Bien qu'une tâche particulière était précisée (regarder le détail en haut à gauche), un élément contrastant avec le fond a continuellement attiré l'attention des observateurs. Cette expérience met en exergue la forte influence et le caractère non transitoire du mécanisme *Bottom-Up*.

3.5 Conclusion

La vocation de ce chapitre était de proposer une extension de la modélisation à la dimension temporelle. Comparativement à la détermination de la saillance spatiale, la saillance temporelle est plus facile à calculer car le concept sous-jacent est relativement simple : une zone temporellement saillante est une zone en contraste de mouvement. Dans l'approche proposée, le contraste de mouvement est déterminé à partir d'une estimation locale et d'une estimation globale du mouvement ; la différence, appelée mouvement relatif, indique les zones temporellement saillantes.
La densité de saillance finale est ensuite obtenue grâce à un procédé de fusion des densités de saillance spatiale et temporelle. Ce procédé décrit dans le chapitre précédent est basé sur la redondance et la complémentarité des données.
La seconde partie de ce chapitre était consacrée à l'évaluation des performances du modèle proposé vis à vis d'une vérité terrain acquise lors d'expérimentations oculométriques. Des comparaisons qualitatives indiquant une certaine ressemblance ont été confirmées une nouvelle fois par des méthodes objectives de comparaison. Deux méthodes ont été utilisées : une fonction de probabilité cumulée et une matrice de confusion. Si un chiffre devait être retenu, ce serait la précision moyenne du modèle obtenue à partir des matrices de confusion : en moyenne, 77% des pixels d'une image sont correctement classés (82%) lorsque la fonction gaussienne est appliquée).
Ces différentes évaluations ont permis également d'apporter des compléments d'information concernant l'activation du mécanisme *Bottom-Up*. Les résultats préliminaires de D. Parkhurst sont confirmés, étendus à des durées d'observation importantes et à la dimension temporelle. Le mécanisme *Bottom-Up* n'est pas un mécanisme transitoire qui voit son influence complètement annihilée par le mécanisme *Top-Down*. Dans un contexte de *free-viewing*, il peut certes se dissiper momentanément mais son action reste en moyenne très importante. Une étude complémentaire serait maintenant de refaire l'évaluation du modèle lorsqu'un ordre précis est donné aux observateurs.

Troisième partie

Applications

Chapitre 1

Codage vidéo à qualité différenciée basé sur la saillance visuelle

1.1 Introduction

L'objet de ce chapitre concerne une application de compression de la vidéo avec une qualité visuelle différenciée pilotée par les cartes de saillance. Ce type de compression est communément appelée compression sélective ou compression avec régions d'intérêt. Contrairement aux approches conventionnelles de compression d'images distribuant de façon homogène les ressources de codage, la compression sélective répartit les ressources de codage de façon adaptée directement ou indirectement. Dans un contexte de compression avec pertes, la distribution adaptée des ressources de codage peut permettre d'accroître substantiellement la qualité globale perçue. L'idée est simple puisqu'elle consiste à favoriser la qualité des zones les plus importantes visuellement. Bien évidemment cela nécessite de disposer d'informations a priori sur la scène à coder. Dans ce contexte d'études, nous couplons un schéma de compression sélective avec le modèle d'attention visuel décrit précédemment.
Le principe général de la compression sélective ainsi qu'un bref état de l'art sont d'abord détaillés. Deux types de compression sélective indirecte et directe basées sur le standard H.264 sont ensuite abordés. Afin de valider l'utilisation d'une densité de saillance dans un schéma de compression vidéo, il est nécessaire d'examiner l'invariance de la stratégie visuelle ainsi que l'invariance de notre modèle à des artefacts de compression. Comme une densité de saillance sert uniquement à hiérarchiser les zones d'une image en terme de capacité d'attraction de l'attention visuelle, certaines modifications doivent être envisagées pour dédier la carte de saillance à une application de codage. Enfin, une application de compression sélective indirecte puis une application de compression sélective directe sont détaillées. La compression sélective indirecte concerne des méthodes de simplification du signal à coder. Pour la compression sélective directe, il s'agit de modifier le coeur et la stratégie de codage.

1.2 Principe général de la compression sélective

Le principe d'un codage sélectif ainsi que ses grandes étapes ont été définies par E. Nguyen dans sa thèse [Nguyen 95]. Il introduit deux notions, un a priori de sélection et un a priori de compression, décrites ci-dessous :

- *un a priori de sélection* permet de définir les zones de l'image à privilégier. Ces zones peuvent être définies implicitement via des connaissances a priori sur le contenu des images (par exemple dans le cadre de la visiophonie ou les images sont de type tête et épaules). Cet a priori peut être défini de façon explicite c'est à dire que sa détermination est basée sur le signal image en faisant intervenir des mesures d'importance locales, des opérateurs d'analyse (segmentation, classification, reconnaissance de formes, ...), des critères psycho-visuels...
- *un a priori de compression* qui caractérise d'une part la nature du codage et d'autre part le critère d'allocation des ressources de codage. Autrement dit, cet a priori concerne la traduction de l'a priori de sélection en terme de compression sélective.

On distingue également deux façons, qui peuvent être complémentaires, d'effectuer une compression sélective : la première dite indirecte consiste à modifier le signal à coder de façon à réduire la quantité d'information sur les zones de non intérêt. La deuxième façon, dite directe, modifie directement le coeur de codage en fonction de la connaissance des régions d'intérêt. Une façon de mettre en oeuvre une compression sélective directe et indirecte est détaillée dans les sections suivantes.

1.3 Approches de compression sélective les plus marquantes

Ce bref état de l'art liste les grandes axes de recherche concernant la compression sélective hybride basée bloc appliquée à la vidéo. La plupart des algorithmes de compression sélective jouent sur l'optimisation de la quantification. Les applications basées sur la norme JPEG2000 ne sont pas décrites.

Optimisation de la quantification via une approche débit-distorsion

L'idée consiste à augmenter le nombre de bits alloué à la région d'intérêt en modifiant dynamiquement la consigne de quantification [Yang 00, Leontaris 04]. Pour ce faire, une optimisation Lagrangienne débit-distortion est effectuée de la façon suivante [Shoham 88, Ramchandran 94, Ortega 94].

Supposons que K macroblocs doivent être encodés. On suppose également connaître les valeurs de distorsion $D_i^{q_i}$ et le coût de codage $C_i^{q_i}$ de chaque macrobloc i et cela quel que soit l'index de quantification q_i. L'objectif d'une optimisation débit-distorsion est de trouver, parmi toutes les configurations possibles $Q = (q_1, \ldots, q_K)$, la solution optimale Q^* devant minimiser la distorsion finale sous une contrainte spécifique :

$$Q^* = arg \min_Q \sum_{i=1}^{K} D_i^{q_i} \text{ sachant que } \sum_{i=1}^{K} C_i^{(q_i)^*} \leq R \tag{1.1}$$

R étant le débit total.

Cette contrainte peut être directement pris en compte en utilisation la méthode des mul-

tiplicateurs de Lagrange. L'idée de base est de transformer le problème contraint exprimé par la relation (1.1) par un problème non contraint en intégrant un facteur Lagrangien λ :

$$J = \min_{q_i} \sum_{i=1}^{K} (D_i^{q_i} + \lambda \cdot C_i^{q_i}) \tag{1.2}$$

ou J représente le coût Lagrangien.
Si on fait l'hypothèse que la fonction débit-distortion est monotone, ce second problème (1.2) est équivalent à :

$$J = \sum_{i=1}^{K} \min_{q_i} (D_i^{q_i} + \lambda \cdot C_i^{q_i}) \tag{1.3}$$

Finalement, l'optimisation globale peut donc être menée de façon indépendante pour chaque macrobloc. Le paramètre λ peut être déterminé via différentes approches (itératives via une approche dichotomique [Shoham 88], programmation dynamique [Ortega 94]...).
Lorsque la position spatiale des zones d'intérêt est connue, il est possible de prendre en compte cette information pour favoriser les zones d'intérêt au détriment des zones in-intéressantes. La relation (1.3) est alors modifiée de la façon suivante :

$$J = \sum_{i=1}^{K} \min_{q_i} (D_i^{q_i} + \theta_i \cdot \lambda C_i^{q_i}) \tag{1.4}$$

La valeur θ_i représente le degré d'intérêt visuel du macrobloc i. Un macrobloc présentera un intérêt maximal lorsque θ sera nul.

Remarquons que les travaux de K. Ramchandran [Ramchandran 94] sont tout à fait intéressant dans le sens où l'aspect temporel est pris en compte dans l'optimisation débit-distorsion. En fait, ces travaux se basent sur le fait que la qualité de codage d'un macrobloc prédit à un instant t dépend de la qualité de codage des images servant de référence. La détermination des index de quantification ne peut être optimale que si on effectue une allocation de débit prenant conjointement la dimension spatiale et temporelle [Liu 05].

Optimisation de la quantification basée sur des attributs bas niveau

Une autre solution, pour modifier la consigne de quantification, consiste à effectuer une pré-analyse de la séquence à coder [Fabri 05], [Tang 04]. Par exemple, C. Tang et al. [Tang 04] définissent un index VDSI de sensibilité à la distorsion visuelle (VDSI : *Visual Distortion Sensitivity Index*) afin de déterminer un index de quantification adapté. La valeur de VDSI, définie pour chaque macrobloc, est obtenue à partir d'un indice spatial et d'un indice temporel : le premier est obtenu à partir de deux détecteurs de contours, utilisés pour caractériser les macroblocs comme soit texturés soit uniformes. Les zones fortement texturées sont intéressantes d'un point de vue codage puisqu'elles agissent comme un signal masquant le bruit de quantification. L'indice temporel tente, très simplement, de détecter les régions en contraste de mouvement. Nous avons précédemment vu que ces régions attirent l'attention visuelle.

Optimisation de la quantification basée sur une détection d'informations de haut niveau

L'application de compression sélective la plus connue est sans aucun doute celle qui se base sur la détection de visage. Dans un contexte de visiophonie, de vidéoconférence, l'intérêt de préserver la qualité d'images sur le visage est évident. Il existe de nombreux papiers abordant ce sujet. Juste à titre d'exemple, citons les nombreuses publications de Chai et al. [Chai 00] dans ce domaine.

Par modification du codage autre que la quantification

La plupart du temps, la compression sélective est traitée comme un problème d'ajustement de la consigne de quantification. Mais, ce n'est pas la seule façon de traiter le problème. Par exemple, A. Leontaris et al. [Leontaris 04] proposent d'augmenter les ressources mémoires en créant une référence constituée uniquement de zones d'intérêt prises à des moments temporels différents. Cette référence long terme doit permettre d'accroître les performances en terme de prédictions d'un algorithme de codage vidéo.
Afin de préserver la qualité des régions d'intérêt, M. Hannuksela et al. [Hannuksela 04] ont mis en oeuvre un procédé permettant de limiter la propagation d'erreurs de transmission. Il est clair que le codage de type prédictif permet d'accroître les performances d'un codage. L'aspect déplaisant du codage prédictif réside dans la capacité du codage prédictif à propager les erreurs de transmission. Pour atténuer ce problème, ces auteurs introduisent le concept de régions isolées. Ces régions se suffisent à elles mêmes pour être décodées. Cela signifie que la prédiction intra image n'est possible que dans les limites imposées par les frontières de la zone isolée. Par ailleurs, pour un type de prédiction inter image, un macrobloc de la région isolée ne peut être prédit qu'à partir des macroblocs appartenant à une région isolée d'une autre image. La notion de groupe d'images isolées est donc introduite.

1.4 Densité de saillance dédiée pour le codage

Dans un cadre de compression, il est nécessaire de modifier la densité de saillance DS pour prendre en compte un certain nombre de contraintes. Dans le contexte d'études que nous nous sommes fixés, celui de la compression vidéo H.264, la première que nous avons identifiée concerne l'obtention d'une valeur unique de saillance par macrobloc. La seconde concerne les zones découvertes. Pour des raisons de simplicité et de complexité, on se limite ici aux zones découvertes situées aux bords de l'image. Ces deux modifications conduisent à la détermination d'un densité de saillance modifiée, notée DS^{Mod}.
L'annexe C présente les grandes lignes de la compression H.264.

1.4.1 Passage au niveau macrobloc

L'obtention d'une valeur de saillance par macrobloc s'effectue simplement en moyennant les différentes valeurs de saillance localisées dans un macrobloc. Notons que d'autres opérateurs tels que la valeur médiane pourraient être utilisés.

1.4.2 Gestion des zones découvertes sur les bords de l'image

La gestion des zones découvertes d'une image est un problème complexe et difficile à traiter. On propose de détecter uniquement les zones découvertes apparaissant sur les bords de l'image ; l'objectif est d'attribuer une valeur de saillance maximale à ces zones et une valeur de saillance minimale aux zones recouvertes. En d'autres termes, on cherche à augmenter sensiblement la qualité des zones ayant potentiellement une durée de vie importante.
Ce problème de détection revient à caractériser le mouvement dominant en terme de translation (horizontale, verticale, et translation oblique) et de zoom (avant, arrière).
La première étape consiste alors à identifier les paramètres du mouvement dominant, donné par le vecteur θ ($\theta = [a_1, a_2, a_3, a_4, a_5, a_6]$), paramètres du mouvement affine), ayant une véritable influence sur le résultat. On cherche donc à savoir si une composante a_k de θ apporte une information significative. Deux hypothèses sont donc testées :

$$\begin{cases} H_0 : a_k & = & 0 \\ H_1 : a_k & \neq & 0 \end{cases} \tag{1.5}$$

L'hypothèse H_0 sera retenue si l'annulation d'un coefficient ne modifie pas substantiellement le modèle paramétrique. Dans le cas contraire, l'hypothèse H_1 est retenue. Par exemple si $k = 2$, ceci revient à comparer les modèles $(a_1, a_2, a_3, a_4, a_5, a_6)$ et $(a_1, 0, a_3, a_4, a_5, a_6)$. Les tests à mettre en oeuvre sont basés sur une loi de Student à $(N - p)$ degré de liberté, telle que décrite dans [Lebart 82]. Ici, la valeur N, c'est à dire le nombre d'observations, tend vers l'infini et la valeur p est égale à 6. La loi de Student tend alors vers la loi normale réduite. La valeur de chaque paramètre est testée par rapport au seuil de confiance 0.1.
Soit p_k la probabilité tirée de la distribution de Student P correspondant à la valeur de la variable de Student t_k (associée au paramètre a_k et obtenue via une "studentisation") prise par t :

$$p_k = P(|t| \geq t_k) \tag{1.6}$$

L'hypothèse H_0 est rejetée si la probabilité p_k est inférieure au seuil de confiance 0.1. Sinon, l'hypothèse selon laquelle le paramètre a_k a une d'influence est conservée.
Les paramètres jugés non pertinents sont mis à zéro. A partir du nouveau jeu de paramètres, le mouvement dominant est typé de la façon suivante :

- tous les paramètres sont nuls, excepté a_1, signifiant que le mouvement dominant est une translation horizontale. La direction de la translation est donnée par le signe du coefficient ;
- tous les paramètres sont nuls, excepté a_4, signifiant que le mouvement dominant est une translation verticale ;
- tous les paramètres sont nuls, excepté a_1 et a_4, signifiant que le mouvement dominant est une translation oblique ;
- tous les paramètres sont nuls, excepté a_2 et a_5, signifiant que le mouvement dominant est un zoom avant ou arrière suivant le signe des coefficients ;
- tous les paramètres sont nuls. La séquence vidéo traitée est un plan statique.

En fonction du type de mouvement, la densité de saillance est modifiée. Par exemple, les bandes blanches apparaissant sur la figure 1.1 traduisent la présence d'un mouvement

translationnel vers la gauche et vers le bas.

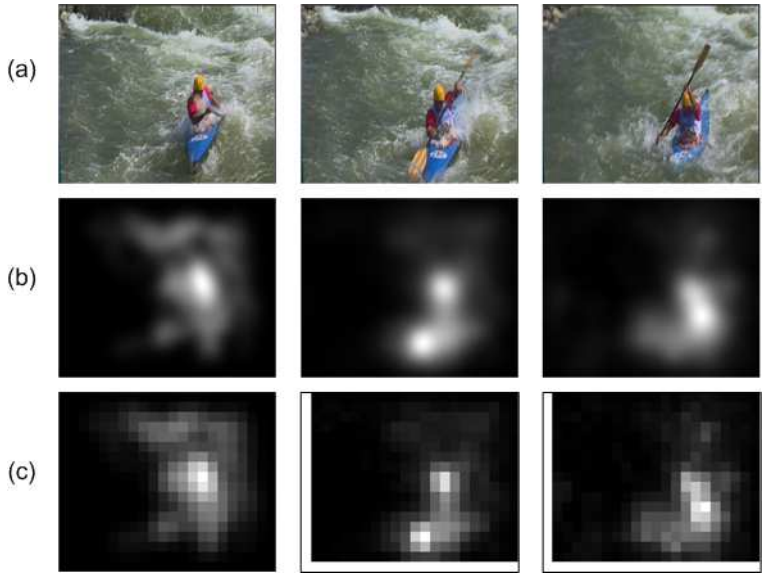

FIG. 1.1: Résultats de la modification de la densité de saillance : (a) image originale ; (b) densité de saillance DS ; (c) densité de saillance modifiée DS^{Mod}

1.5 Invariance de la carte de saillance à des artefacts de compression

Avant de décrire deux façons de mettre en oeuvre une compression sélective indirecte et directe, une question se pose et mérite qu'on s'y attarde : le modèle d'attention visuelle que nous proposons est-il invariant aux dégradations causées par un système dégradant de type compression vidéo ?
Si le résultat du modèle d'attention visuelle décrit auparavant est dépendant d'un système de compression, il va de soit que l'exploitation de densité de saillance dans un contexte de compression s'avère très délicate. Ce qu'on conçoit assez bien, mais faut-il encore le montrer, concerne le fait que l'attention d'un observateur peut être détournée par un artefact de codage.

Sensibilité de la stratégie visuelle d'un panel d'observateurs sur images fixes à un système de compression

Des tests oculométriques ont été menés sur deux images *Barba* et *Isabe* pour éclaircir ce problème. Ces deux images sont codées avec un codeur de type JPEG et un codeur de type JPEG2000. Le résultat de codage ainsi que les images originales, présentées à la figure 1.2, sont montrés à différents instants des tests oculométriques. Une densité de saillance est obtenue pour chaque image. Afin de comparer les stratégies visuelles et évaluer l'impact d'un codage sur la stratégie visuelle d'un observateur, le coefficient de corrélation est calculé entre la densité de saillance obtenue avec l'image originale et celle obtenue avec une image dégradée. Ce calcul est effectué pour trois durées de visualisation (2, 8 et

FIG. 1.2: Images *Barba* (première ligne) et *Isabe* (seconde ligne) : (a) image originales ; (b) images codées avec JPEG ; (c) images codées JPEG2000.

Image-Durée d'observation	2s	8s	14s
barba	0.986	0.979	0.981
isabe	0.988	0.991	0.987

TAB. 1.1: Évolution temporelle du coefficient de corrélation calculé entre la densité originale et la densité dégradée par un codage jpeg (r5).

14 secondes). Les résultats sont donnés dans les tableaux 1.1 et 1.2. Les coefficients de corrélation sont très élevés et cela quelle que soit la durée d'observation. Cela signifie que la stratégie visuelle n'a pas fondamentalement été modifiée par les dégradations causées par la compression.
La figure 1.3 montre les points fixés par les observateurs en fonction du temps d'observation (au plus une fixation par pixel). On constate que la distribution des points de fixation est plus bruitée lorsque l'image est dégradée. Cette différence de distribution reste néanmoins peu significative.

Image-Durée d'observation	2s	8s	14s
barba	0.981	0.984	0.98
isabe	0.991	0.979	0.979

TAB. 1.2: Évolution temporelle du coefficient de corrélation calculé entre la densité originale et la densité dégradée par un codage jpeg2000 (r4).

En conclusion, la stratégie visuelle semble être peu modifiée par un système de compression, pour des taux de compression moyens. Des tests complémentaires sont à mener sur des séquences d'images pour confirmer ce résultat.

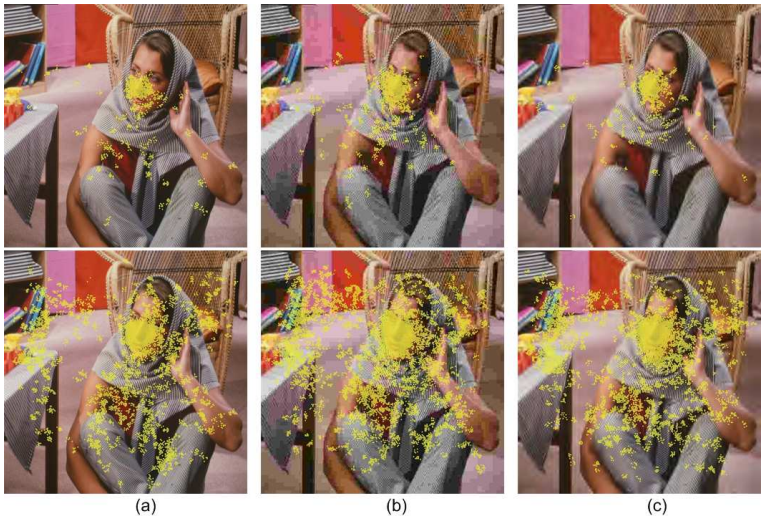

FIG. 1.3: Superposition des points fixés (au plus une fixation par pixel) sur l'image *Barba* pour deux durées d'observation (2 et 14 secondes) : (a) image originales ; (b) images codées avec JPEG ; (c) images codées JPEG2000.

Sensibilité à des artefacts de codage vidéo du modèle d'attention visuelle proposé

Afin d'examiner l'invariance de la densité de saillance prédite vis à vis d'artefacts de codage, on utilise la séquence *F1 Car* et ses versions dégradées utilisées par le groupe de normalisation VQEG (*Video Quality Expert Group*). Cette séquence (*src6_ref__625*), présentant de forts mouvements et des couleurs saturées est une séquence entrelacée de 625 lignes ayant une cadence temporelle de 25 images par seconde. Les séquences dégradées sont obtenues via des procédés bien précis et détaillées dans le rapport [VQEG]. Les quinze séquences sont numérotées de 1 à 15, précédant le nom de la séquene et du sigle HRC (*Hypothetical Reference Circuits*). Les séquences *src6_hrc1_625* à *src6_hrc9_625* sont le résultat d'un codage qualifié haute qualité (débit allant de 3Mb/s à 50Mb/s). Les autres sont obtenues pour un débit allant de 768kb/s-4.5Mb/s et sont considérées comme étant de basse qualité.

Nous avons calculé le coefficient de corrélation linéaire entre la densité de saillance issue de la source et la densité de saillance provenant d'une séquence dégradée. Les valeurs de corrélation sont données dans le tableau 1.3 pour toutes les dégradations envisagées. Quelle que soit la dégradation, le coefficient de corrélation est supérieur à 0.95. Cela signifie que les dégradations apportées par les différents codages n'entraînent pas de modifications majeures dans les densités de saillance. Le modèle proposé semble donc être invariant aux dégradations de type codage, même pour d'assez fortes dégradations.

TAB. 1.3: Coefficient de corrélation calculé entre la densité de saillance issue de la source et une densité de saillance provenant d'une séquence dégradée. Les séquences dégradées et le type de dégradation sont identifiés par le terme *hrc* suivi d'une numérotation [VQEG].

	hrc1	hrc2	hrc3	hrc4	hrc5	hrc6	hrc7	hrc8
cc	0,96	0,98	0,98	0,97	0,98	0,97	0,98	0,97
	hrc9	hrc10	hrc11	hrc12	hrc13	hrc14	hrc15	hrc16
cc	0,96	0,98	0,97	0,97	0,96	0,97	0,95	0,96

1.6 Compression sélective indirecte

1.6.1 Objectif

Dans le même ordre d'idée que celui proposé par L. Itti [Itti 04], la compression sélective indirecte consiste à réduire la quantité d'information de la séquence d'images à coder tout en conservant l'information originale sur les zones de focalisation. Un prétraitement piloté par la carte de saillance est généralement utilisé. L'objectif est de diminuer le débit de codage tout en maintenant une bonne qualité sur les zones visuellement intéressantes des images. Le choix du type de prétaitement à appliquer est important car il doit être complémentaire du codage et plus particulièrement de l'opérateur de quantification. Un prétaitement non-linéaire est certainement le choix le plus judicieux puisqu'il est peu redondant avec une quantification.

1.6.2 Définition du prétraitement utilisé

Le prétraitement que nous utilisons appartient à la famille des filtres morphologiques connexes. Le filtre, appelé nivellement [Meyer 98], est la combinaison de deux opérateurs morphologiques qui sont une ouverture par reconstruction et une fermeture par reconstruction. Ces deux filtres sont détaillés dans les paragraphes suivants. Ce filtre est intéressant à la fois pour sa non-linéarité et sa capacité à conserver les structures.
En reprenant la terminologie classique utilisée pour caractériser des filtres morphologiques et celle de C. Gomila [Gomila 01], les notations sont :

- f fonction originale ;
- g fonction marqueur ;
- g' fonction filtrée ;
- B élément structurant ;
- $\wedge$ est l'opérateur minimum ;
- $\vee$ est l'opérateur maximum ;
- ϵ_B érosion sur l'élément structurant B ;
- δ_B dilatation sur l'élément structurant B.

L'opération morphologique d'ouverture, érosion suivie d'une dilatation $\delta_B\,[\epsilon_B(f)]$, (respectivement de fermeture, une dilatation suivie d'une érosion $\epsilon_B\,[\delta_B(f)]$) est relativement bien connue. L'ouverture morphologique (respectivement fermeture) permet de supprimer les parties claires trop petites pour contenir l'élément structurant (respectivement de supprimer les parties sombres de l'image trop petites vis à vis de la taille de l'élément

structurant). Ces deux opérateurs peuvent être contraints par une fonction marqueur g. On parle alors d'ouverture et de fermeture par reconstruction.

Ouverture par reconstruction

g' est une ouverture par reconstruction de f, noté $Rec(g', f, g)$, si $g' \leq f$ et $g' = g' \wedge g$. Pour construire la fonction marqueur g, il faut simplement avoir à l'esprit que $g \geq g'$. Classiquement, g est obtenue par dilatation de la fonction g'.

Fermeture par reconstruction

g' est une fermeture par reconstruction de f, noté $\widetilde{Rec}(g', f, g)$, si $g' \geq f$ et $g' = g' \vee g$. Comme précédemment, pour construire la fonction marqueur g, il faut que $g \leq g'$. Classiquement, g est obtenue par érosion de la fonction g'.
A partir de ces deux opérateurs, le filtre de nivellement est défini.

Le nivellement

Le filtre de nivellement est obtenu en combinant une ouverture et une fermeture par reconstruction. Sa définition est la suivante :
g' est un nivellement de la fonction f, si et seulement si, la relation suivante est vérifiée :

$$\forall s, \quad f_s \wedge \delta g_s \leq g_s \leq f_s \vee \epsilon g_s$$

f_s (respectivement g_s) représente la valeur de la fonction f (respectivement g) au site s. En pratique, la fonction g' est le plus fort nivellement de f. La fonction g' est obtenue en itérant le critère d'assignation suivant et en modifiant la fonction marqueur g (extrait de la thèse de C. Gomila [Gomila 01]) :

- lorsque $g < f$, on doit remplacer g par δg jusqu'à la création d'une zone plate ou jusqu'à ce que la fonction g' soit égale à la fonction f ;
- lorsque $g > f$, on doit remplacer g par ϵg jusqu'à la création d'une zone plate ou jusqu'à ce que la fonction marqueur g' soit égale à la fonction f

Un exemple de nivellement est donné à la figure 1.4. Le filtre de nivellement peut s'implanter de façon itérative ou par files d'attente [Gomila 01].

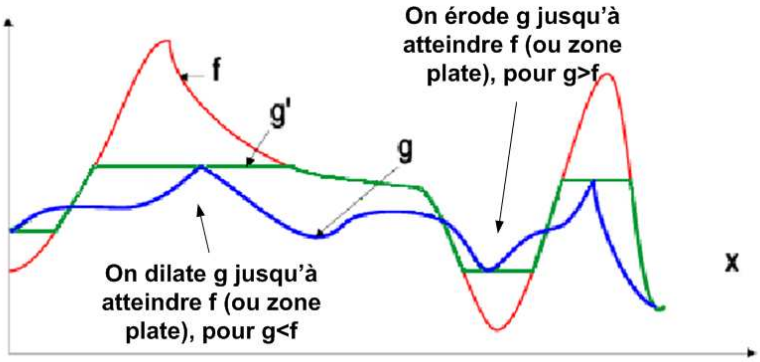

FIG. 1.4: Construction d'un nivellement à partir d'une fonction marqueur (extrait de [Gomila 01]).

1.6.3 Le nivellement couplé à une carte de saillance

La compression sélective indirecte a pour objectif de simplifier le contenu à coder tout en conservant le maximum d'information sur les zones d'intérêt. Pour préserver la qualité des zones d'intérêt, le nivellement est désactivé sur les zones de forte saillance. Une zone de saillance est définie comme une zone circulaire d'un rayon de 1 degré visuel, centrée sur un maximum local. L'inhibition de cette zone permet de déterminer une seconde zone de saillance et ainsi de suite. Le nombre de zones saillante N est un paramètre important : un nombre élevé de zones saillantes conduira à une réduction de débit négligeable. Un nombre faible, favorisant une zone localisée de la séquence, permettra de réduire considérablement le débit de codage.
Un exemple de nivellement sans et avec carte de saillante, pour deux tailles d'élément structurant 3×3 et 11×11, est donné figure 1.5. Les deux tailles d'éléments structurant sont prises sciemment très différentes.

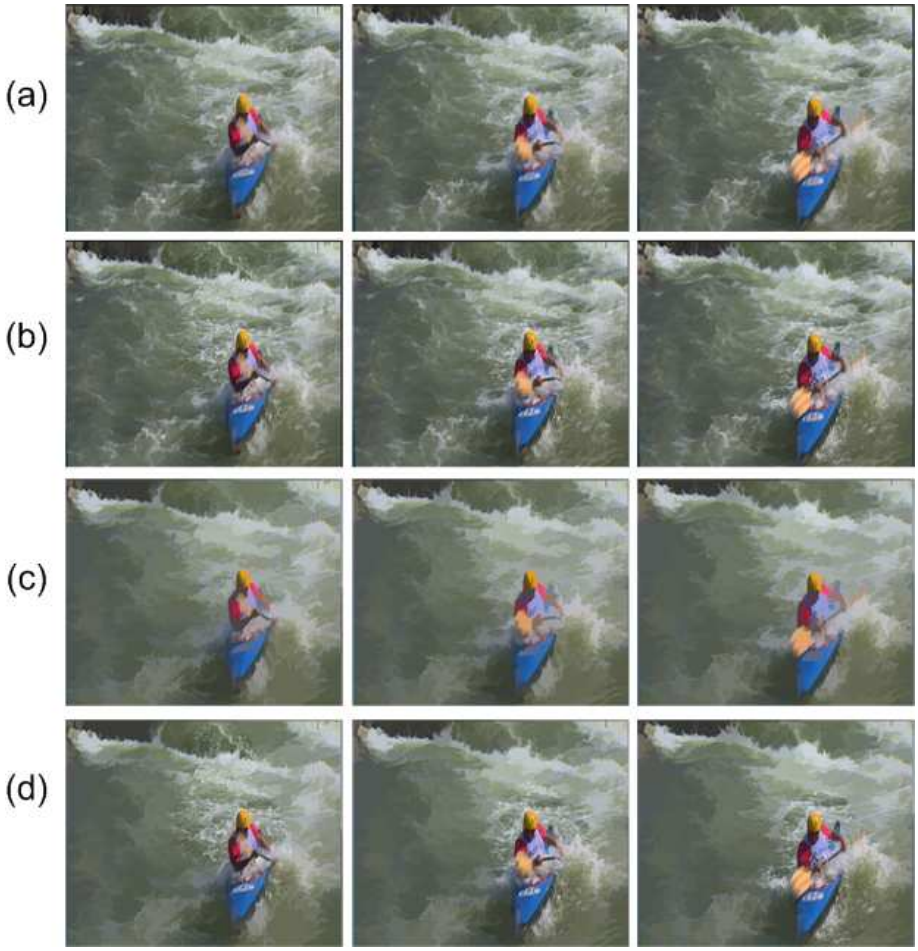

FIG. 1.5: Exemple de nivellement sans et avec la carte de saillance : (a) sans carte de saillance, taille de l'élément structurant 3×3 ; (b) avec 4 zones de saillance, taille de l'élément structurant 3×3 ; (c) sans carte de saillance, taille de l'élément structurant 11×11 ; (d) avec 4 zones de saillance, taille de l'élément structurant 11×11.

TAB. 1.4: Débit de codage et gain en pourcentage vis à vis de la référence. Ces valeurs sont données en fonction du nombre de zones de saillance considérées et de la taille de l'élément structurant.

	3x3		5x5		7x7		11x11	
	Débit	Gain	Débit	Gain	Débit	Gain	Débit	Gain
4 zones	436	24.5%	396	31.5%	374	35.3%	352	39%
8 zones	482	16.6%	454	21.4%	437	24.4%	423	27%
12 zones	514	11.1%	492	14.9%	481	16.78%	469	18.9%

1.6.4 Évaluation de l'approche

La compression sélective indirecte est ici testée en faisant varier à la fois la taille de l'élément structurant B et le nombre de zones saillantes N. L'évaluation consiste à comparer le coût de codage d'une séquence sans et avec prétraitement. La séquence *Kayak* (format CIF, *Common Intermediate Format*, 352×288) est codée en boucle ouverte avec un index de quantification de 37, conduisant à un débit de 580 kb/s (50 images sont considérées).
La figure 1.6 donne le gain de débit (en pourcentage) par rapport à notre référence codée (580 kb/s). Après filtrage et codage, la qualité des zones saillantes est exactement la même que celle de la référence codée.

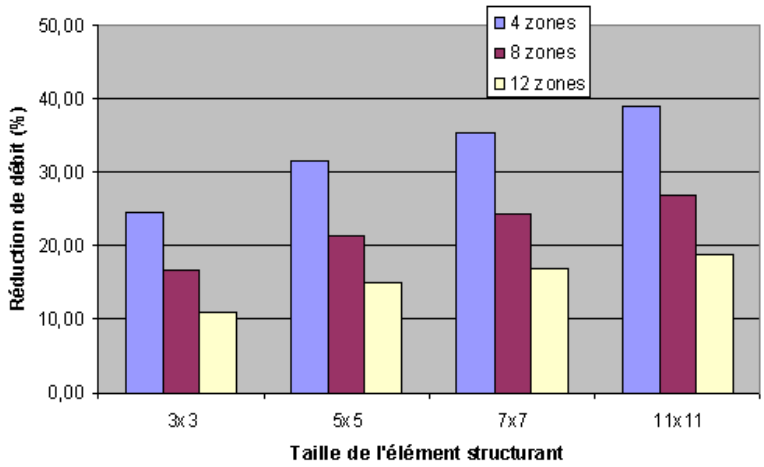

FIG. 1.6: Gain en terme de débit sur la séquence *Kayak* en fonction de la taille de l'élément structurant et du nombre de zones saillantes.

Les valeurs numériques sont données dans le tableau 1.4 : Les résultats sont à interpréter avec précaution, puisqu'il faut considérer l'application visée. Si elle consiste à favoriser clairement les zones de saillance (pour une application de télé-surveillance, par exemple), le débit le plus faible possible peut être souhaité. Dans ce cas, le nombre de zones de saillance à considérer est faible et le prétraitement agressif. Dans un contexte de diffusion vidéo, les dégradations apportées par le prétraitement ne doivent pas être gênantes. Un

élément structurant de petite taille est souhaitable si le nombre de zones est faible. Dans le cas contraire, un élément structurant de taille plus grande peut être utilisé pour filtrer fortement et localement.
Finalement, l'intérêt d'une telle approche réside d'une part dans la possibilité de préserver la qualité des zones saillantes et d'autre part dans le fait d'être indépendant de la technologie de compression utilisée.

1.7 Compression sélective directe

1.7.1 Objectif

L'objectif de la compression sélective directe est de contrôler la distribution des ressources de codage en fonction de l'intérêt visuel de chaque macrobloc afin d'accroître la qualité visuelle perçue. A. Bradley [Bradley 03] a d'ailleurs montré qu'une compression sélective sur images fixes permettait d'améliorer la qualité subjective d'une part lorsque les zones d'intérêt sont de tailles relativement faibles et d'autre part, lorsque pour une approche classique de codage, le débit de consigne provoque l'apparition d'artefacts de codage sur les zones saillantes. La plupart du temps, le paramètre sur lequel on agit est la consigne de quantification. En d'autres termes, un macrobloc présentant un intérêt visuel faible sera quantifié plus grossièrement qu'un macrobloc ayant un intérêt visuel important. Afin de conserver le débit cible imposé par l'utilisateur, il est impératif de contrôler le surcoût engendré par une quantification plus fine ainsi que le gain engendré par une quantification plus forte. Par ailleurs, il est également intéressant de connaître la variation de la distortion engendrée par la modification de la consigne de quantification. Pour ces deux raisons et pour obtenir un ajustement local de la consigne de quantification, il est nécessaire de connaître pour chaque macrobloc sa fonction débit-distortion. Avant de décrire l'algorithme d'adaptation de l'index de quantification envisagée, la détermination de la fonction débit-distortion est détaillée.

1.7.2 La fonction débit-distortion

La fonction débit-distortion est obtenue en codant chaque macrobloc avec différents index de quantification. Pour chaque codage, un point débit-distortion, dit point de contrôle, est obtenu. Ici, la distortion est représentée par l'erreur quadratique moyenne (EQM ou SSE pour le terme anglais). L'ensemble de ces points est ensuite placé dans les repères index de quantification en fonction du coût $(Q, Cout)$ et index de quantification en fonction de la distortion (Q, SSE). La figure 1.7 donne un exemple des deux courbes déduites de points débit-distortion. Le coût d'un macrobloc est la somme des coûts des coefficients quantifiés et des coûts de syntaxe.
Afin d'obtenir une granularité à l'index de quantification près, un modèle d'approximation de la courbe est effectué. La modélisation que nous avons envisagée est linéaire pour les deux premiers et les deux derniers points de contrôle. Le reste de la courbe est obtenu via une modélisation par spline cubique. La modélisation par fonction spline cubique, fortement inspirée des travaux de L. Lin [Lin 97], consiste à déterminer entre deux points de contrôle successifs (x_i, y_i) et (x_{i+1}, y_{i+1}) un polynôme de la forme (x_i et y_i représentent

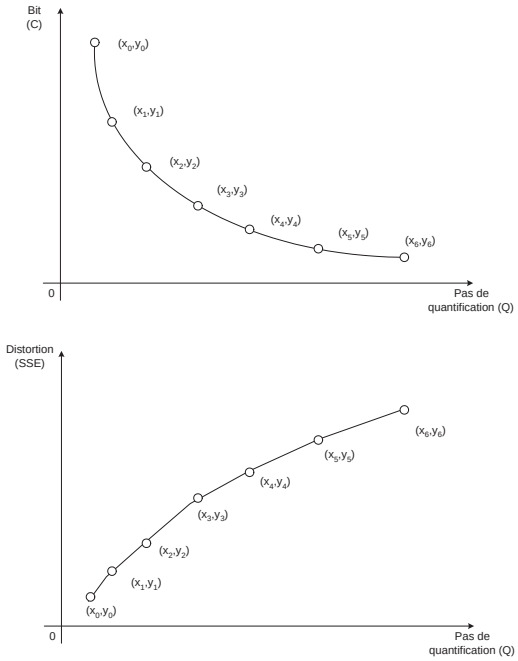

FIG. 1.7: Points de contrôle pour les courbes $Cout = f(Q)$ et $SSE = f(Q)$. Exemple donné lorsque 7 points de contrôle ont été préalablement déterminé.

respectivement l'index de quantification et le débit ou la distortion selon qu'on considère la courbe $Cout = f(Q)$ ou $SSE = f(Q)$) :

$$f_i(x) = a_i \cdot x^3 + b_i \cdot x^2 + c_i \cdot x + d_i \tag{1.7}$$

Etant donné qu'on a besoin de 4 points de contrôle pour la modélisation spline (x_{i-1}, y_{i-1}), (x_i, y_i), (x_{i+1}, y_{i+1}) et (x_{i+2}, y_{i+2}), les deux premiers et les deux derniers points de contrôle ne sont pas considérés. Pour $K+1$ points de contrôle, on détermine donc $K-2$ polynômes. Les paramètres a_i, b_i, c_i et d_i se déduisent des contraintes suivantes :

– la fonction spline modélisée doit passer par les points (x_i, y_i) et (x_{i+1}, y_{i+1}) :

$$a_i \cdot x_i^3 + b_i \cdot x_i^2 + c_i \cdot x_i + d_i = y_i \tag{1.8}$$

$$a_i \cdot x_{i+1}^3 + b_i \cdot x_{i+1}^2 + c_i \cdot x_{i+1} + d_i = y_{i+1} \tag{1.9}$$

– la dérivée première doit être continue aux points de contrôle. Cette condition peut être obtenue en définissant la pente au point de contrôle x_i :

$$f_i'(x_i) = f_{i-1}'(x_i) = \frac{y_{i+1} - y_{i-1}}{x_{i+1} - x_{i-1}} \tag{1.10}$$

En prenant la dérivée de (1.7) et en remplaçant dans (1.10) pour les deux points de contrôle considérés, on obtient les deux nouvelles conditions suivantes :

$$3a_i \cdot x_i^2 + 2b_i \cdot x_i + c_i = \frac{y_{i+1} - y_{i-1}}{x_{i+1} - x_{i-1}} \tag{1.11}$$

$$3a_i \cdot x_{i+1}^2 + 2b_i \cdot x_{i+1} + c_i = \frac{y_{i+2} - y_i}{x_{i+2} - x_i} \tag{1.12}$$

A partir de (1.8), (1.9), (1.11) et (1.12), les paramètres a_i, b_i, c_i et d_i peuvent être déduits. Toutefois, pour simplifier le calcul, un changement de variable peut être réalisé pour translater x_i vers 0 et x_{i+1} vers 1, en utilisant

$$z(x) = \frac{x - x_i}{x_{i+1} - x_i} \tag{1.13}$$

Le polynôme (1.7) devient donc

$$f_i(z) = a_i \cdot z^3 + b_i \cdot z^2 + c_i \cdot z + d_i \tag{1.14}$$

On obtient ainsi les valeurs suivantes :

$$\begin{aligned} d_i &= y_i, && (1.15)\\ c_i &= \frac{y_{i+1} - y_{i-1}}{x_{i+1} - x_{i-1}}, && (1.16)\\ b_i &= 3(y_{i+1} - y_i) - 2 \cdot c_i - \beta, && (1.17)\\ a_i &= -2(y_{i+1} - y_i) + c_i + \beta, && (1.18)\\ \beta &= \frac{y_{i+2} - y_i}{x_{i+2} - x_i}. && (1.19) \end{aligned}$$

Enfin, pour les deux premiers et les deux derniers points de contrôle, une interpolation linéaire est effectuée. Les paramètres du polynôme sont pour les points de contrôle (x_i, y_i) et (x_{i+1}, y_{i+1}) :

$$\begin{aligned} d_i &= y_i - c_i \cdot x_i, && (1.20)\\ c_i &= \frac{y_{i+1} - y_i}{x_{i+1} - x_i}, && (1.21)\\ b_i &= 0, && (1.22)\\ a_i &= 0. && (1.23) \end{aligned}$$

1.7.3 Modification du coeur de codage

La modification du coeur de codage se fait en deux étapes. La première étape consiste à déterminer l'index de quantification de chaque macrobloc pour satisfaire un coût minimal imposé. La détermination des index de quantification doit permettre, dans la mesure du possible, d'obtenir une qualité homogène sur l'image. La seconde étape consiste à redistribuer le surplus de débit en favorisant les zones saillantes de l'image à coder.

Notations

Les notations que nous allons utiliser sont les suivantes :

- $C_{Consigne}$, le coût consigne, c'est à dire le coût de codage autorisé pour coder l'image courante ;
- C_{Init}, le coût initial, imposé arbitrairement. Dans un premier temps, ce coût représente 50% du coût de codage de consigne. A partir de ce côut, un index de quantification est déterminé pour chaque macrobloc ;
- $c^k(i)$ coût associé au macrobloc i pour un index de quantification égal à k ;
- $q_{Init}(i)$, index de quantification du macrobloc i, obtenu en fonction de C_{Init} ;

- DS^{Mod}, densité de saillance modifiée en vue d'un codage ;
- $d_v^k(i)$, distortion visuelle associée au macrobloc i pour un index de quantification de valeur k ;
- $\overline{d_v}$, distortion visuelle moyenne sur l'image ;
- σ_v^2, variance de la distortion visuelle sur l'image ;
- $d_{vi}^k(i)$, distortion visuelle avec prise en compte des zones d'intérêt perceptuelles associée au macrobloc i pour un index de quantification égal à k : d_{vi} est fonction de la distortion visuelle d_v et de la densité de saillance modifiée DS^{Mod}. Cette fonction a une valeur nulle si la valeur DS^{Mod} associée est nulle ;
- ΔC, coût à distribuer sur les zones d'intérêt perceptuel : $\Delta C = C_{Consigne} - C_{Init}$.

Ajustement local des index de quantification en fonction de C_{Init}

A partir du coût de codage initial, noté C_{Init}, un index de quantification est associé à tous les macroblocs de l'image. Chaque macrobloc i est donc associé à un index de quantification $q_{Init}(i)$. Afin d'uniformiser la qualité (en terme d'EQM), un ajustement local des index de quantification est effectué. L'uniformisation de la qualité est effectuée pour tenir compte de la règle suivante : la qualité perçue est maximale lorsque la qualité locale perçue est la même partout. Bien évidemment, cet algorithme fonctionne avec un critère de qualité très simple et peu représentatif d'un jugement subjectif. Néanmoins, à notre connaissance, il n'existe pas à l'heure actuelle de métrique de qualité suffisamment simple et pertinente pour la remplacer. L'algorithme utilisé, illustré figure 1.8, est le suivant :

1. les macroblocs (représentant $2p\%$ de la population totale de macrobloc) ayant des valeurs de qualité les plus faibles et les plus fortes sont considérer ;
2. pour ces macroblocs, on augmente (respectivement on diminue) l'index de quantification pour diminuer (respectivement augmenter) leur qualité. Ensuite, la valeur moyenne de la distorsion $\overline{d_v}$ ainsi que la variance σ_v^2 sont recalculées. Si la différence entre la valeur de variance précédente et actuelle est jugée très faible (inférieure à une quantité epsilon), la redistribution est terminée. Sinon, on retourne à l'étape 1 pour continuer la redistribution.

Distribution de la quantité ΔC en fonction d'une carte de saillance modifiée DS^{Mod}

La distribution de la quantité ΔC consiste à déterminer itérativement le macrobloc de l'image qui apporterait le plus en terme de qualité pour un surcoût minimum. Si l'indice j représente l'indice d'itération, la variation $\lambda_j^{k\to(k-1)}(i)$ en terme de distortion et de débit, lorsque l'index de quantification est décrémenté, est calculée pour chaque macrobloc i de la façon suivante :

$$\lambda_j^{k\to(k-1)}(i) = \frac{d_{vi}^k(i) - d_{vi}^{(k-1)}(i)}{c^{(k-1)}(i) - c^k(i)} \tag{1.24}$$

Cette quantité est bien évidement positive puisque la distortion $d_{vi}^k(i)$ est supérieure à $d_{vi}^{(k-1)}(i)$ et que le coût $c^{(k-1)}(i)$ est supérieur à $c^k(i)$.
Lorsque la variation $\lambda_j^{k\to(k-1)}$ est déterminée pour tous les macroblocs, le macrobloc qui est le plus intéressant à coder finement est celui qui possède la plus grande variation :

$$\lambda_{max} = \max_{i,j} \lambda_j^{k\to(k-1)}(i) \tag{1.25}$$

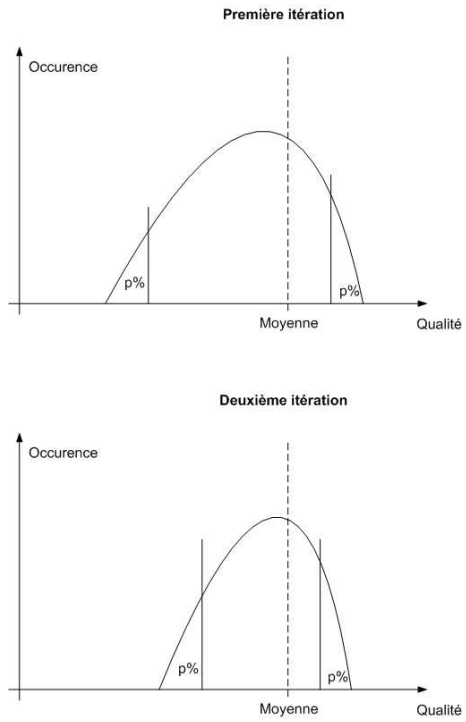

FIG. 1.8: Algorithme d'uniformisation de la qualité.

L'index de quantification du macrobloc ayant la variation maximale est donc décrémenté. La quantité ΔC est alors mise à jour. La boucle itérative continue tant que la quantité ΔC est positive.

1.7.4 Résultats

TAB. 1.5: Mesures de PSNR moyen calculés sur l'image complète et sur les zones de saillance pour trois séquences.

Séquence	Paramètres de codage	$PSNR_{image}$		$PSNR_{RoI}$	
		Classique	Adapté	Classique	Adapté
Kayak	1 Mb/s, CIF, 50 images	30.95	28.1	30.46	32.72
	550 kb/s, CIF, 50 images	28.87	27.59	28.66	30.13
Stefan	1 Mb/s, CIF, 50 images	35.71	33.43	34.52	37.45
	550 Mb/s, CIF, 50 images	32.02	30.41	30.65	31.63
Raid	1.4Mb/s, 640×352, 125 images	34.99	34.46	34.4	35.23

Le fonctionnement de cette approche a été tout d'abord validé sur deux séquences au format CIF et sur une séquence ayant une résolution de 640×352 pixels. Il s'agit ici de comparer la qualité de la zone d'intérêt lorsqu'on utilise un codage classique ou un codage adapté. La métrique de qualité utilisée est le PSNR (*Peak Signal Noise Ratio*). Le seul avantage de cette métrique est sa simplicité. Les inconvénients sont nombreux et bien connus : faible corrélation avec les tests visuels de qualité, incapacité de la métrique

à prendre en compte des effets de masquage des dégradations... En dépit de cela, la validation du fonctionnement de l'algorithme a été effectuée avec ce type de métrique. Le tableau 1.5 présente les différents résultats obtenus. La qualité globale en terme de PSNR est fortement diminuée lorsqu'on considère une approche adaptée, ce qui était attendu. En contre partie, la qualité des zones d'intérêt est nettement améliorée. Les résultats de

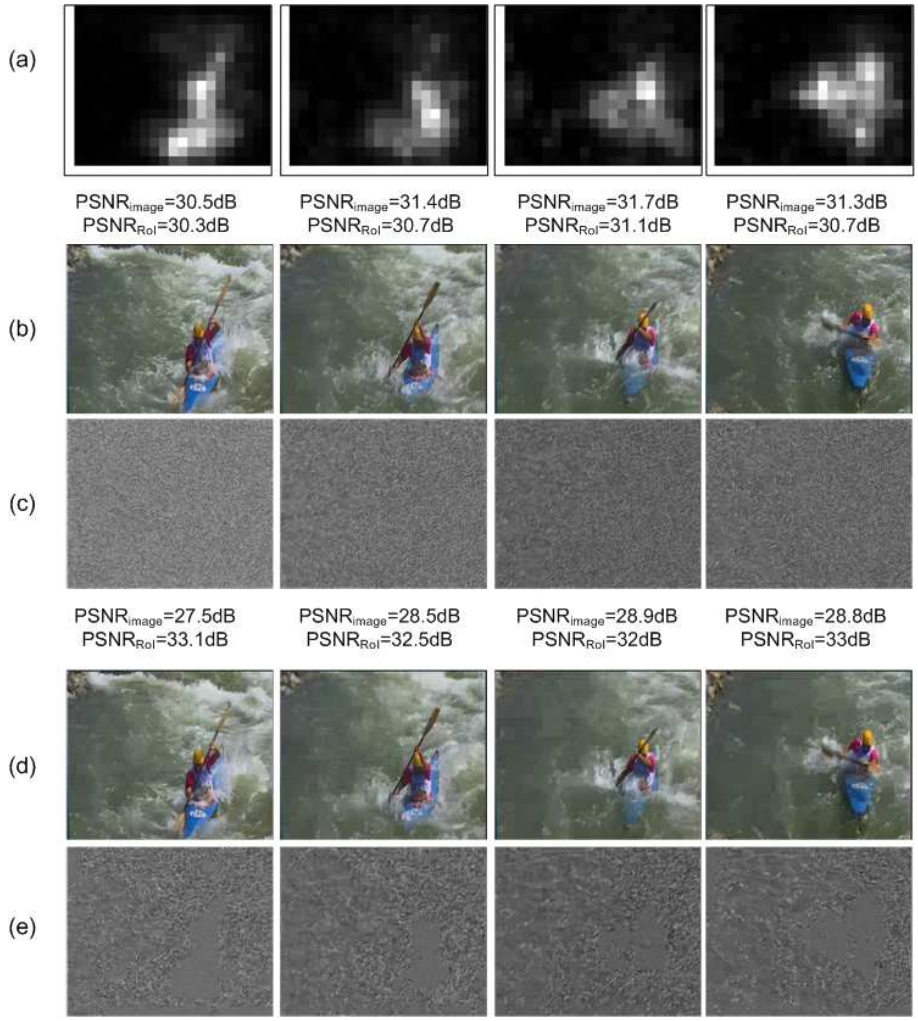

FIG. 1.9: Exemple de résultats sur la séquence *Kayak*. (a) carte de saillance modifiée pour un codage ; (b) codage classique ; (c) erreur engendrée par un codage classique ; (d) codage adapté ; (e) erreur engendrée par un codage adapté.

compression obtenus par un codage classique et un codage adapté sont donnés figure 1.9 pour la séquence *Kayak*, codée à 1 Mb/s. Il est particulièrement intéressant de considérer les erreurs engendrées par un codage classique (c) et un codage adapté (e). On constate, en effet, que pour le premier, l'erreur est uniforme alors que pour le second la distribution de l'erreur de codage est fonction de la carte de saillance. L'objectif que nous souhaitions est donc atteint : favoriser la qualité des zones d'intérêt au détriment des zones de non intérêt.

La figure 1.10 présente un résultat de codage obtenu sur la séquence *Stefan*. Les cartes d'erreur témoignent particulièrement bien de la nette amélioration de la zone saillante. Les jambes, les bras et le visage du joueur sont particulièrement mieux codés par la méthode proposée que par la méthode classique. La séquence *Stefan* est une séquence type pour laquelle le codage sélectif peut grandement améliorer la qualité perçue. En effet, cette séquence se compose d'une zone saillante de taille relativement faible et d'un arrière plan contenant des zones fortement texturées, difficile à coder. Ce type d'arrière plan présente deux intérêts : tout d'abord, le codage de l'arrière plan est fortement consommateur de débit. Il est donc possible de venir puiser du débit dans cette partie pour l'affecter à la zone saillante. Le second point intéressant est la présence d'une forte texture ; ce type de texture a une capacité intrinsèque à masquer le bruit de quantification et donc par voie de conséquence peut supporter une quantification plus forte sans introduire de défauts perceptibles. Ce second point ne peut être atteint que si la métrique de distortion utilisée est fortement évoluée, basée sur des propriétés du SVH et travaillant au niveau macrobloc.

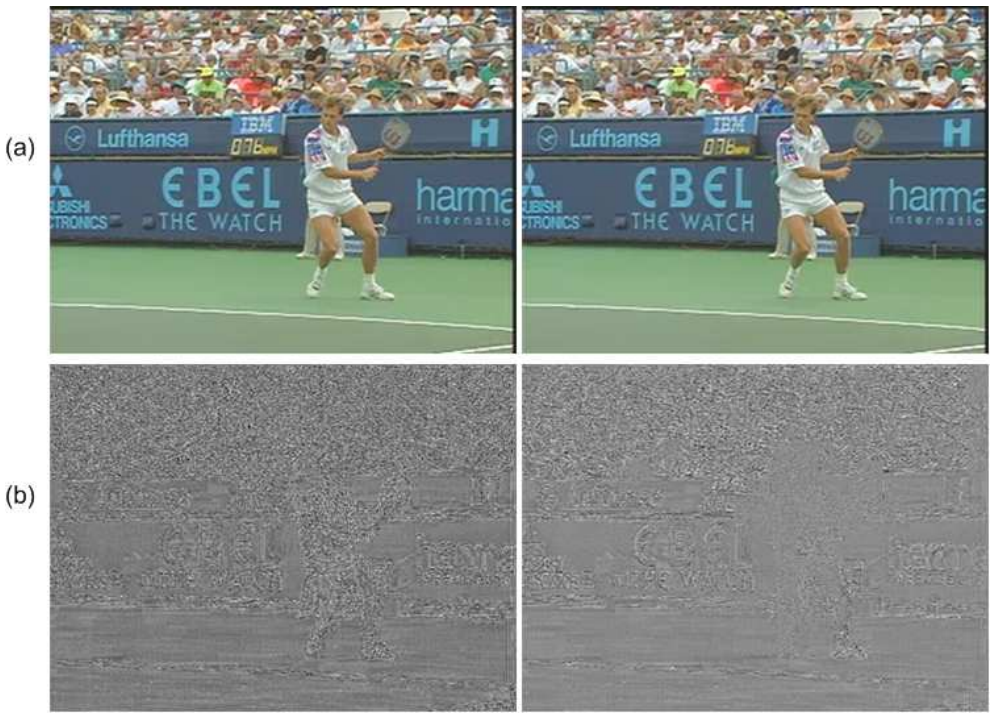

Fig. 1.10: Exemple de résultats de codage (1 Mb/s CIF) classique (image de gauche (a)) et adapté (image de droite (a)) associés avec les cartes d'erreur (b).

La figure 1.11 présente la distribution des coûts de codage macrobloc au niveau image. Un codage classique distribue les ressources de codage en fonction de la complexité de l'image. On constate par exemple que l'arrière plan de la séquence *Stefan* s'octroie la majeure partie des ressources. La compression sélective permet de concentrer ces ressources de codage sur les zones d'intérêt, conduisant à l'amélioration de leur qualité.

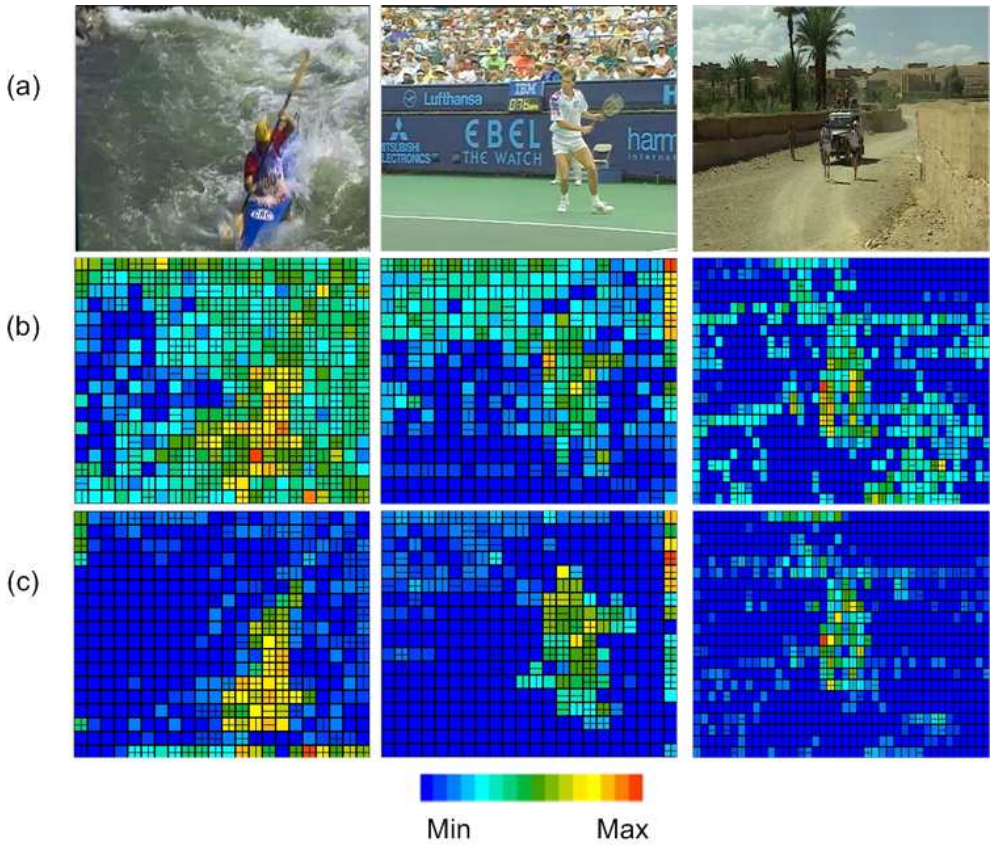

FIG. 1.11: Coût de codage pour trois images issues de trois séquences différentes : (a) images codées via une approche classique ; (b) coûts de codage macrobloc obtenus pour un codage classique ; (c) coûts de codage macrobloc pour l'approche de compression sélective proposée.

1.7.5 Limitations de l'approche

Cette approche de compression sélective décrite précédemment soulève plusieurs questions :
La première réside dans la nécessaire connaissance de la courbe débit-distortion pour chaque macrobloc. Actuellement, il n'y a pas vraiment de solutions efficaces pour déterminer cette fonction, autre que celle de coder systématiquement chaque macrobloc avec plusieurs index de quantification. Cette complexité calculatoire est considérable et la mise en oeuvre d'un tel procédé s'avère donc difficile. Il existe bien entendu des façons de simplifier l'obtention de la courbe débit-distortion. Par exemple, dans de nombreuses situations [Uz 93], [Frimout 93], [Chen 93], [Hang 97], l'utilisation de distributions Gaussiennes, Laplaciennes ou encore des Gaussiennes généralisées sont utilisées pour modéliser l'évolution du coût de codage en fonction de l'index de quantification. Concernant la distortion, elle est le plus souvent considérée comme linéairement dépendante de la consigne de quantification. Bien évidemment, ces approximations affectent clairement les résultats tant du point de vue débit que distortion. Une autre solution en cours d'investigation est d'utiliser une régulation basée sur le ρ-domaine [He 01a, He 01b, Milani 03]. Cette

solution permettra d'obtenir facilement l'évolution du débit en fonction de la consigne de quantification.

La seconde question concerne la métrique de qualité utilisée pour évaluer la gêne liée à la distortion. Une métrique telle que l'EQM (Erreur Quadratique Moyenne) calculée entre l'image originale et l'image dégradée, ne peut être satisfaisante pour réaliser une adaptation correcte de l'index de quantification. Tout d'abord, cette mesure n'est pas invariante à de légères transformations appliquées sur l'image dégradée (rotation, translation, zoom...). De la même façon, un éclaircissement ou un assombrissement engendrent une forte augmentation de l'EQM. Par ailleurs, même si on fait l'hypothèse que l'image dégradée n'a pas subi ce type de transformations, les résultats issus de l'EQM, via le PSNR, sont faiblement corrélés avec les tests subjectifs. Bien que la maîtrise de la qualité perçue soit un facteur déterminant pour de nombreux domaines et notamment pour la compression sélective, il n'existe pas ou peu de méthodes de qualité basées macrobloc simples et efficaces.

Enfin, la compression sélective directe, dans un contexte de diffusion, n'est pertinente que pour des configurations particulières :

- pour des situations confortables de codage : ce type de situation, où le coût de la syntaxe est considéré négligeable devant le coût de codage des coefficients transformés quantifiés, permet d'envisager sereinement un transfert de débit des zones non saillantes vers les zones saillantes. La qualité de reconstruction des zones saillantes mesurée via une métrique classique sera alors incontestablement améliorée. Mais, l'objectif, qui est d'accroître la qualité perçue à débit équivalent, n'est pas forcément atteint. En effet, cette situation est paradoxale puisque le gain en qualité subjective est plus facilement perceptible à bas débit qu'à moyen débit. L'effort à fournir pour atteindre un gain doit alors être plus important ;
- pour des séquences particulières ayant une structure proche de celle de la séquence *Stefan* : on y retrouve une zone d'intérêt de taille relativement faible (on retrouve une conclusion des travaux de A. Bradley [Bradley 03]) et un arrière plan consommateur de débit, présentant des capacités de masquage visuel du bruit de quantification. Pour ce type de séquences, et particulièrement les séquences de sport intégrant des plans sur des tribunes, la compression sélective directe peut être extrêmement bénéfique sur deux aspects complémentaires : le premier concerne l'amélioration de la qualité perçue à débit équivalent et le second concerne la conservation de la qualité perçue à débit fortement réduit. Sur ce dernier point, les travaux de L. Huguenel [Huguenel 05] ont montré qu'il était possible de réduire le débit de séquences de sport intégrant des tribunes de 30% tout en conservant la qualité subjective constante. La première rangée de la figure 1.12, provenant des travaux de L. Huguenel, correspond à un codage classique, utilisant un index de quantification uniforme (image de droite de la rangée (a)). La seconde rangée présente la même image codée en adaptant l'index de quantification en fonction de l'activité spatiale des zones. Des zones à forte activité spatiale sont plus fortement codées, sans introduire de gêne particulière. L'image (b) présente un coût de codage inférieur de 40% à celui de l'image (a).

Principalement à cause des deux premières limitations, et en attendant la mise en

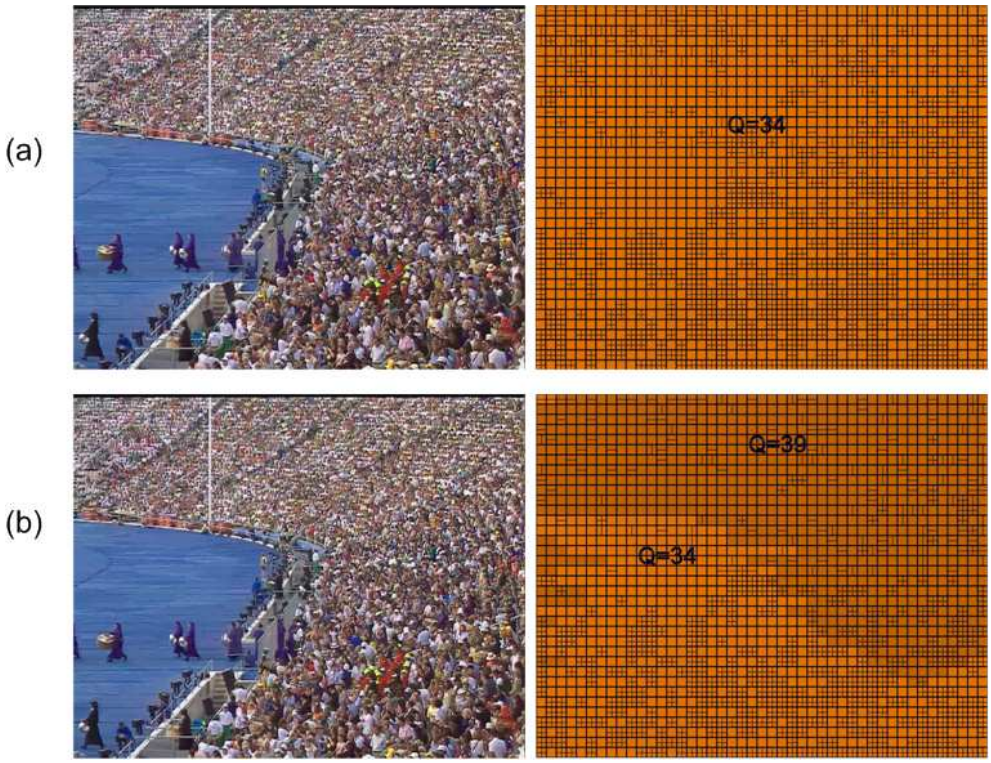

FIG. 1.12: Adaptation de l'index de quantification en fonction de l'activité spatiale : (a) codage classique avec un index de quantification de 34 ; (b) codage adapté conduisant à une réduction du coût de codage de 40%.

place d'une régulation de type ρ-domaine, les investigations dans ce domaine n'ont pas été approfondies.

1.8 Conclusion

L'objectif de ce chapitre était de décliner un cadre général de compression vidéo utilisant une carte de saillance. Après avoir défini la compression sélective, nous avons vérifié l'invariance de la stratégie visuelle d'observateurs en présence d'artefacts de codage. Cette vérification a été menée sur images fixes. L'extension à la vidéo n'a malheureusement pas été réalisée. Néanmoins, en faisant l'hypothèse que le précédent résultat se vérifierait également pour de la vidéo, la carte de saillance, moyennant quelques modifications, a été couplée de deux façons différentes à un schéma de codage.
La première méthode concerne la compression sélective indirecte qui consiste à modifier la source à coder. L'objectif est ici de réduire le débit de codage tout en conservant une qualité très acceptable sur les régions visuellement importantes. Le pré-traitement mis en oeuvre est donc piloté par une carte de saillance. Avec un filtre de pré-traitement morphologique, nous avons montré que le débit pouvait être réduit jusqu'à 39%. Evidemment, la dégradation des zones inintéressante est très forte et ne peut convenir que pour certaines applications du type visiophonie et télé-surveillance.

La deuxième méthode concerne la compression sélective directe. Pour ce type de compression, c'est le coeur et la stratégie de codage qui sont modifiés. L'objectif est d'améliorer la qualité perçue comparativement à une approche classique de codage. L'approche envisagée permet d'améliorer significativement la qualité en terme de *PSNR* des régions d'intérêt (+2 dB en moyenne). Des tests subjectifs devront être effectués afin de confirmer l'intérêt ou non d'une compression sélective directe.
La compression sélective psycho-visuellement adaptée a longtemps été délaissée se résumant à sa plus simple expression. Ce manque d'intérêt s'explique essentiellement par le manque de méthodes permettant de définir correctement et automatiquement les régions à privilégier. Les nouveaux travaux de modélisation de l'attention visuelle, que ce soit les nôtres, ceux de L. Itti, D. Parkhurst ou encore A. Chauvin, permettent d'entrevoir la définition d'applications de compression plus performantes. Nous contribuons à la définition de ces nouvelles approches en déclinant deux méthodes de compression sélective. Ajoutons pour finir, qu'il peut également être intéressant de coupler compression sélective indirecte et directe afin de tirer profit des avantages de chacune des deux méthodes.

Chapitre 2

Construction d'images miniatures

2.1 Introduction

Ce chapitre décrit une application récente et prometteuse basée sur la saillance visuelle. Cette application concerne la construction d'images et de séquences d'images miniatures facilitant à la fois la recherche de contenu et améliorant la satisfaction visuelle des utilisateurs. A première vue, cette application peut être considérée comme marginale. Mais, face à l'explosion des contenus multimédia et aux offres pléthoriques de création de contenu, un outil d'aide à la recherche de contenu et permettant d'augmenter la satisfaction visuelle des observateurs va très vite s'avérer nécessaire. La recherche de contenu peut être facilitée si on ne considère, pour effectuer la recherche, que les zones les plus intéressantes, les plus saillantes de l'image. Par ailleurs, le confort visuel peut être amélioré lorsque un utilisateur travaille sur un écran de taille réduite (téléphone portable, PC de poche...). Dans ce contexte, il est intéressant d'afficher seulement les parties d'intérêt d'une image ou d'une séquence d'images.
Des travaux récents ont déjà été réalisé pour les objectifs précédemment évoqués. Les principaux sont décrits dans la première partie. La seconde partie est consacrée à la description et à l'évaluation de l'approche développée pour des images fixes. L'extension à des séquences d'images est détaillée dans la troisième partie. Un bilan et des pistes d'améliorations sont donnés en conclusion.

2.2 Des travaux récents

Les premières travaux dans ce domaine sont ceux de X. Fan et al. [Fan 03]. L'objectif était de faciliter l'exploration d'une base d'images affichées sur l'écran d'un téléphone portable. Les zones d'intérêt sont déterminées à partir d'une carte de saillance. Cette dernière, issue du modèle de L. Itti, est couplée avec des informations de haut niveaux telles que la détection de texte et la détection de visage. La zone la plus saillante est alors extraite de l'image. Une amélioration de ce procédé a été très vite proposée par H. Liu [Liu 03]. Elle consiste à prendre en compte le fait qu'une image peut présenter plusieurs zones d'intérêt spatialement éloignées. Dans un tel contexte, lorsqu'un observateur regarde l'image, l'oeil effectue des séquences de saccades et de fixations entre les différentes régions. Pour reproduire ce phénomènes, H. Liu et al. [Liu 03] proposent d'afficher séquentiellement chaque

zone d'intérêt en fonction de leur saillance. Un paramètre, appelé MPT pour *Minimal Perceptible Time*, est intégré. La durée d'affichage est fonction de la saillance et des préférences de l'utilisateur.
L'extension de ces techniques à la vidéo a été peu abordée. L'une des premières, et des seules à notre connaissance, est proposée par J. Wang et al. [Wang 04]. Le principe consiste à construire une séquence de zones d'intérêt. Cette séquence est constituée des sous-séquences suivantes :

- la séquence originale avec une cadence image réduite,
- une séquence présentant un zoom avant sur la région d'intérêt,
- une séquence de de fixation sur cette région d'intérêt,
- une séquence présentant un zoom arrière pour revenir à la séquence originale avec une cadence image réduite.

Les deux phases de zoom permettent de lisser la transition entre la scène originale et la zone d'intérêt. Par ailleurs, lors de la séquence de fixations, une version fortement sous-échantillonnée de la séquence originale apparaît en bas à droite de l'image. Comme les auteurs le signalent, le problème est simplifié en ne considérant qu'une seule région d'intérêt. L'application visée est ici la vidéo surveillance.
Ces travaux, bien que très séduisants sur certains aspects, ne mettent pas assez l'accent sur la nécessité d'adapter la taille de la miniature en fonction de la distribution de la saillance de l'image. Pour pallier ce problème, nous proposons une solution entièrement automatique permettant d'adapter la taille de la miniature en fonction notamment de la distribution de la saillance.

2.3 Images miniatures centrées sur les zones visuellement intéressantes

La création d'images miniatures doit répondre à plusieurs exigences. La plus importante, à notre avis, concerne la taille de la miniature. Doit-elle être fixe et centrée sur le maximum global ou adpaté en fonction de certains paramètres telle que la distribution de la saillance ? Si on considère une image ayant un taux de couverture faible, la taille de la miniature peut être petite favorisant un effet de zoom sur la zone d'intérêt. Par contre, si le taux de couverture est important, la miniature doit avoir une taille raisonnable pour englober la majorité des zones saillantes. Nous proposons de répondre à ce problème.

2.3.1 Sélection des sites les plus saillants

A partir de la densité de saillance, un algorithme de type *Winner-Take-All* est utilisé pour déterminer les maximums locaux. Le maximum local $n+1$ est déterminé en inhibant une zone circulaire $\mathcal{C}$ centrée autour du maximum local n et d'un rayon de un degré visuel. Chaque zone circulaire représente donc une partie intéressante de l'image. Cette zone circulaire, centrée sur un maximum, est repositionnée sur son centre d'inertie *ci* avant

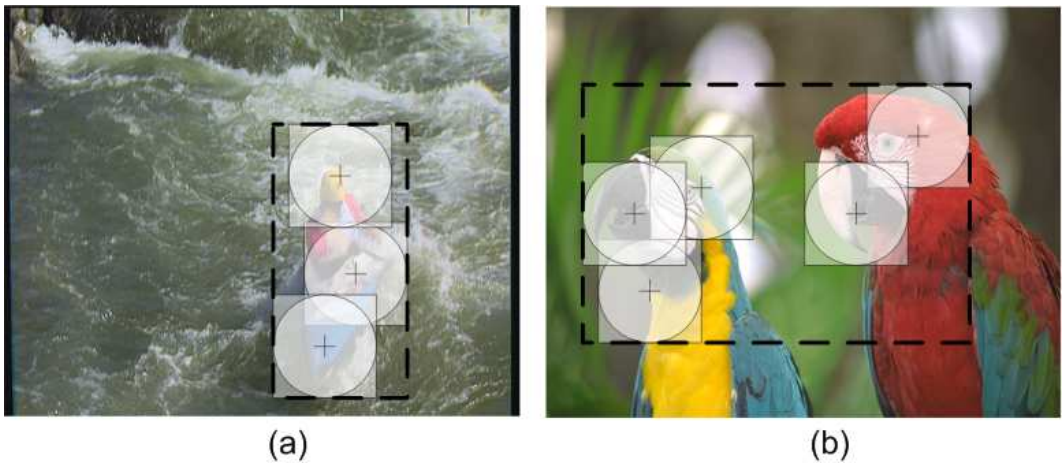

FIG. 2.1: Exemple de miniatures centrées sur les zone saillantes : (a) image *Kayak* avec trois maximums locaux et (b) image *Parrots* avec cinq maximum locaux. L'image miniature résultante est représentée par la rectangle en pointillé noir.

d'être inhibée. Ce dernier est calculé de la façon suivante :

$$ci_x = \frac{1}{\sum_{(x,y)\in\mathcal{C}} DS^{SP}(x,y)} \sum_{(x,y)\in\mathcal{C}} DS^{SP}(x,y) \times x \tag{2.1}$$

$$ci_y = \frac{1}{\sum_{(x,y)\in\mathcal{C}} DS^{SP}(x,y)} \sum_{(x,y)\in\mathcal{C}} DS^{SP}(x,y) \times y \tag{2.2}$$

avec,
ci_x et ci_y les coordonnées du centre d'inertie ci. DS^{SP} représente la densité de saillance spatiale.
La recherche des maximums locaux s'arrête lorsque la saillance inhibée contenu dans les cercles recentrés représente plus de $P\%$ de la saillance totale de l'image. La valeur de $P\%$ est importante car c'est elle qui conditionne la taille finale de la miniature. Cette valeur est adaptée en fonction du taux de couverture de l'image traitée. Le taux de couverture qualifie la distribution de la saillance ; une valeur faible tend à montrer qu'il existe une et une seule régions d'intérêt alors qu'une valeur forte signifie que l'image contient aucune zone saillante.

2.3.2 Construction de l'image miniature

A partir des coordonnées des coins des carrés dans lesquels sont inscrits les cercles centrés sur un maximum local, une miniature est déterminée. Les coordonnées des coins en haut à gauche et en bas à droite de l'image miniature sont déduites afin d'inclure toutes les zones carrés centrées sur des maximums locaux pertinents. La figure 2.1 présente deux situations différentes. La taille de la miniature dépend donc du nombre de maximums choisi mais également de leur position spatiale relative. La première situation concerne l'image *Kayak*. Cette image contient une et une seule zone saillante. Les différents maximums locaux pertinents se situent tous autour de cette zone saillante. Dans ce cas, la pertinence

TAB. 2.1: Evaluation subjective de la pertinence de l'algorithme proposé et de l'algorithme de type *Windows* : test 1 et test 2

Test 1 miniatures de type *Windows*		Test 2 méthode proposée	
Exploitable	Inexploitable	Exploitable	Inexploitable
31 (62%)	19	34 (67%)	16

visuelle de la miniature est très élevée. La seconde situation concerne l'image *Parrots* pour laquelle deux zones saillantes apparaissent clairement. L'image miniature obtenue présente une pertinence moyenne car elle intègre de nombreuses zones visuellement moins ou peu intéressantes.

2.3.3 Évaluation qualitative

A partir d'une base de 50 images, présentant des contenus variés, nous avons mené une évaluation qualitative. On se place dans un contexte de recherche de contenus. La figure 2.2 présente une partie des miniatures utilisées pour réaliser l'étude qualitative.

Trois tests ont été effectués :

1. Test 1 : les miniatures obtenues via une approche classique de décimation sont montrées à différents observateurs. Ces derniers doivent dénombrer le nombre de miniatures exploitables rapidement ;
2. Test 2 : les miniatures obtenues via les cartes de saillance sont montrées à différents observateurs. Comme précédemment, les observateurs doivent dénombrer le nombre de miniatures exploitables rapidement ;
3. Test 3 : ce dernier test est un test comparatif. Il s'agit de comparer les miniatures du test 1 avec celles du test 2 afin de déterminer qu'elle est la miniature la plus pertinente.

Le tableau 2.1 donne les résultats des tests 1 et 2. Sur une base de 50 images et pour 30 observateurs, le nombre de miniatures exploitable rapidement provenant du test 2 est supérieur au nombre de miniatures provenant du test 1 : 67% des miniatures construites à partir de l'approche proposée sont jugées rapidement exploitables contre 64% des miniatures provenant de l'approche classique. La méthode proposée est donc légèrement plus performante que la méthode classique.

Le tableau 2.2 donne le résultat du test 3. On constate que les observateurs préfèrent les miniatures construites à partir des cartes de saillance : 29 miniatures sur 50 (58%) proviennent de la méthode proposée.

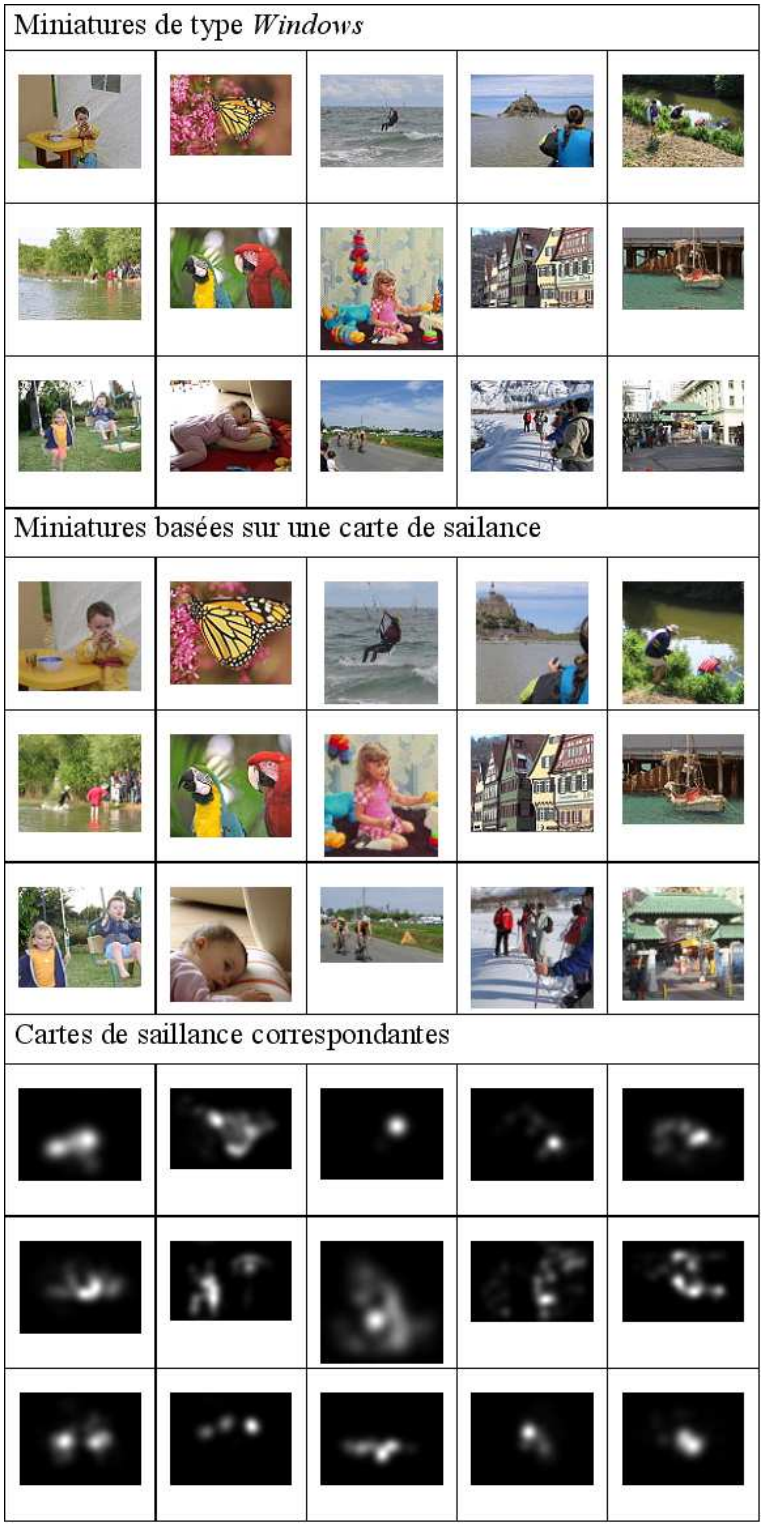

FIG. 2.2: Exemple d'images miniatures obtenues via une approche classique de décimation et via la méthode proposée. Les cartes de saillance associées aux images sont également données.

TAB. 2.2: Évaluation comparative des miniatures de l'algorithme proposé et de l'algorithme de type *Windows* : test 3

Test 3	
Nombre de miniatures préférées provenant de l'algorithme de type *Windows*	Nombre de miniatures préférées provenant de l'algorithme proposé
21 (42%)	29 (58%)

2.4 Séquences d'images miniatures centrées sur les zones visuellement intéressantes

La création de séquences vidéo miniatures, centrées sur les zones les plus saillantes, peut être intéressante pour des applications bien particulières telles que la recherche de contenu vidéo et la création de mosaïque de séquences vidéo. L'application la plus intéressante concerne certainement l'affichage de séquences vidéo sur des écrans de taille très limitée tels que les téléphones portables.
Alors que la création d'images miniatures est assez directe, la création de séquences d'images est plus complexe.
Bien que le principe général de construction d'une séquence d'images miniatures soit le même que celui présenté auparavant, de nombreux problèmes apparaissent. Les deux principaux problèmes sont les suivants :

- d'une image à l'autre, les positions des maximums locaux de saillance peuvent être différentes. Si aucune précaution n'est prise, un phénomène de saccade ou de tremblement peut apparaître et peut être fort désagréable ;
- la distribution de la carte de saillance est susceptible d'être différente d'une image à l'autre. La taille de la miniature est donc adaptée d'une image à l'autre. Pour construire la séquence finale, toutes les miniatures doivent avoir la même résolution. Un remise à l'échelle de toutes les images est donc nécessaire. Cette dernière peut avoir des conséquences néfastes puisque des zooms avant et arrière peuvent apparaître d'une image à l'autre.

Une version simplifiée de construction de séquences d'images miniatures est présentée dans le paragraphe suivant. Elle utilise une approche qui résout le problème de légers tremblements.

2.4.1 Version simplifiée de construction de séquences d'images miniatures

Pour éviter d'avoir à remettre à l'échelle toutes les images, l'algorithme proposé consiste à fixer arbitrairement la résolution de chaque image de la séquence d'images miniatures. De ce fait, il reste simplement à déterminer la position de la miniature dans l'image source en fonction de la carte de saillance. Le centre de la fenêtre est donc positionné sur le maximum global de la carte de saillance. Une heuristique est ajoutée permettant de limiter les légers tremblements : si la position spatiale du maximum de saillance de l'image à l'instant t est

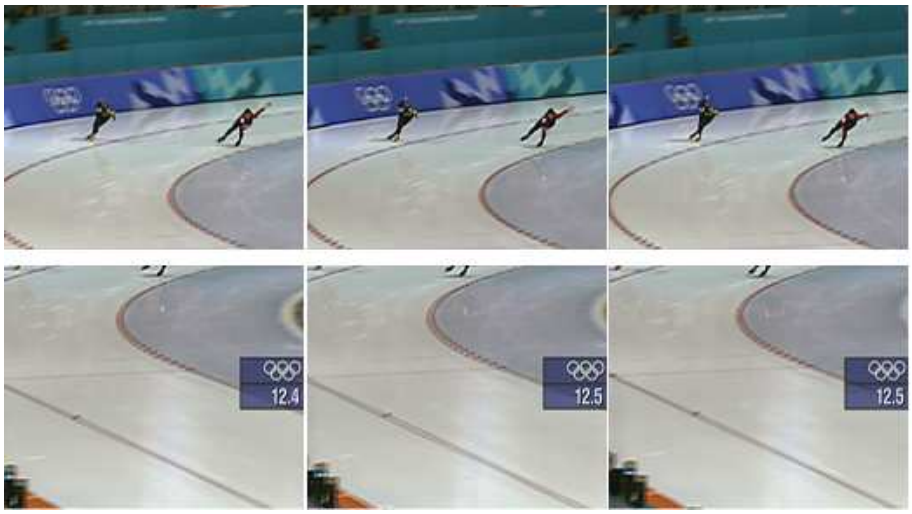

FIG. 2.3: Images miniatures de la séquence *PatinVitesse* obtenues avec l'algorithme développé.

proche (inférieur à un seuil arbitrairement fixé) de celui de l'image de l'instant $t-1$, la position du maximum de l'image $t-1$ est conservée. Cette façon de procéder ne permet pas d'éviter les fortes saccades, comme nous le verrons dans le paragraphe suivant. Notons pour finir qu'un recentrage de la fenêtre est effectué si la position du maximum est située à proximité d'un bord de l'image.

2.4.2 Évaluation qualitative de l'algorithme sur séquences d'images

Les figures 2.4 et 2.5 présentent des images miniatures de la séquence *Raid* provenant respectivement d'un algorithme de décimation classique et de l'algorithme proposé.
Il semble assez évident que la méthode proposée offre un confort visuel plus important. L'interprétation de la scène est rendue plus facile. La qualité de la séquence est meilleure car aucun sous-échantillonnage n'est utilisé. Néanmoins, un problème de saccade demeure : la miniature initialement centrée sur les deux coureurs est translatée intempestivement vers le haut de l'image.

Le même défaut, illustré à la figure 2.3 apparaît dans la séquence *PatinVitesse* mais il est du au fait qu'il existe trois zones d'intérêt : deux patineurs de vitesse sur glace et une incrustation dans laquelle se trouve un chronomètre. La qualité de la miniature est très moyenne puisque successivement l'incrustation et les patineurs sont choisis comme les zones les plus saillantes. Cette qualité est d'autant plus moyenne que le recentrage de la miniature n'est pas pertinent lorsque l'incrustation est choisie comme zone saillante.
Ces défaut nuisent considérablement à la qualité finale des résultats. En tout état de cause, l'algorithme de création de miniature se doit d'être plus robuste.

FIG. 2.4: Images miniatures de la séquence *Raid* obtenues après décimation.

FIG. 2.5: Images miniatures de la séquence *Raid* obtenues avec l'algorithme développé.

2.5 Conclusion

Ce chapitre a présenté une application de création d'images ou de séquences d'images miniatures basée sur la saillance visuelle. Confronté à l'explosion des données multimédia, il devient important d'avoir des outils aidant et facilitant la recherche de contenu tout en conservant les données les plus importantes.
Bien que les deux approches décrites auparavant soient simples, elles mettent clairement en exergue l'utilité de la saillance visuelle. Le confort visuel apporté par la création d'images miniatures basée sur la saillance visuelle est incontestable. Les résultats montrent que les observateurs préfèrent les miniatures centrées sur les zones saillantes aux miniatures classiques de type *Windows*. On reste toutefois confronté à deux problèmes : le premier concerne la modélisation de l'attention visuelle. Outre le fait qu'elle n'intègre pas des données sémantiques tels que le texte et les visages, le modèle est mis en défaut sur certaines images. Il est donc nécessaire d'analyser les problèmes inhérents à la modélisation. Par ailleurs, l'algorithme de construction des images miniatures n'est pas suffisamment robuste (détermination du seuil $P\%$). Des améliorations sont à apporter. La première consisterait à adapter la taille de la miniature en fonction de la taille finale de la miniature (96 par 96 pixels par défaut pour *Windows*).
Concernant la création de séquences d'images miniatures, le problème est encore plus délicat. Les problèmes de saccades et de tremblements sont à considérer. Les tremblements sont dus à une légère variations de la position du maximum de saillance. Ce problème est relativement simple à résoudre. Des saccades peuvent apparaître lorsqu'il y a plusieurs régions d'intérêt dans la séquence. Le maximum se déplace d'une région à une autre région, d'une image à l'autre. La résolution de ce problème est délicat et la solution est certainement liée à l'application visée. En dépit de cela, les premiers résultats sont encourageants.

Conclusion générale et perspectives

Nos travaux de cette thèse portaient sur la modélisation de l'attention visuelle pré-attentive. Nous avons développé un modèle fonctionnant sur les images fixes couleurs et sur les séquences d'images. Un soin particulier a été apporté à l'évaluation de ses performances. Deux applications utilisant le modèle proposé sont décrites. Dans cette conclusion générale, nous présentons tout d'abord une synthèse de ces travaux. Ce bilan est suivi par plusieurs perspectives tant sur la modélisation, l'évaluation des performances que sur l'exploitation des cartes de saillance.

Synthèse des travaux effectués

Les principales contributions de nos travaux concernent les quatre points suivants :
- la **modélisation de l'attention visuelle sur images fixes** permettant de détecter les zones spatialement saillantes ;
- la **modélisation de l'attention visuelle sur séquences d'images** permettant de détecter les zones spatio-temporellement saillantes ;
- l'**évaluation des performances d'un modèle d'attention visuelle** aussi bien sur images fixes que sur séquences d'images ;
- l'**utilisation d'un modèle d'attention visuelle dans deux applications**.

Modélisation de l'attention visuelle sur images fixes

Afin de pallier les problèmes rencontrés par les modèles d'attention visuelle existants, nous avons défini un cadre original et cohérent de modélisation. Il s'agit de définir un espace psycho-visuel dans lequel les données achromatique et chromatiques extraites de l'image à analyser sont exprimées en fonction du seuil différentiel de visibilité. Cet espace présente deux caractéristiques importantes. La première est d'exprimer des données provenant de différentes dimensions visuelles avec la même unité (la visibilité). La seconde concerne la hiérarchisation implicite des données obtenues via le procédé de normalisation utilisé.
A partir des données de cet espace, nous avons proposé plusieurs transformations pour détecter les zones visuellement importantes. De ce point vue, il semble alors tout à fait pertinent et cohérent de considérer les données en fonction de leur degré de visibilité pour construire la saillance. Une donnée peu visible ne va pas générer de saillance alors qu'une donnée visible peut générer de la saillance.

A partir de ces transformations, trois cartes de saillance (une pour la composante achromatique et une pour chaque composante chromatique) sont obtenues. Nous avons proposé une méthode de fusion permettant de construire une carte de saillance finale. Comparativement à une fusion classique, le procédé proposé améliore substantiellement les performances de la modélisation.

Modélisation de l'attention visuelle sur séquences d'images

Nous avons proposé une extension du modèle à la dimension temporelle. Les zones temporellement saillantes sont celles qui présentent un contraste de mouvement. Pour déterminer ce contraste de mouvement, nous utilisons un estimateur de mouvement local et un estimateur de mouvement global. La différence entre le mouvement local et le mouvement global permet d'obtenir un indicateur pertinent de la valeur de la saillance temporelle. Idéalement, cette différence de mouvement représente le mouvement réel de la scène filmée par une caméra fixe.
Une carte de saillance spatio-temporelle est alors déterminée en fusionnant la carte de saillance spatiale précédemment calculée et la carte de saillance temporelle. La même méthode de fusion, défini lors de la conception de la modélisation de l'attention visuelle sur images fixes, est de nouveau utilisé.

Evaluation des performances à partir d'expérimentations oculométriques

L'évaluation des performances d'un modèle d'attention visuelle n'est pas triviale. Tout d'abord, il a fallu construire une référence à partir d'expériences oculométriques menées sur images fixes et séquences d'images. Nous avons donc défini un protocole d'expérimentation ainsi qu'une façon de construire la vérité terrain.
Concernant l'évaluation des performances globales du modèle sur images fixes, nous avons utilisé trois métriques : le coefficient de corrélation linéaire, la divergence de Kullback-Leibler et une méthode de classification. Comparativement à un modèle de l'état de l'art, les trois métriques donnent la même tendance, positionnant avantageusement nos travaux.
L'un des aspects intéressants de cette évaluation concerne la durée d'observation. Nous montrons que, même en considérant une durée d'observation importante (14s), les performances du modèle sont correctes. Cela signifie donc que les mécanismes inhérents à l'attention de type *Bottom-Up* persistent dans le temps et ne disparaissent pas au profit d'un mécanisme cognitif.
Concernant l'évaluation des performances du modèle spatio-temporel de saillance, deux métriques (une fonction de probabilité cumulée et une méthode de classification) sont utilisées. Les résultats donnés par ces métriques confirment l'évaluation qualitative des résultats. La modélisation permet de détecter plus de 75% des zones saillantes.

Applications des cartes de saillance

Une méthode automatique de détection de zones saillantes peut permettre d'accroître les performances de nombreuses applications. Deux de ces applications ont été décrites.
La première concerne le codage d'une séquence d'images pour laquelle la qualité des zones attirant le regard est favorisée. L'objectif est d'améliorer la qualité perçue par des observateurs. Une méthode de compression sélective a donc été développée sur la base de la

norme de codage H.264/AVC. Elle consiste à adapter en fonction de la carte de saillance la consigne de quantification au niveau macrobloc. Les résultats sont prometteurs car la qualité visuelle de certaines séquences est indéniablement accrue. Toutefois, la compression sélective uniquement basée sur une carte de saillance n'est pas suffisante puisque la dégradation des zones inintéressantes peut engendrer l'apparition d'artefacts et donc la création de nouveaux points de fixation.
La seconde application concerne la création d'images miniatures (*Thumbnails* en anglais) ou de séquences d'images miniatures. L'objectif de cette application est d'augmenter le confort visuel lors d'une recherche de contenu. L'idée est simple puisqu'il s'agit de centrer l'image miniature sur la zone la plus attractive perceptuellement. Nous montrons que l'application simple de création d'images miniatures est qualitativement préférée à celle utilisant un sous-échantillonnage classique. Nous avons par ailleurs proposé les bases d'un algorithme de création de séquences d'images miniatures.

Perspectives

Les perspectives que nous envisageons dans le prolongement de ces travaux de thèse sont au nombre de trois. Tout d'abord, il est intéressant de poursuivre des travaux dans le domaine de la modélisation de l'attention visuelle. La seconde concerne l'évaluation des performances d'un modèle d'attention visuelle. La dernière perspective concerne la multitude d'applications pouvant bénéficier de ces travaux.

Modélisation de l'attention visuelle

La modélisation de l'attention visuelle est un domaine en plein essor. Nous y avons contribué mais les voies à explorer restent nombreuses pour améliorer le modèle :

- la première concerne la mise en place d'expérimentations psychophysiques nécessaires pour optimiser de nombreux paramètres du modèle. Ce type d'expériences permettrait également d'affiner la technique de fusion que nous avons mis en place ;
- la seconde concerne l'utilisation d'autres informations de bas niveau (texture, netteté d'un contour...) ;
- la troisième voie concerne les informations de haut niveau. Il va de soit que coupler le modèle avec des informations visuelles de haut niveau (visage, teinte chaire, texte...) peut améliorer les performances de la modélisation. A ces informations visuelles, on peut ajouter les informations auditives notamment en étudiant la direction du son. Deux personnages saillants discutant peuvent être alternativement plus ou moins saillants lorsqu'il parle ou non. Par ailleurs, la prise en compte du contexte via une reconnaissance de scènes permettant de détecter les objets incongrus dans la scène peut être intéressant. Ce sont typiquement les travaux d'A. Torralba [Torralba 03].

Les performances du modèle de type *Bottom-Up* que nous proposons peuvent radicalement être modifiées dans de nombreuses configurations si le travail d'optimisation des paramètres via des expériences psychophysiques ainsi que le couplage avec des informations de haut niveau est réalisé.

Evaluation des performances

L'évaluation des performances est un véritable problème. Comme actuellement il n'y a aucun consensus, chacun utilise une méthode particulière pour estimer les performances. Nous proposons d'utiliser plusieurs métriques complémentaires pour rendre "robuste" l'évaluation. Outre ce premier problème, un second plus ennuyeux concerne à la fois le protocole expérimental utilisé pour les tests oculométriques et la façon de créer la carte de saillance des observateurs.
Sur ces deux points, un travail fédérateur donnant les bases de l'évaluation des performances, à l'instar de ce qu'il se fait en estimation de la qualité d'images [VQEG], est nécessaire.

Applications des cartes de saillance

Le développement d'outils automatiques de détection de zones saillantes représente un enjeux important pour de nombreuses applications. Deux applications ont été présentées dans cette thèse. Il en existe beaucoup d'autres :

- le tatouage numérique : les cartes de saillance peuvent piloter un procédé de tatouage numérique en indiquant à quel endroit il est judicieux de placer une information. On peut alors distinguer le tatouage invisible du tatouage visible. Pour le premier, l'idée est d'intégrer la marque dans des zones inintéressantes visuellement. Les applications visées sont l'identification du propriétaire du média tatoué, du contenu, contrôle de copie... Le second type de tatouage est de placer une marque visible sur les zones saillantes dans le but de rendre l'image ou la séquence d'images inexploitable ;
- l'estimation de qualité : actuellement, l'estimation de qualité d'images se fait en attribuant le même poids à toutes les erreurs. Cependant, il est clair qu'un artefact sur une zone saillante est plus gênant qu'un artefact apparaîssant sur une zone inintéressante. Des travaux ont démarré sur ce sujet mais actuellement l'introduction de la saillance n'a pas encore permis d'améliorer les performances ;
- la recherche de contenu : l'idée est d'accélérer la recherche d'un contenu en utilisant une imagette représentative des zones saillantes.

D'autres applications existent mais les applications que nous avons citées c'est à dire codage, image miniature, tatouage, qualité et recherche de contenu, forment à notre avis les 5 plus importants domaines d'applications.

Quatrième partie

Annexes

Annexe A

Résultats des tests oculométriques sur images fixes

A.1 Images de test

La figure A.1 présente les images de tests utilisées pour les tests oculométriques.

FIG. A.1: Images originales utilisées pour les tests. Première ligne : *Bikes*, *Paintedhouse*, *Zebre*, *Lighthouse*, *Dancers* ; teconde ligne : *Manfishing*, *Parrots*, *Plane*, *Rapids*, *Sailing1* ; troisième ligne : *Vautours629Couleur*, *Vautours538Couleur*, *Vautours825Couleur*, *Kayak-Couleur*, *Ocean* ; dernière ligne : *PatinCouleur*, *ChurchAndCapitol*, *Stream*.

A.2 Résultats sur images fixes

Pour chaque durée d'observation (4, 10 et 14 secondes) et pour chaque image, les fixations superposées sur l'image originale et la carte de saillance de l'observateur moyen sont données.

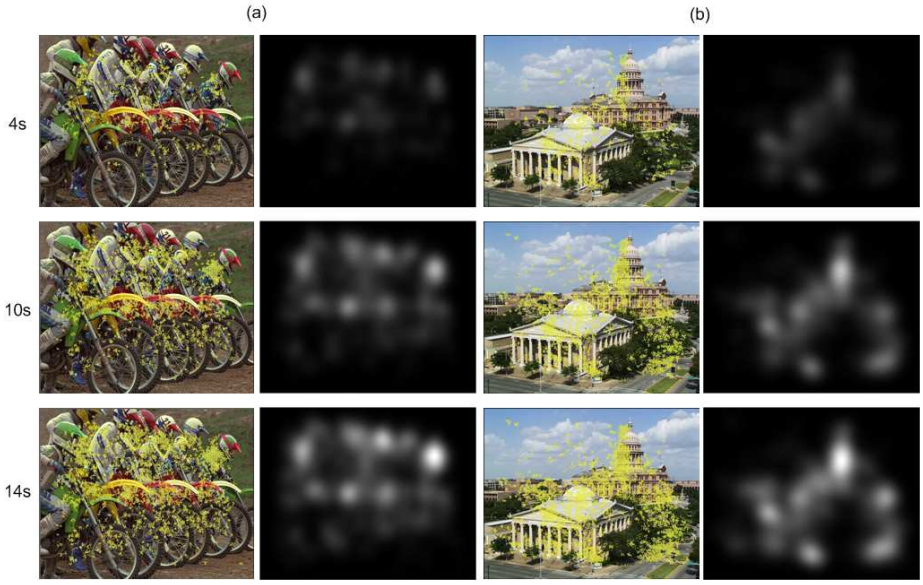

FIG. A.2: Résultats oculométriques pour les images *Bikes* (a) et *Churchandcapitol*(b).

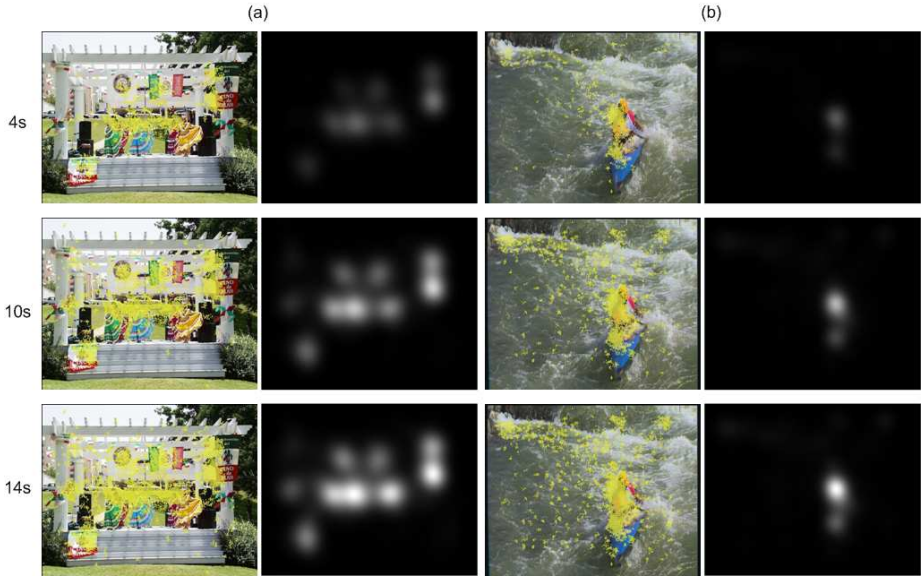

FIG. A.3: Résultats oculométriques pour les images *Dancers* (a) et *KayakCouleur*(b).

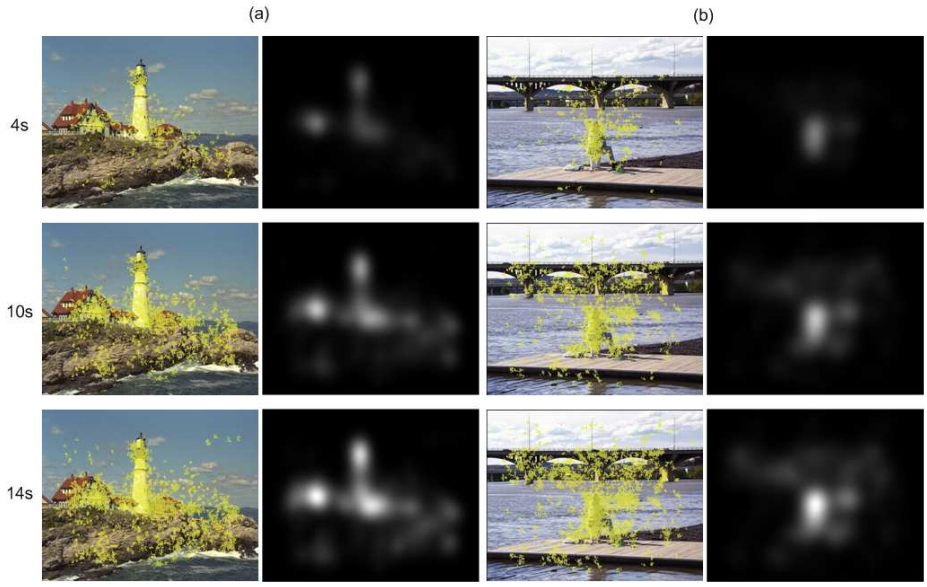

FIG. A.4: Résultats oculométriques pour les images *Lighthouse* (a) et *Manfishing*(b).

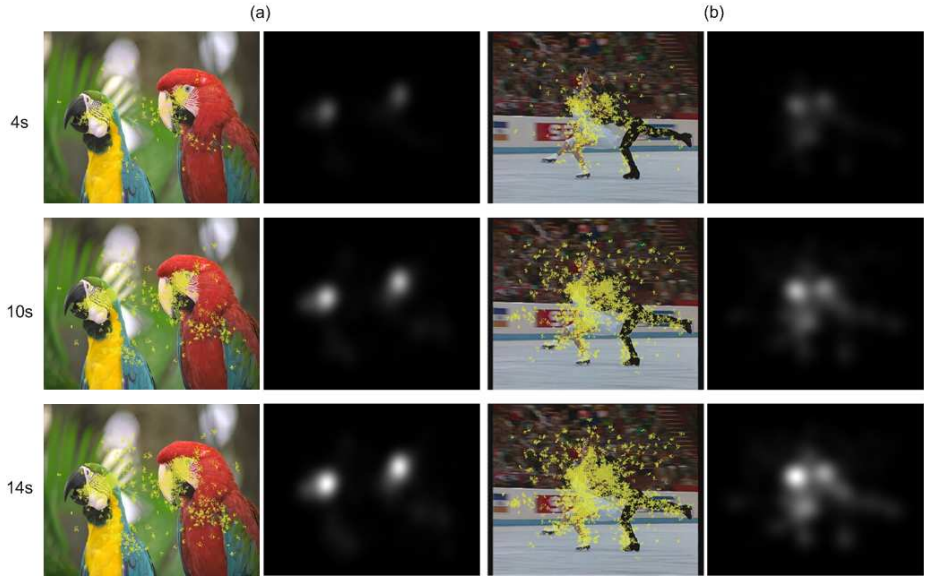

FIG. A.5: Résultats oculométriques pour les images *Parrots* (a) et *PatinCouleur*(b).

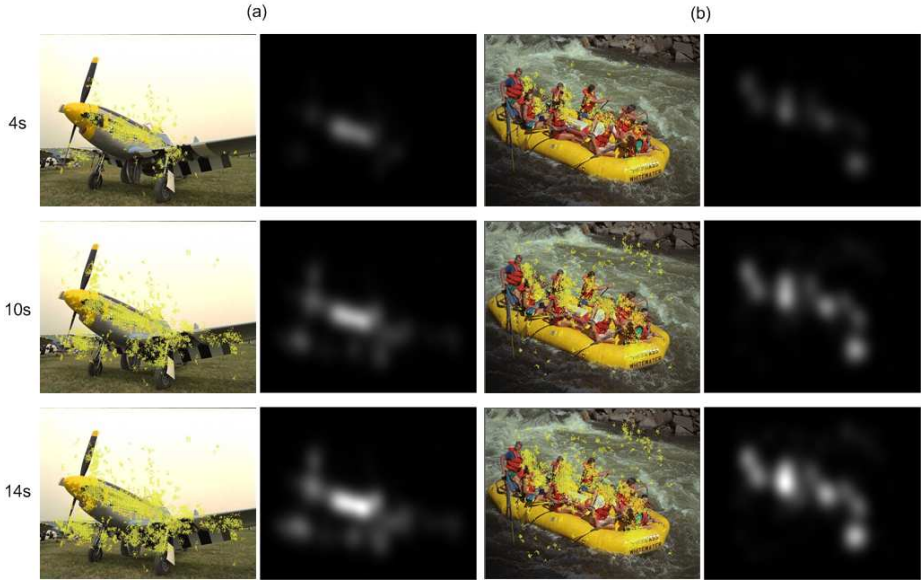

FIG. A.6: Résultats oculométriques pour les images *Plane* (a) et *Rapids*(b).

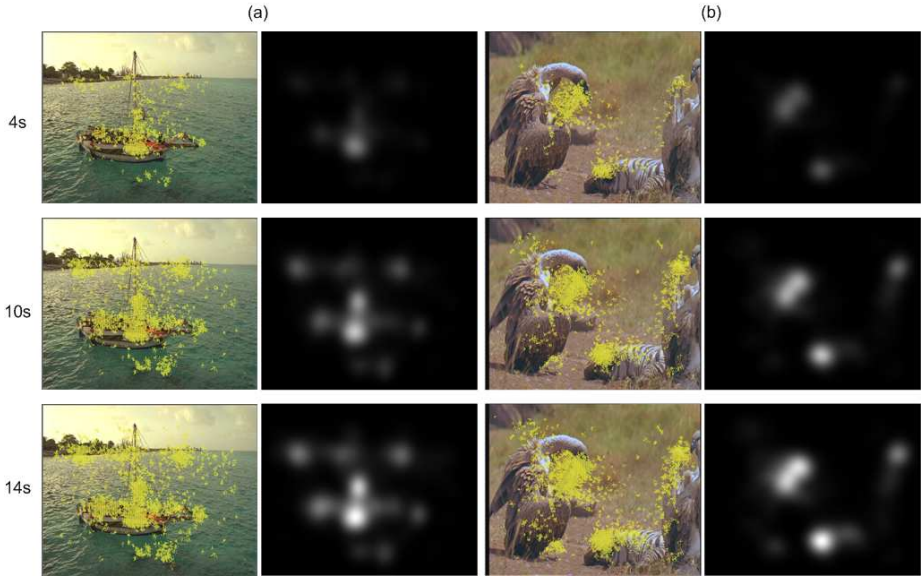

FIG. A.7: Résultats oculométriques pour les images *Sailing* (a) et *Vautour825Couleur*(b).

Annexe B

Paramètres des modèles de masquage visuel

TAB. B.1: Valeurs des paramètres du modèle de masquage intra canal pour chaque sous bande de chaque composante Cr

	Cr_1		Cr_2	
	SB I	SB II,n	SB I	SB II,n
a	0.45	0.48	0.72	0.13
b	0.06	0.03	0.22	0.05
c	1.22	0.82	2.78	0.42

TAB. B.2: fonctions de masquage prises en compte dans le modèle (le couple $[Y, Z]$ signifie que la sous bande Y peut influencer sensiblement le seuil de visibilité des éléments contenus dans la sous bande Z). L'exemple en gras corresponds à la sous bande II,n de la composante $Cr1$ peut être masquée par la sous bande I de la composante A.

	Composante masquée		
Masquant	A	$Cr1$	$Cr2$
A	masquage intra	$[I,I]$ $[I,IIn]$ $[IIn,I]$ $[IIn,IIn]$	négligée
$Cr1$	$[I,I]$ $[I,IIn]$ $[I,IIIn]$ $[IIn,I]$ $[IIn,IIn]$	masquage intra	$[I,I]$ $[I,IIn]$
$Cr2$	négligée	$[I,I]$ $[I,IIn]$	masquage intra

TAB. B.3: Valeurs des paramètres du modèle de masquage entre composantes Cr

	Cr_1 sur Cr_2		Cr_2 sur Cr_1	
	I sur I	I sur II	I sur I	I sur II
modèle	A	B	A	B
a	0.136	1.88	0.09	1.16
b	0.004	0.8	0.1	0.16
c	0.196	0.02	0.82	0.048

TAB. B.4: Valeurs des paramètres du modèle de masquage de la composante achromatique par la composante Cr_1

	I sur I	I sur II,n	I sur III,n	II,n sur I	II,n sur II,n	II sur III,n
modèle	A	A	A	B	B	B
a	1.89	0.089	0.141	1.7	2	0.019
b	0.03	0.0005	0.002	0.7	1	0.0008
c	2	0.1	0.238	0.22	0.0016	0.064

Annexe C

La norme de compression vidéo H.264/AVC

C.1 Introduction

La norme H.264/AVC (Advanced Video Coding) a été finalisée en 2003 par le groupe JVT (*Joint Video Team*), regroupant des experts de l'ITU (*International Telecommunication Union*) et de MPEG (*Moving Picture Expert Group*). Cette norme permet de coder de la vidéo au format 4 :2 :0, 4 :2 :2 et 4 :4 :4. L'objectif de cette normalisation est de définir un système de codage vidéo présentant des performances accrues comparativement aux normes en vigueur, et plus particulièrement MPEG-2. Un rapport deux en terme de débit distortion est espéré. Ses grandes caractéristiques sont rappelées dans les paragraphes suivants.

C.2 Les grandes caractéristiques de l'algorithme de codage H.264/AVC

La figure C.1 présente le schéma de codage H.264/AVC. En bleu, apparaissent les éléments non normatifs. Ces éléments sont à considérer comme des degrés de libertés sur lesquels on peut agir pour améliorer l'efficacité de compression. Comme on peut le constater, l'algorithme de base est classique puisque c'est une technique hybride de prédiction inter image, exploitant la redondance statistiques temporelles, et de codage par transformée de l'erreur de prédiction, exploitant la redondance spatiale. L'erreur de prédiction est la différence entre le bloc original et le bloc prédit. Les nouveautés concernent en fait l'intégration de nombreux raffinements qui sont brièvement décrits dans la suite. Pour de plus amples informations, le lecteur pourra consulter le livre de I. Richardson [Richardson 03].

C.2.1 Division en macrobloc (MB) et type de codage

Comme pour la précédente norme de compression vidéo MPEG-2, l'image est d'abord partitionnée en macroblocs de dimension fixe couvrant une zone rectangulaire de 16×16

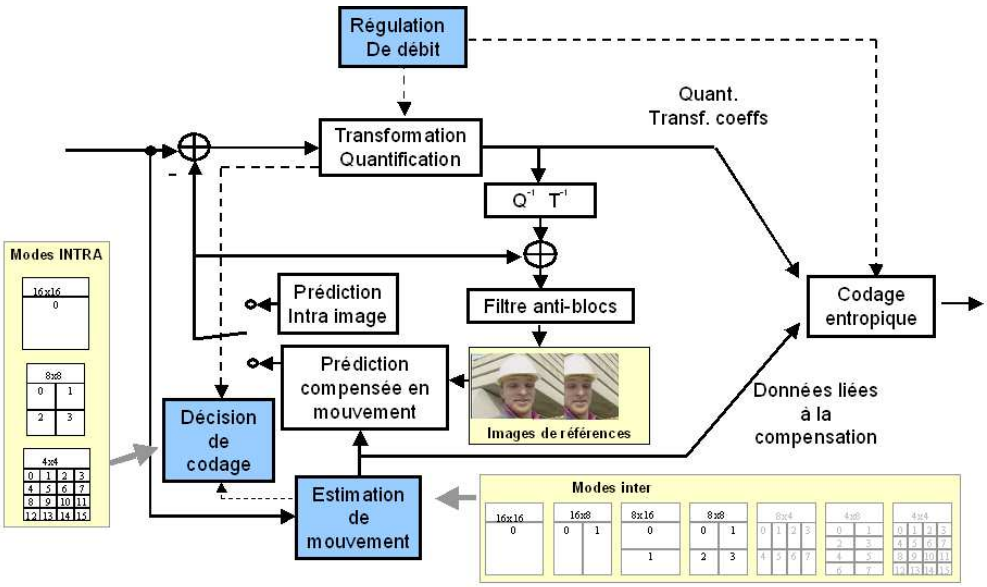

FIG. C.1: Synoptique d'un codeur H.264/AVC (en bleu, les parties non normatives).

échantillons de luminance et de 8×8 échantillons pour chaque composantes de chrominance.
Ils existent trois types de codage possibles. Le premier est le codage Intra , c'est à dire que le macrobloc est codé sans référence à d'autres images. Le deuxième est le codage prédictif (à compensation de mouvement), plus communément appelé inter ou P. Ce type de codage fait appel à une image de référence afin de déterminer une prédiction. Enfin, le dernier type est le codage bi-prédiction, plus communément appelé B. Ce codage fait appel à deux images de références pour déterminer la prédiction. L'erreur résiduelle de prédiction obtenue avec un codage de type P ou B est ensuite codée.

C.2.2 Prédiction intra image

L'une des premières différences avec la norme MPEG-2 se situe dans la façon de déterminer la prédiction intra. Alors que cette prédiction se faisait dans le domaine transformée pour MPEG-2, la prédiction intra image de la norme H.264/AVC s'effectue dans le domaine spatial. Elle se détermine à partir des échantillons des blocs voisins causaux (déjà codés). Lorsque le codage intra se fait sur des blocs de taille 4×4, préconisé pour coder des contours, il a neuf modes de prédictions possibles. Outre la prédiction *DC* (valeur moyenne des pixels de prédiction), huit modes de prédiction directionnelle sont spécifiés. Pour le mode de codage 16×16, très efficace pour coder des zones d'une image à faible activité, quatres prédictions sont possibles (mode vertical, horizontal, DC et plane (somme pondérée des prédictions)).

C.2.3 Prédiction compensée en mouvement

C.2.3.1 Codage des macroblocs d'une image P

Chaque macrobloc de type P est partitionné en 16×16, 16×8, 8×16 et 8×8 (les blocs 8×8 pouvant être sous partitionnés en 8×4, 4×8 et 4×4). Par ailleurs, un macrobloc de type P peut également être codé en intra.
La prédiction de chaque bloc de luminance $m \times n$ est obtenue en recherchant dans l'image de référence le bloc minimisant un certain critère distortion, débit ou plus généralement débit-distortion. La norme H.264/AVC permet d'utiliser plus d'une image préalablement codée comme image de référence. En d'autres termes, chaque bloc du macrobloc peut pointer sur une image de référence particulière.
Enfin, un macrobloc de type P peut être codé selon le mode SKIP (issu du mode inter image 16×16). Lorsque ce mode est choisi, ni l'erreur de prédiction quantifiée ni le le vecteur de mouvement sont transmis.

C.2.3.2 Codage des macroblocs d'une image B

Le codage des macroblocs de type B est différent de celui des macroblocs de type P dans le sens ou l'erreur de prédiction est calculée à partir d'une somme pondérée de deux prédictions ; c'est un codage bi-prédit. Outre les modes de codage Intra et Inter 16×16, 16×8, 8×16 et 8×8, un mode direct est possible. Pour le mode de prédiction directe, aucune donnée relative au mouvement est transmise. Elle est déduite soit à partir du vecteur de mouvement du macrobloc colocalisé d'une image P (mode directe temporel) soit à partir des vecteurs de mouvement des macroblocs voisins (mode direct spatial). Enfin, lorsque l'erreur de prédiction liée au mode directeest nulle, le macrobloc est codé en mode SKIP.

C.2.4 Transformation et quantification

C.2.4.1 Transformation

Après le calcul de l'erreur de prédiction, une transformée est appliquée. Dans le cas de cette norme, la transformée est appliquée sur des blocs 4×4. Au lieu d'une transformée en cosinus discrète, classiquement utilisée dans les précédentes normes, une transformée en entiers séparables présentant des propriétés similaires est utilisée. En prédiction intra 16×16, une transformée de Hadamard 4×4 est également utilisée pour transformer les coefficients DC de chaque bloc 4×4.

C.2.4.2 Quantification

La norme H.264/AVC utilise une quantification scalaire. Il y a 52 index ou paramètres de quantification (QP). Le pas de quantification est donné par la relation suivante :

$$Q = 0.67 \times 2^{\frac{QP}{6}} \tag{C.1}$$

Les pas de quantification obtenus sont disposés de telle sorte qu'un incrément de l'index de quantification provoque l'augmentation du pas de quantification de 12.5%. L'augmenation

	Principales propriétés de H.264/AVC
Codage INTRA	▪ *Tailles de blocs : 16x16 et 4x4* ▪ *Modes de prédiction spatiale :* • *9 modes de prédiction pour les blocs 4x4* • *4 modes de prédiction pour les blocs 16x16*
Codage INTER	▪ *4 partitions possibles (16x16, 16x8, 8x16, 8x8) et 4 sous partitions (8x8, 8x4, 4x8, 4x4)* 16x16 16x8 8x16 8x8 8x8 8x4 4x8 4x4 ▪ *Résolution des vecteurs jusqu'au quart de pixel* ▪ *Plusieurs images de référence par partition* ▪ *Prédictions pondérées*
Transformée	▪ *Transformée entière sur les blocs 4x4*
Quantification	▪ *Quantification scalaire précise (52 index de quantification)*
Filtre anti-blocs	▪ *Filtre adapté de lissage d'effets de blocs*
Codage entropique	▪ *Codage CAVLC (Context Adaptive VLC)* ▪ *Codage CABAC (Context-Based Adaptive Binary Arithmetic Coder)*

FIG. C.2: Tableau récapitulatif des propriétés de H.264/AVC.

de l'index d'une valeur de 6 multiplie le pas de quantification par 2.
Après la quantification, les coefficients sont lus en zigzag et transmis au codage entropique.

C.2.5 Codage entropique

Deux méthodes de codage entropique sont disponibles : CAVLC (*Context-Adaptive Variable Length Coding*) et CABAC (*Context-Adaptive Binary Arithmetic Coding*). La technique CAVLC est une première amélioration du codage à longueur variable (VLC). Elle consiste à remplacer les tables VLC en fonction des éléments syntaxiques déjà transmis. Cette optimisation permet d'améliorer les performances. La seconde technique utilise le codage arithmétique ce qui permet d'assigner un nombre de bits non entier à chaque symbole. Par ailleurs, le CABAC s'adapte en fonction du contexte.

C.2.6 Filtre anti-blocs

Pour atténuer l'un des défauts des codeurs basés bloc, un filtre anti-blocs est utilisé dans la boucle de codage. La force du filtrage dépend de l'index de quantification, du mode de codage, des vecteurs de mouvement et de l'erreur de prédiction codée.

C.2.7 Tableau récapitulatif

Le tableau C.2 récapitule les grandes caractéristiques de la norme de codage vidéo H.264/AVC.

Bibliographie

[Ballard 91] D. Ballard. – Animate vision. *Artificial intelligence*, 86 :48–57, 1991.

[Barten 04] P.G. Barten. – Formula for the contrast sensitivity of the human eye. – *SPIE Human Vision and Electronic Imaging*, San Jose, CA, 2004.

[Bavelier 00] D. Bavelier, A. Tomann, C. Hutton, T. Mitchell, G. Liu, D. Corina, H. Neville. – Visual attention to the periphey is enhanced in congenitally deaf individuals. *Journal of Neuroscience*, 20(17) :1–6, 2000.

[Bedat 98] L. Bedat. – *Aspects psychovisuels de la perception des couleurs. Application au codage d'images couleurs fixes avec compression del'information.* – Université de Nantes, PhD. Thesis, IRESTE, 1998.

[Bodmann 80] H. Bodmann, P. Haubner, A. Marsden. – A unified relationship between brightness and luminance. – *CIE Proceedings Kyoto Session 1979*, pp. 99–102, Paris, 1980.

[Bonds 89] A.B. Bonds. – Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex. *Visual Neuroscience*, 2(1) :41–55, 1989.

[Bradley 03] A.P. Bradley. – Can region of interest coding improve overall perceived image quality ? – *Proceedings of APRS Workshop on Digital Image Computing*, 2003.

[Braun 90] J. Braun, D. Sagi. – Vision outside the focus of attention. *Perception and Psychophysics*, 48 :45–58, 1990.

[Braun 94] J. Braun. – Visual search among items of different salience : removal of visual attention mimics a lesion in extrastriate area v4. *Jour. Neurosci.*, 14 :554–567, 1994.

[Braun 98] J. Braun. – Divided attention : Narrowing the gap between brain and behavior. *The Attentive Brain*, éd. par Ed. Parasuraman, R., pp. 327–352. – Cambridge, MA, MIT Press, 1998.

[Bruce 03] N. Bruce, E. Jernigan. – Evolutionary design of context-free attentional operators. – *Proceedings ICIP-03 (IEEE International Conference on Image Processing)*, 2003.

[Burt 83] P.J. Burt, E.H. Adelson. – The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31 :532–540, 1983.

[Callet 01] P. Le Callet. – *Critères objectifs avec référence de qualité visuelle des images couleur.* – Université de Nantes, PhD. Thesis, Ecole Polytechnique de l'Université de Nantes, 2001.

[Canosa 03] R. Canosa. – *Seeing, sensing, and selection : modeling visual perception in complex environments.* – USA, PhD. Thesis, Rochester Institute of Technology, 2003.

[Chai 00] D. Chai, A. Bouzerdoum. – Coding videophone sequences at better perceptual quality by using face localization and bit redistribution. – *IEEE ISPACS 2000*, vol. 1, pp. 22–26, 2000.

[Chauvin 00] A. Chauvin, J. Hérault, C. Marendaz, C. Peyrin. – Natural scene perception : visual attractors and image processing. – *7th Neural Computation and Psychology Workshop*, 2000.

[Chauvin 02] A. 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.* – Grenoble, PhD. Thesis, Université Joseph Fourier, 2002.

[Chen 93] J.-J. Chen, H.M. Hang. – A transform video coder source model and its application. – *Proceedings ICIP-93 (IEEE International Conference on Image Processing)*, 1993.

[Csiszar 67] I. Csiszar. – Information-type measures of difference of probability distributions and indirect observations. *Stu. Sci. Math. Hungar.*, 2 :299–318, 1967.

[Daly 93] S. Daly. – The visible differences predictor : An algorithm for the assessment of image fidelity. *Digital Images and Human Vision*, chap. 14, pp. 179–206. – MIT Press, 1993.

[Daugman 80] J.G. Daugman. – Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, 20 :847–856, 1980.

[Dhavale 03] N. Dhavale, L. Itti. – Saliency-based multi-foveated mpeg compression. – *IEEE Seventh International Symposium on Signal Processing and its Applications*, 2003.

[Fabri 05] S.N. Fabri, A. M. Kondoz. – Perceptually-weighted error-resilient coding for mpeg-4. – *London Communications Symposium*, University College London, 2005.

[Fan 03] X. Fan, X. Xie, W.Y. Ma, H.J. Zhang, H.Q. Zhou. – Visual attention based image browsing on mobile devices. – *in Proc. of ICME 2003*, vol. 1, 2003.

[Faugeras 76] O. D. Faugeras. – *Digital color image processing and psychophysics within the framework of human visual system.* – PhD. Thesis, University of utah, 1976.

[Felleman 81] D. Felleman. – *A comparison of the receptive field properties of neurons in the middle temporal visual area (MT) and striate cortex of the owl monkey, Aotus trivirgatus.* – Nashville TN, PhD. Thesis, Vanderbilt University, 1981.

[Flanagan 90] P. Flanagan, P. Cavanagh, O. E. Favreau. – Independent orientation-selective mechanisms for cardinal directions of color space. *Vision Research*, 30(5) :769–778, 1990.

[Foley 94] John M. Foley, Geoffrey M. Boynton. – A new model of human luminance pattern vision mechanisms : Analysis of the effects of pattern orientation, spatial phase and temporal frequency. – *SPIE Human Vision and Electronic Imaging*, vol. 2054, pp. 32–42, 1994.

[Fredericksen 97] R.E. Fredericksen, R.F. Hess. – Temporal detection in human vision : dependence on stimulus energy. *J. Opt. Soc. Am. A.*, 14(10) :2557–2569, 1997.

[Fredericksen 98] R.E. Fredericksen, R.F. Hess. – Estimating multiple temporal mechanisms in human vision. *Vision Research*, 38(7) :1023–104, 1998.

[Frimout 93] E.D. Frimout, J. Biemond, R.L. Lagendijk. – Forward rate control for mpeg encoding. – *SPIE Visual Communication and Image Processing*, 1993.

[Gilbert 89] C. D. Gilbert, T. N. Wiessel. – Columnar specificity of intrinsic horizontal and cortico-cortical connections in cat visual cortex. *Jour. Neurosci.*, 9(7) :2432–2442, 1989.

[Girod 89] B. Girod. – The information theoretical significance of spatial and temporal masking in video signals. – *SPIE Human Vision and Electronic Imaging*, vol. 1077, pp. 178–187, 1989.

[Gomila 01] C. Gomila. – *Mise en correspondance de partitions en vue du suivi d'objets.* – Paris, PhD. Thesis, Ecole des Mines, 2001.

[Green 03] C. S. Green, D. Bavelier. – Action video game modifies visual selective attention. *Nature*, 4(23) :543–537, 2003.

[Guyader 03] N. Guyader. – *Scènes visuelles : catégorisation basée sur des modèles de perception.* – Grenoble, PhD. Thesis, Université Joseph Fourier, 2003.

[Hammett 92] S.T. Hammett, A.T. Smith. – Two temporal channel or three ? a re-evaluation. *Vision Research*, 32(2) :285–291, 1992.

[Hang 97] H.-M. Hang, J.-J. Chen. – Source model for transform video coder and its application. *IEEE Trans. on Circuit and System for Video Technology*, 7(2) :287–311, 1997.

[Hannuksela 04] M. M. Hannuksela, Y.K. Wang, M. Gabbouj. – Isolated regions in video coding. *IEEE Trans. on Multimedia*, 6(2) :259–267, 2004.

[Hansen 02] T. Hansen. – *A neural model of early vision : contrast, contours, corners and surface.* – PhD. Thesis, University of ULM, 2002.

[He 01a] Z. He, Y. K. Kim, S.K. Mitra. – A novel source modeling framework for bit rate and picture quality control in dct visual coding. – *Picture Coding Symposium*, 2001.

[He 01b] Z. He, S.K. Mitra. – A unified rate-distortion analysis framework for transform coding. *IEEE Trans. on Circuit and System for Video Technology*, 11(12), 2001.

[Heeger 93] D.J. Heeger. – Modelling simple-cell direction selectivity with normalised, half-squared, linear operators. *Journal of Neurophysiology*, 70(5), 1993.

[Henderson 99a] J.M. Henderson, A. Hollingworth. – High-level scene perception. *Annu. Rev. Psychol.*, 50 :243–271, 1999.

[Henderson 99b] J.M. Henderson, P.A. Weeks, A. Hollingworth. – Effects of semantic consistency on eye movements during scene viewing. *Jour. Exp. Psychol. Hum. Percept. Perf.*, 25 :210, 1999.

[Hillstrom 94] A.P. Hillstrom, S. Yantis. – Visual motion and attentional capture. *Perception Psychophysic*, 55 :399–411, 1994.

[Hérault 01] J. Hérault. – De la rétine biologique aux circuits neuromorphiques. *Les systèmes de vision.* – Edition Hermès, J.M Jolion., 2001.

[Huguenel 05] L. Huguenel. – Codage par zones d'intérêt dans le cadre d'un codeur mpeg4 avc, 2005. Diplôme de Recherche Technologique.

[Hurst 04] C. Hurst. – The hurst model of vision balances : a new approach, 2004. Optometry Today.

[Itti 98] L. Itti, C. Koch, E. Niebur. – Model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(11) :1254–1259, 1998.

[Itti 00a] L. Itti, C. Koch. – A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40 :1489–1506, 2000.

[Itti 00b] L. Itti, C. Koch, J. Braun. – Revisiting spatial vision : toward a unifying model. *J. Opt. Soc. Am. A.*, 17(11) :1899–1917, 2000.

[Itti 01a] L. Itti, C. Koch. – Computational modelling of visual attention. *Nature Rev Neuroscience*, 2(3) :194–203, 2001.

[Itti 01b] L. Itti, C. Koch. – Feature combination strategies for saliency-based visual attention systems. – *SPIE Human Vision and Electronic Imaging*, 2001.

[Itti 04] L. Itti. – Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, 13(10) :1304–1318, Oct 2004.

[James 90] W. James. – The principles of psychology. – New York, Holt, 1890.

[Jameson 55] D. Jameson, L. M. Hurvich. – Some quantitative aspects of an opponent-color theory : I. chromatic responses and spectral saturation. *Journal of the Optical Society of America*, 45 :546–552, 1955.

[Kandel 91] E. R. Kandel, J. R. Schwarz, T. M. Jessel. – Principles of neural science. – New York, Elsevier, 1991.

[Kapadia 95] M. K. Kapadia, M. Ito, C. D. Gilbert, G. Westheimer. – Improvement in visual sensitivity by changes in local context : parallel studies in human observers and in v1 of alert monkeys. *Neuron*, 15(4) :843–856, 1995.

[Köhler 29] W. Köhler. – Gestalt psychology. – New York, Liverigth, 1929.

[Klein 88] R. Klein. – Inhibitory tagging system facilitates visual search. *Nature*, 334 :430–431, 1988.

[Klein 99] R. Klein, W.J. Mac Innes. – Inhibition of return is a foraging facilitator in visual search. *Psychological Science*, 10 :346–352, 1999.

[Koch 84] C. Koch, S. Ullman. – Selecting one among the many : a simple network implementing shifts in selective visual attention. *Massachusetts Institute of Teechnology, CBIP Paper 003*, 1984.

[Koch 85] C. Koch, S. Ullman. – Shifts in selection in visual attention : towards the underlying neural circuitry. *Human Neurobiology*, 4(4) :219–227, 1985.

[Koffka 35] K. Koffka. – Principles of gestalt psychology. – New York, Hartcourt, 1935.

[Kolb 96] H. Kolb, E. Fernandez, R. Nelson. – Webvision : The organization of the retina and the visual system. – http ://webvision.med.utah.edu/, 1996.

[Kowaliski 90] P. Kowaliski. – *Vision et mesure de la couleur.* – Masson, 2 édition, 1990. actualisée par F. Viénot et R. Sève.

[Krauskopf 82] J. Krauskopf, D. R. Williams, D. W. Heeley. – Cardinal direction of color space. *Vision Research*, 22 :1123–1131, 1982.

[Lambrecht 96] C J. Van Den Branden Lambrecht. – *Perceptual models and architectures for video coding applications.* – PhD. Thesis, Ecole polytechnique fédérale de Lausanne, EPFL, 1996.

[Lebart 82] L. Lebart, A. Morineau, J.P. Fénelon. – *Traitement des données statistique : méthodes et programmes.* – Dunod, 1982.

[Leontaris 04] A. Leontaris, P. C. Cosman. – Region-of-interest video compression with a composite and a long-term frame. – *Proceedings of the Seventh IASTED International Conference Computer Graphics and Imaging*, Hawaï, USA, 2004.

[Lin 97] L.J. Lin. – *Video bit-rate control with spline approximated rate-distortion characteristics.* – PhD. Thesis, University of Southern California, 1997.

[Liu 03] H. Liu, X. Xie, W.Y. Ma, H.J. Zhang. – Automatic browsing of large pictures on mobile devices. – *in ACM Multimedia conference*, 2003.

[Liu 05] S. Liu, C.C. Kuo. – Joint temporal-spatial bit allocation for video coding dependency. *IEEE Trans. on Circuits and Systems for Video Techno.*, 15(1), 2005.

[Livingstone 90] M. Livingstone. – Segregation of form, color, movement, and depth processing in the visual system : anatomy, physiology, art and illusion. *Vision and the brain : the organization of the central visual system*, pp. 119–138. – Raven Press, 1990.

[Luo 00] J. Luo, A. Singhal. – On measuring low-level saliency in photographic images. – *IEEE Conf. On Computer Vision and Patten Recognition*, 2000.

[Mannan 96] S.K. Mannan, K.H. Ruddock, D. S. Wooding. – The relationship between the locations of spatial features and those fixations made during visual examination of briefly presented images. *Spatial Vision*, 10(3) :165–188, 1996.

[Mannan 97] S.K. Mannan, K.H. Ruddock, D. S. Wooding. – Fixation sequences made during visual examination of briefly presented 2d images. *Spatial Vision*, 11(2) :157–178, 1997.

[Mannos 74] J. L. Mannos, D. J. Sakrison. – The effects of a visual fidelity criterion on the encoding of images. *IEEE Transactions of Information Theory*, 20(4) :525–535, 1974.

[Marcelja 80] S. Marcelja. – Mathematical description of the responses of simple cortical cells. *J. Opt. Soc. Am. A.*, 70 :1297–1300, 1980.

[Marichal 96] X. Marichal, T. Delmot, C. De Vleeschouwer, V. Warscotte, B. Macq. – Automatic detection of interest areas of an image or of a sequence of images. – *IEEE Inernational Conference on Image Processing*, 1996.

[Meyer 98] F. Meyer. – From connected operators to levelings. – *Proceedings of the fourth international symposium on Mathematical morphology and its applications to image and signal processing*, 1998.

[Milanese 92] R. Milanese, J.M. Bost, T. Pun. – A bottom-up attention system for active vision. – *ECAI92, 10th European Conference on Artificial Intelligence*, pp. 808–810, 1992.

[Milanese 93] R. Milanese. – *Detecting salient regions in an image : from biological evidence to computer implementation.* – Suisse, PhD. Thesis, Université de Genève, 1993.

[Milani 03] S. Milani, L. Celetto, G.A. Mian. – A rate control algorithm for the h264 encoder. – *Sixth Baiona Workshop on Signal Processing in Communications*, 2003.

[Mizobe 01] K. Mizobe, U. Polat, M.W. Pettet, T. Kasamatsu. – Facilitation and suppression of single striate-cell activity by spatially discrete pattern stimuli presented beyond the receptive field. *Visual Neuroscience*, 18 :377–391, 2001.

[Muir 03] L. Muir, I. Richardson, S. Leaper. – Gaze tracking and its application to video coding for sign language. – *Picture Coding Symposium*, 2003.

[Nakayama 89] K. Nakayama, M. Mackeben. – Sustained and transient components of visual attention. *Vision Research*, 29 :1631–1647, 1989.

[Neisser 67] U. Neisser. – Cognitive psychology. – New York, Appleton, 1967.

[Nguyen 95] E. Nguyen. – *Compression sélective et focalisation visuelle : application au codage hybride de séquences d'images.* – PhD. Thesis, Université de Rennes I, 1995.

[Nisbett 05] R.E. Nisbett, A. Norenzayan, E.E. Smith, B.J. Kim. – Cultural preferences for formal versus intuitive reasoning. *Cognitive Science*, 7, 2005.

[Odobez 95] J.M. Odobez, P. Bouthemy. – Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4) :348–365, 1995.

[Oliva 01] A. Oliva, A.B. Torralba. – Modeling the shape of the scene : a holistic representation of the spatial envelope. *International Jounral of Computer Vision*, 43(3) :145–175, 2001.

[Oliva 03] A. Oliva, A. Torralba, M. S. Castelhano, J. M. Henderson. – Top-down control of visual attention in object detection. – *IEEE Inernational Conference on Image Processing*, 2003.

[Ortega 94] A. Ortega, K. Ramchandran, M. Vetterli. – Optimal trellis-based buffered compression and fast approximations. *IEEE Trans. Image Proc.*, 3(1) :26–40, 1994.

[Osberger 98] W. Osberger, A.J. Maeder. – Automatic identification of perceptually important regions in an image. – *14th International Conference on Pattern Recognition*, pp. 701–704, 1998.

[Parkhurst 02] D.J. Parkhurst. – *Selective attention in natural vision : using computational models to quantify stimulus-driven attentional allocation.* – Baltimore, Maryland, USA, PhD. Thesis, The Johns Hopkins University, 2002.

[Parkhurst 04] D.J. Parkhurst, E. Niebur. – Texture contrast attracts overt visual attention in natural scenes. *European Journal of Neuroscience*, 19 :783–789, 2004.

[Pattanaik 89] S.N. Pattanaik, J.A. Ferwerda, M.D. Fairchild, D.P. Greenberg. – A multiscale model of adaptation and spatial vision for realistic image display. – *Proceedings of the SIGGRAPH 98*, pp. 287–298, 1989.

[Peli 93] E. Peli, L. E. Arend, G. M. Young, R. B. Goldstein. – Contrast sensitivity to patch stimuli : effects of spatial bandwith and temporal presentation. *Spatial Vision*, 7(1) :1–14, 1993.

[Peters 05] R. J. Peters, A. Iyer, L. Itti, C. Koch. – Components of bottom-up gaze allocation in natural images. *Vision Research*, 2005.

[Polat 93] U. Polat, D. Sagi. – Lateral interactions between spatial channels : suppression and facilitation revealed by lateral masking experiments. *Visual Research*, 33(7) :993–999, 1993.

[Polat 94] U. Polat, D. Sagi. – The architecture of perceptual spatial interactions. *Visual Research*, 34(1) :73–78, 1994.

[Posner 80] M.I. Posner. – Orienting of attention. *Quaterly Journal of Experimental Psychology*, 32 :3–25, 1980.

[Posner 84] M. I. Posner, Y. Cohen. – Components of visual orienting. *Attention and performance X*, éd. par H. Bouma, D.G. Bouwhuis, pp. 531–556. – 1984.

[Privitera 00] C. M. Privitera, L. Stark. – Algorithms for defining visual regions-of-interest : Comparison with eye fixations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22 :970–982, 2000.

[Ramchandran 94] K. Ramchandran, A. Ortega, M. Vetterli. – Bit allocation for dependent quantization with applications to multiresolution and mpeg video coders. *IEEE Trans. Image Proc.*, 3(5) :533–545, 1994.

[Reinagel 99] P. Reinagel, A.M. Zador. – Natural scene statistics at the center of gaze. *Network : Computation in Neural Systems*, 10 :341–350, 1999.

[Richardson 03] I.E.G. Richardson. – *H.264 and MPEG-4 video compression : video coding for next-generation multimedia.* – Wiley, 2003.

[Salvucci 99] D.D. Salvucci. – *Mapping eye movements to cognitive processes.* – Pittsburgh, PhD. Thesis, Carnegie Mellon University, 1999.

[Seyler 59] A.J. Seyler, Z.L. Budrikis. – Measurements of temporal adaptation to spatial detail vision. *Nature Rev Neuroscience*, 184 :1215–1217, 1959.

[Seyler 65] A.J. Seyler, Z.L. Budrikis. – Details perception after scene changes in television image presentations. *IEEE Trans. Inform. Theory*, 11(1) :31–43, 1965.

[Shoham 88] Y. Shoham, A. Gersho. – Efficient bit allocation for an arbitrary set of quantizers. *IEEE Trans. Acoust., Speech, Signal Processing*, 36 :1445–1453, 1988.

[Sénane 96] H. Sénane. – *Représentation d'images en sous-bandes visuelles. Application au codage d'images de télévision sans défaut visuel.* – Université de Nantes, PhD. Thesis, IRESTE, 1996.

[Tam 95] W.J. Tam. – Visual masking at video scene cuts. – *SPIE Human Vision and Electronic Imaging*, vol. 2411, pp. 111–119, 1995.

[Tang 04] C.W. Tang, C.H. Chen, Y.H. Yu, C.J. Tsai. – A novel visual distortion sensitivity analysis for video encoder bit allocation. – *Proceedings ICIP-04 (IEEE International Conference on Image Processing)*, Singapour, 2004.

[Torralba 03] A. B. Torralba. – Modeling global scene factors in attention. *J Opt AM A Opt Image Sci Vis*, 20(7) :1407–1418, 2003.

[Treisman 80] A. Treisman, G. Gelade. – A feature integration theory of attention. *Cognitive Psychology*, 12 :97–136, 1980.

[Ts'o 86] D. Ts'o, C. D. Gilbert, T. N. Wiessel. – Relationships between horizontal interactions and functionnal architecture in cat striate cortex as revealed by cross-correlation analysis. *Jour. Neurosci.*, 6 :1160–1170, 1986.

[Tsotsos 95] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y.H. Lai, N. Davis, F. Nuflo. – Modelling visual attention via selective tuning. *Artificial intelligence*, 78 :507–545, 1995.

[Uz 93] K.M. Uz, J.M. Shapiro, M. Czigler. – Optimal bit-allocation in the presence of quantizer feedback. – *Proc. of ICASSP'93*, vol. 5, 1993.

[Valois 58] R. De Valois, C. Smith, S. Kitai, S. Karoly. – Responses of single cells in different layers of the primate lateral geniculate nucleus to monochromatic light. *Science*, 127 :238–239, 1958.

[Valois 88] R. De Valois, K. K. De Valois. – Spatial vision. *Spatial vision.* – Oxford University Press, 1988.

[Valois 92] R. De Valois, K. K. De Valois. – A multi-stage color model. *Vision Research*, 33(8) :1035–1035, 1992.

[Valois 00] R. De Valois, , N.P. Cottaris, L.E. Mahon, S.D. Elfar, J.A. Wilson. – Spatial and temporal receptive fields of geniculate and cortical cells and diretional selectivity. *Vision Research*, 20 :3685–3702, 2000.

[Varela 98] F. Varela. – Le cerveau n'est pas un ordinateur, 1998. La Recherche.

[Varela 99] F. Varela. – La conscience : la tempête sous un crâne, 1999. Interview de I. Brisson Le Figaro.

[VQEG] VQEG. – Final report from the video quality experts group on the validation of objective models of video quality assessment. Site officiel VQEG http ://www.its.bldrdoc.gov/vqeg/.

[Walker 99] G. A. Walker, I. Ohzawa, R.D. Freeman. – Asymetric suppression outside the classical receptive field of the visual cortex. *Jour. Neurosci.*, 19 :10536–10553, 1999.

[Wang 04] J. Wang, M. Reinders, R. Lagendijk, J. Lindenberg, M. Kankanhalli. – Video content representation on tiny devices. – *IEEE Conference on Multimedia and Expo.*, Taiwan, 2004.

[Watson 87] A. B. Watson. – The cortex transform : Rapid computation of simulated neural images. *Computer Vision, Graphics and Image Processing*, 39 :311–327, 1987.

[Watson 98] A.B. Watson. – Toward a perceptual video quality metric. – *SPIE Human Vision and Electronic Imaging*, vol. 3299, pp. 946–949, 1998.

[Webster 90] M. A. Webster, K. K. De Valois, E. Switkes. – Orientation and spatial-frequency discrimination for luminance and chromatic gratings. *Journal of the Optical Society of America*, 7(6) :1034–1049, 1990.

[Wertheimer 38] M. Wertheimer. – Laws of organization in perceptual forms. *A source book of Gestalt psychology*, éd. par W. Ellis, pp. 71–88. – http ://psychclassics.yorku.ca, 1938.

[Wolfe 89] J.M Wolfe, K.R. Cave, S.L. Franzel. – Guided search : an alternative to the feature integration model for visual search. *Journal Exp. Psychol. Hum. Percept. Perform*, 15(3) :419–433, 1989.

[Wolfe 04] J.M. Wolfe, T.S. Horowitz. – What attributes guide the deployement of visual attention and how do they do it ? *Nature Rev Neuroscience*, 5, 2004.

[Wooding 02] D. S. Wooding. – Eye movements of large population : Ii. deriving regions of interest, coverage, and similarity using fixation maps. *Behavior Research Methods, Instruments and Computers*, 34(4) :509–517, 2002.

[Wässle 91] H. Wässle, B. B. Boycott. – Functional architecture of mammalian retina. *Physiol. Rev.*, 71(2) :447–480, 1991.

[Yang 00] Y. Yang, S. S. Hemami. – Rate-distortion optimizations for region and object based wavelet video coding. – *in Proc. 34th Asilomar Conference on Signals, Systems, and Computers*, 2000.

[Yantis 96] S. Yantis, J. Jonidas. – Attentional capture by abrupt onsets and selective attention : evidence from visual search. *Jour. Exp. Psychol. Hum. Percept. Perf.*, 20 :1505–1513, 1996.

[Yarbus 67] A. Yarbus. – *Eye movements and vision.* – L.A. Riggs, Trans., New-York :Plenum Press, 1967.

[Yee 01] H. Yee, S. Pattanaik. – Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. – *IACM*, 2001.

[Zettl 90] H. Zettl. – Sight, sound, motion : applied media aesthetics, 1990. Belmont, CA :Wadsworth.

[Zhao 96] J. Zhao, Y. Shimazu, K. Ohta, R. Hayasaka, Y. Matsushita. – An outstandingness oriented image segmentation and its application. – *Inernational Symposium on Signal Processing and its Applications*, 1996.

Contributions scientifiques

Journaux :

1. O. Le Meur, P. Le Callet, D. Barba et D. Thoreau, *A coherent computational approach to model bottom-up visual attention*, soumis à IEEE Transactions on Pattern Analysis Machine Intelligence, juillet 2004 en seconde révision mineure (août 2005).

2. O. Le Meur, P. Le Callet et D. Barba, *Selective H.264 video coding based on a saliency map*, soumis à EURASIP, Special Issue on Advanced Video Technologies and Applications for H.264/AVC and Beyond, juillet 2005.

3. O. Le Meur, P. Le Callet et D. Barba, *A spatio-temporal model of bottom-up visual selective attention : description and assessment*, en préparation pour soumission dans Vision Research.

Colloques et conférences avec comité de lecture et actes :

1. O. Le Meur, P. Le Callet, D. Barba, D. Thoreau et E. Francois, *From low level perception to high level perception, a coherent approach for visual attention modeling*, Proc. SPIE Human Vision and Electronic Imaging IX (HVEI'04), San Jose, CA, (B. Rogowitz, T. N. Pappas Ed.), pp. 284-295, Janvier 2004.

2. O. Le Meur, P. Le Callet, D. Barba et D. Thoreau, *Masking effect in visual attention modeling*, Proc. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Lisbonne, Portugal, Avril 2004.

3. O. Le Meur, P. Le Callet, D. Barba et D. Thoreau, *Bottom-up attention modeling : quantitative comparison of predicted saliency maps with observers eye-tracking data*, ECVP 2004, Budapest, Hongrie, Août 2004.

4. O. Le Meur, P. Le Callet, D. Barba et D. Thoreau, *Performance assessment of a visual attention system entirely based on a human vision modeling*, Proc. ICIP-04 (IEEE International Conference on Image Processing), pp. 2327-2330, Singapore, Octobre 2004.

5. O. Le Meur, P. Le Callet, D. Barba et D. Thoreau, *A human visual model-based approach of the visual attention and performance evaluation*, Proc. SPIE Human Vision and Electronic Imaging X (HVEI'05), San Jose, CA, (B. Rogowitz, T. N. Pappas Ed.), pp. 258-267, Janvier 2005.

6. O. Le Meur, P. Le Callet, D. Barba et D. Thoreau, *A spatio-temporal model of the selective human visual attention*, Proc. ICIP-05 (IEEE International Conference on Image Processing), pp. 1188-1191, Gêne, Italie, Septembre 2005.

Conférence nationale avec comité de lecture et actes :

1. O. Le Meur, P. Le Callet, D. Barba et D. Thoreau, Modélisation spatio-temporelle de l'attention visuelle, CORESA 2005 (COmpression et REprésentation des SIgnaux Audiovisuels), Rennes, septembre 2005.

Présentations à journée d'études :

1. O. Le Meur, P. Le Callet, D. Barba et D. Thoreau, *From low level perception to high level perception, a coherent approach for visual attention modeling*, séminaire de présentations des thèses en codage de l'image, France Télecom Rennes, 24 juin 2004.

2. O. Le Meur, P. Le Callet, D. Barba et D. Thoreau, *L'attention visuelle, critère déterminant pour la compression sélective : application au codage H.264 à qualité différenciée*, GdR ISIS, thèmes compression, Paris, 3 mars 2005.

Brevets :

1. O. Le Meur, P. Le Callet, D. Barba, D. Thoreau et E. François, *Automatic determination of fixation points based on the luminance signal and on the chrominance signal*, THOMSON LICENSING S.A, EP053471, 18 décembre 2003.

2. O. Le Meur, P. Le Callet, D. Barba, D. Thoreau et P. Guillotel, *Selective H.264 video coding based on side-information : a bottom-up approach*, THOMSON LICENSING S.A, FR0552345.

3. O. Le Meur, P. Le Callet et D. Barba, *Structural image distortion metric based on a saliency map*, THOMSON LICENSING S.A, EP05291498, 25 mai 2005.

4. O. Le Meur, D. Thoreau, J. Kypreos et C. Chevance, *Advanced video coding using a pre-filtering driven by a saliency map*, THOMSON LICENSING S.A, EP05261126, 11 juillet 2005.

5. O. Le Meur, P. Guillotel et J. Haddad, *Computation of reduced video centered on conspicuous areas*, THOMSON LICENSING S.A, EP0552362.