



DEA Imagerie Vision Robotique

Mémoire de stage

ATTENTION VISUELLE MULTI-ECHELLE

Aurélie Bugeau

Tuteur : Bill Triggs

Rapporteur extérieur : Sylvain Meignen

Remerciements :

Je tiens à remercier en premier lieu Bill Triggs pour m'avoir proposé ce stage et m'avoir encadré pendant ces 5 mois. Je le remercie pour ses idées, conseils et explications, et pour la liberté de recherche qu'il a bien voulu me laisser.

Je remercie l'équipe Learning et Recognition de l'INRIA Rhône-Alpes pour m'avoir accueilli.

Mes remerciements vont aussi à Sylvain Meignen du laboratoire LMC pour avoir accepté de tenir le rôle de rapporteur extérieur de ce mémoire.

Un grand merci à tous ceux qui de près ou de loin m'ont aidé pendant cette période, et en particulier Tijmon et à Hans pour leur aide précieuse sur les petits problèmes informatiques.

Résumé :

L'attention visuelle est la capacité d'un système de vision, qu'il soit humain ou artificiel, à sélectionner rapidement les informations les plus pertinentes de son environnement. Son but est de choisir quelle zone de l'image analyser avant les autres car elle serait potentiellement plus intéressante. Ainsi elle permet de réduire la quantité d'informations à traiter, et par conséquent d'accélérer l'ensemble du processus de vision. Dans ce mémoire, un système d'attention visuelle [13], basé sur les processus bottom-up, est présenté. Dans ce modèle, différentes caractéristiques de l'image de départ sont combinées en une carte de saillance, dont les pics codent les points d'attention. Une brève évaluation est donnée, et diverses améliorations possibles sont proposées. Ces modifications permettent la mise en place d'un nouveau système d'attention visuelle, cette fois-ci multi-échelle, basé sur [13].

Table des matières

1	Introduction	7
2	Le laboratoire d'accueil	8
2.1	Axes de recherche	8
2.2	Domaines d'applications	9
3	Etat de l'art	10
3.1	Le système visuel humain	10
3.2	Les systèmes de couleurs	12
3.3	L'attention visuelle humaine	13
3.4	Un modèle d'attention visuelle basé sur le processus humain	14
4	Modèle d'attention visuelle basée sur la saillance	16
4.1	Description du modèle	16
4.1.1	Extraction des cartes de caractéristiques	16
4.1.2	Extraction des cartes d'évidence	19
4.1.3	Extraction de la carte de saillance et des points saillants	20
4.2	Evaluation de la méthode	21
5	Analyse et améliorations du modèle	22
5.1	Système multi-échelle	22
5.2	Cartes de caractéristiques	25
5.2.1	Espaces de couleurs	25
5.2.2	Nombre d'orientations	26
5.2.3	Pyramide gaussienne	27
5.3	Carte de saillance	30
5.3.1	Combinaison des cartes d'évidence	30
5.3.2	Normalisation	31
6	Système d'attention visuelle multi-échelle	34
7	Conclusion et Perspectives	38
7.1	Ajout des symétries	38
7.2	Zone de recherche du point saillant suivant	39
7.3	Ajout du processus top-down	39
7.4	Utilisation du contexte	39
7.5	Application à la détection d'objets	39
A	Résultats obtenus avec différents espaces de couleur	41
B	Résultats obtenus avec différentes orientations	43

C	Influence des cartes d'évidence	44
C.1	Résultats de l'influence du contraste d'intensité	44
C.2	Résultats de l'influence de la couleur	46
C.3	Résultats de l'influence de l'orientation	48

Table des figures

3.1	Oppositions Rouge/Vert et Jaune/Bleu	11
3.2	Filtre de Gabor 1 dimension puis 2 dimensions	12
3.3	Mode d'exploration oculaire une image. Dans le cas de portraits se sont les yeux et le nez qui sont particulièrement visés [17].	13
3.4	Trouvez le T rouge a) impossible de ne pas le voir : processus bottom-up ; b) Plus difficile qu'à coté car cela demande un processus actif mettant en jeu diverses "stratégies" perceptives : processus top-down.	14
3.5	Système d'attention visuelle bottom-up, d'après [12]	15
4.1	Architecture du modèle défini dans [13]	17
4.2	a) Structure pyramidale; b) Exemple d'une pyramide dyadique; c) Résultat obtenu avec l'opérateur "centre - région contournante" (pyramide laplacienne de gaussiennes)	18
4.3	Opérateur de normalisation.	20
5.1	Mise en place du système multi-échelle.	23
5.2	Correspondance entre des points saillants à différentes échelles.	24
5.3	Résultats obtenus avec le système multi-échelles a) système initial; b) Utilisation de 2 échelles $c'=2,3$; b) Utilisation de 3 échelles $c'=2,3,4$; b) Utilisation de 4 échelles $c'=2,3,4,5$	24
5.4	Taille de l'angle définissant le nombre d'orientations pouvant être détectées par le filtre de Gabor suivant la largeur de la bande passante	27
5.5	Construction des niveaux intermediaires à partir d'une pyramide dyadique.	28
5.6	Résultats obtenus avec a) une pyramide dyadique, b) 1 image intermédiaire, c) 2 images intermédiaires, d) 3 images intermédiaires.	29
5.7	Filtre utilisé pour la normalisation itérative	32
5.8	Résultat de la normalisation par différence gaussienne sur la carte de saillance a) image originale; b) 0 itération; c) 2 itérations d) 4 itérations; e) 8 itérations	33
6.1	Architecture du nouveau modèle	35
6.2	a) Résultats obtenus avec le système décrit au chapitre 4; b) Résultats avec le système d'attention visuelle multi-échelle	36
6.3	Résultat obtenu sur un visage dont les cheveux présentent un contraste assez faible avec le fond.	37
A.1	Résultats obtenus en partant de la méthode [13], avec différents systèmes de couleur. a)méthode décrite dans l'article (système de couleur : RGB normalisé); b)RGB c)Lab.	42
B.1	Résultats obtenus en partant de la méthode [13] , avec plusieurs différentes orientations. a) 4 orientations, comme dans l'article; b) 8 orientations; c) 12 orientations.	43

C.1	a) Résultats obtenus en ne considérant que les caractéristiques de couleur et d'orientations; b) Résultats obtenus avec la méthode Itti; c) Résultats en donnant deux fois plus d'importance au contraste d'intensité.	45
C.2	a) Résultats obtenus en ne considérant que les caractéristiques de contraste et d'orientations; b) Résultats obtenus avec la méthode Itti; c) Résultats obtenus en donnant deux fois plus d'importance aux couleurs qu'au contraste et à l'orientation, c) Résultats obtenus en donnant deux fois plus d'importance aux couleurs	47
C.3	a) Résultats obtenus en ne considérant que les caractéristiques de contraste et de couleurs; b) Résultats obtenus avec la méthode Itti; c) Résultats obtenus en donnant deux fois plus d'importance à l'orientation qu'au contraste et aux couleurs.	49

Chapitre 1

Introduction

La vision est incontestablement le sens le plus développé chez l'homme : Les humains ont une capacité remarquable d'analyse et d'interprétation en temps réels des scènes complexes. Par ailleurs, la vision par ordinateur est devenu ces dernières années un domaine en plein essor de part ses multiples applications en médecine, robotique, surveillance. ... Ainsi, l'étude du système visuel humain ou animal dans le but d'en comprendre et copier les mécanismes pour la vision par ordinateur est particulièrement intéressante. Diverses analyses des systèmes visuels animaux ont été menées ces trente dernières années. Elles montrent que le regard est en premier lieu attiré par certaines caractéristiques d'une scène, afin d'en réduire la complexité d'analyse. C'est ce que l'on appelle l'attention visuelle, capacité d'un système de vision à sélectionner rapidement les informations les plus pertinentes du champ de vision. Il s'agit d'un mécanisme indispensable pour l'analyse temps réel de notre environnement.

Des recherches ont été conduites afin de mettre en place un système informatique d'attention visuelle proche de celui utilisé par les primates. Une partie des modèles d'attention visuelle proposés jusqu'à présent est basée sur les connaissances physiologiques des primates. Chez eux, les zones attentionnelles sont sélectionnées en combinant un balayage de la scène indépendant du but recherché, faisant juste ressortir les points saillants (processus bottom-up) avec un balayage dépendant du but recherché et nécessitant des connaissances préalables (processus top-down). Ces systèmes d'attention visuelle ont pour rôle principal l'accélération et l'amélioration des modèles de vision artificielle. Les applications possibles sont diverses : compression d'images, détection d'objets ou même amélioration des panneaux publicitaires....

Le but du stage de DEA était de construire un système de direction d'attention visuelle, basé sur le mécanisme qui pilote les fixations visuelles humaines, en ne considérant que le processus bottom-up. Ce système doit être capable d'analyser n'importe quelle image présentée et de retourner les coordonnées des points les plus saillants. L'application directe de cette méthode devra être la détection de classe d'objets. Le travail a consisté à reprendre un modèle développé par Itti, Koch et Niebur [13], à l'analyser et à le modifier, en particulier en détectant les points saillants à différentes échelles. Le projet a été réalisé au sein de l'équipe LEAR (Learning and Recognition) du laboratoire INRIA, sous la tutelle du chercheur Bill Triggs.

Ce mémoire présente l'ensemble du travail et des résultats obtenus au cours de ce stage. Il s'articule autour du plan suivant. Une brève description du laboratoire d'accueil est suivie d'un rapide état de l'art présentant le mécanisme d'attention visuelle humaine. Dans le chapitre 4, nous expliquons le modèle sur lequel a été basé le travail, et dans le chapitre 5 sont expliquées les améliorations apportées. Ce rapport se termine par la présentation et l'évaluation du système obtenu à partir de ces modifications.

Chapitre 2

Le laboratoire d'accueil

Le stage de DEA s'est déroulé au sein de l'équipe LEAR du laboratoire GRAVIR/INRIA. LEAR a été créé officiellement le 1er juillet 2003, et est composé de trois enseignants chercheurs, de sept doctorants et post-doctorants et d'un chercheur invité.

2.1 Axes de recherche

Les recherches menées dans cette équipe s'orientent autour de trois axes principaux : la description d'images, l'apprentissage et la reconnaissance :

– *La description d'images :*

De nombreux descripteurs d'image invariants au changement de point de vue ou de luminosité existent déjà. Cependant, des recherches sont encore nécessaires pour obtenir des descripteurs robustes aux différentes transformations d'images. De meilleurs descripteurs rendraient également plus simples la phase d'apprentissage et faciliteraient le choix des données d'apprentissage. En particulier, il serait intéressant d'obtenir, pour une seule image, un nombre important de descripteurs locaux ou globaux invariants aux changements photométriques et géométriques de l'image. Ces descripteurs doivent contenir un grand nombre d'informations afin d'être exploitable pour la reconnaissance d'objets.

L'obtention de descripteurs locaux demande l'extraction de fragments de l'image. Plutôt que d'utiliser des fragments généraux, il serait préférable d'utiliser des points intéressants de l'image, donnant plus d'informations. Ces points doivent de plus pouvoir être retrouvés dans d'autres images, quelle que soit l'échelle.

– *L'apprentissage :*

Les principaux défis actuels de l'apprentissage sont :

- * gérer le nombre très important de données que contiennent les images
- * obtenir un nombre de classe plus important que les deux classes "objet" / "non-objet"
- * réussir à entraîner les classifieurs et descripteurs sur un faible nombre d'images

La recherche dans ce domaine consiste à évaluer les différentes méthodes existantes et à en créer de nouvelles répondant aux critères ci-dessus.

– *La reconnaissance :*

La reconnaissance est en fait la combinaison de la description d'images et de l'apprentissage. Pour l'instant, le choix des descripteurs et des méthodes d'apprentissage est fait à la main. Ainsi un gros challenge de la reconnaissance de classe d'objets ou d'objets est l'automatisation du choix des méthodes et de leurs couplages.

2.2 Domaines d'applications

Les domaines d'applications de la reconnaissance d'objets sont très vastes : robotique, médecine, sécurité, interaction homme-machine, exploration spatiale, collecte de données. ... Le principal domaine d'application pour l'équipe LEAR est l'indexation d'images et de vidéos. Par exemple, LEAR cherche à développer un assistant visuel personnel permettant d'identifier la catégorie à laquelle l'image visualisée appartient et fournissant à l'utilisateur des informations sur cet objet. Une autre application est le développement d'un système qui, à partir d'une série de personnages, actions ou scènes, est capable de retrouver dans une vidéo les endroits où ces éléments apparaissent.

Chapitre 3

Etat de l'art

3.1 Le système visuel humain

Modéliser la vision a toujours été un défi en informatique et en intelligence artificielle. Il apparaît qu'étudier le cerveau et la vision humaine peut faire progresser en vision par ordinateur.

Au niveau de la rétine, plusieurs dizaines de types différents de cellules (photo réceptrices, ganglionnaires...) codent les informations visuelles, chacune réalisant une fonction très spécialisée. La rétine possède par exemple un grand nombre de cellules photoréceptrices : environ 4 millions de cônes pour un peu plus de 100 millions de bâtonnets. Les bâtonnets servent à l'adaptation aux changements de luminosité.

Peu nombreux et moins sensibles à la lumière que les bâtonnets, les cônes sont responsables de la vision haute résolution. On distingue 3 types de cônes : les cônes S sensibles à des longueurs d'onde courtes (short), les cônes M sensibles à des longueurs d'onde moyennes (medium) et les cônes L sensibles à des longueurs d'onde longues (long). Ils sont à l'origine de l'aspect trichromatique de la vision des couleurs [18]. Les cônes L sont sensibles au jaune-vert à rouge, les cônes M au vert et les cônes S au bleu. Leur répartition n'est pas égale : on compte 64% de cônes L, 32% de M et seulement 2% de S.

Ces cellules photoréceptrices sont connectées aux cellules bipolaires, qui à leur tour, sont connectées aux cellules ganglionnaires dont les axones se prolongent dans le nerf optique. Les cellules ganglionnaires assurent une conversion tension-fréquence : le nerf optique transmet des informations au cerveau sous forme de trains d'impulsions modulés en fréquence. A chaque cellule ganglionnaire correspond un champ récepteur : région de la rétine à partir de laquelle on peut influencer un neurone. Les champs récepteurs sont modélisés comme la différence entre deux distributions de Gaussienne, leurs donnant la forme d'un "chapeau mexicain"). Dès 1952, deux types de cellules ganglionnaires ont été répertoriés : les cellules à centre ON et les cellules à centre OFF. Malgré l'obscurité, les cellules ganglionnaires émettent un niveau moyen d'impulsions. Les cellules à centre ON augmentent ce nombre d'impulsions lorsqu'un stimulus éclaire le centre du champ récepteur et deviennent silencieuses si le stimulus éclaire la périphérie. Les cellules à centre OFF montrent un comportement inverse. Les oppositions noir/blanc, rouge/vert et bleu/jaune obtenues par ces cellules ganglionnaires décrivent aussi la vision chromatique humaine. Les cellules ganglionnaires permettent la détection de contours, et sont insensibles à l'orientation du stimulus.

Ainsi la rétine est composée de cônes sensibles au rouge, vert et bleu, mais c'est l'information sur les oppositions de couleur qui est transmise au cerveau (figure 3.1), le jaune étant obtenu par combinaison du rouge et du vert.

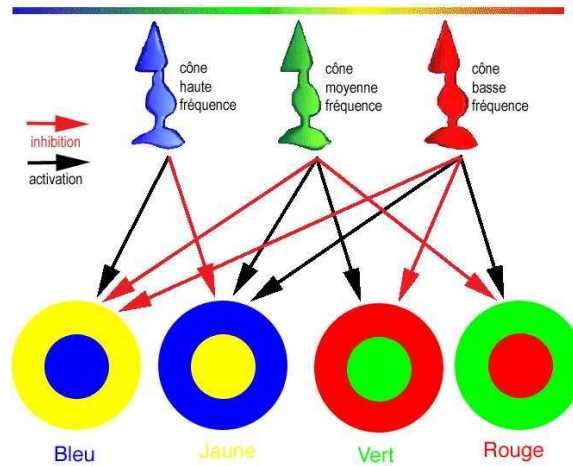


FIG. 3.1 – Oppositions Rouge/Vert et Jaune/Bleu

Les signaux électriques issus des deux rétines empruntent les deux nerfs optiques, qui se rencontrent au chiasma optique, puis gagnent les corps genouillés latéraux, excroissances du cerveau. De là, les signaux rejoignent le cortex visuel pour être reconstruits en images intelligibles. L'analyse du cortex visuel montre que les quelques 10^{10} neurones qui le constituent possèdent un arrangement hiérarchique bien défini avec quelques types de neurones. Ces neurones sont répertoriés comme simples, complexes, hypercomplexes et hypercomplexes d'ordre élevé, avec des propriétés définies selon leur champ récepteur :

- Cellule simple : Le champ récepteur est de forme allongée, avec un centre entouré de 2 régions antagonistes. Le stimulus le plus efficace est alors une barrette orientée dans le même sens que le champ récepteur. Elles sont donc sensibles à l'orientation mais comme la réponse n'est pas invariante par translation, elles sont aussi sensibles à la position.
- Les cellules complexes : Ces cellules utilisent les cellules simples en entrée. Elles sont sensibles à l'orientation mais pas à la position exacte de l'excitation dans le champ visuel : elles signalent l'orientation indépendamment de la position (le nombre approximatif de directions quantifiées doit être de 30).
- Les cellules hypercomplexes : Elles répondent à des discontinuités comme la fin d'une ligne, un coin, un angle droit éclairé d'un côté et sombre de l'autre... Il n'y a réponse que pour une orientation et une discontinuité (fin de ligne, angle...) spécifiques : elles permettent la perception des formes.

Ces trois types de cellules sont donc sensibles à l'orientation : elles répondent aux changements d'intensité spatiale suivant une certaine orientation. D'autres études ont ajouté que les humains utilisent une vingtaine voire une trentaine d'orientations et les primates une douzaine. Macelja et Daugman [14] [3] ont précisé que une partie de ces cellules peut être décrite par la partie réelle d'une fonction de Gabor [8].

La partie réelle des filtres de Gabor, décrivant bien les cellules du cortex visuel, sont le produit d'un cosinus avec une enveloppe gaussienne :

$$Ga(x, y, \theta, \sigma, \phi) = \cos(2\pi f x' + \phi) \cdot e^{-\frac{x'^2 + y'^2}{\sigma^2}} \tag{3.1}$$

avec $x' = x \cdot \cos(\theta) + y \cdot \sin(\theta)$ et $y' = y \cdot \cos(\theta) - x \cdot \sin(\theta)$, f la fréquence de la sinusoïde et σ la largeur de l'enveloppe gaussienne. La figure 3.2 montre l'allure d'un filtre de Gabor $Ga(\theta, \sigma)$. La réponse de ces cellules à une image peut être obtenue en convoluant l'image avec un filtre de Gabor.

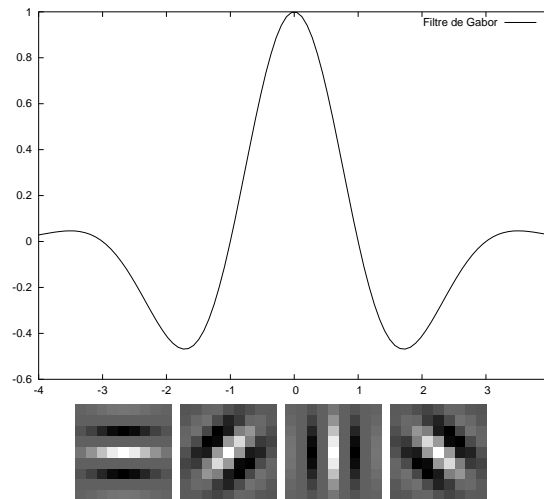


FIG. 3.2 – Filtre de Gabor 1 dimension puis 2 dimensions

Dans cette section, nous avons montré que le cerveau utilise différentes caractéristiques, comme la couleur et l'orientation, pour l'analyse d'une scène.

3.2 Les systèmes de couleurs

On appelle espace de couleurs la représentation mathématique d'un ensemble de couleurs. Nous détaillons ci-dessous les trois qui nous importent dans la suite :

- **RGB** : Le codage RGB, mis au point en 1931 par la Commission Internationale de l'Eclairage (CIE) consiste à représenter l'espace des couleurs à partir de trois rayonnements monochromatiques de couleurs : rouge (de longueur d'onde égale à 700,0 nm), vert (de longueur d'onde égale à 546,1 nm), bleu (de longueur d'onde égale à 435,8 nm). Cet espace de couleur correspond à la façon dont les couleurs sont généralement codées informatiquement ou plus exactement à la manière dont les tubes cathodiques des écrans d'ordinateurs représentent les couleurs.
- $RGB_{normalise}$: Les intensités varient beaucoup avec l'illumination, les chrominances moins. Séparer la chrominance de la luminance est donc souvent utile et ils existent plusieurs espaces de couleurs qui le permettent. L'un d'entre eux est l'espace $RGB_{normalis}$:

$$R_n = \frac{R}{R + G + B}, G_n = \frac{G}{R + G + B}, B_n = \frac{B}{R + G + B} \tag{3.2}$$

- **La*b*** : Les couleurs peuvent être perçues différemment selon les individus et peuvent être affichées différemment selon les périphériques d'affichage. Afin d'établir des standards permettant

de définir une couleur indépendamment des périphériques utilisés, la CIE (Commission Internationale de l'Éclairage) a mis en place des critères basés sur la perception de la couleur par l'œil humain, grâce à un triple stimulus. En 1931 la CIE a élaboré le système colorimétrique xyY représentant les couleurs selon leur chromaticité (axes x et y) et leur luminance (axe Y). Le diagramme de chromaticité (ou diagramme chromatique), issu d'une transformation mathématique représente sur la périphérie les couleurs pures, c'est-à-dire les rayonnements monochromatiques correspondant aux couleurs du spectre (couleurs de l'arc-en-ciel), repérées par leur longueur d'onde. Toutefois ce mode de représentation purement mathématique ne tient pas compte des facteurs physiologiques de perception de la couleur par l'œil humain, ce qui résulte en un diagramme de chromaticité laissant par exemple une place beaucoup trop large aux couleurs vertes. Ainsi, en 1976, afin de pallier les lacunes du modèle xyY , la CIE développe le modèle colorimétrique La^*b^* (aussi connu sous le nom de CIELab), dans lequel une couleur est repérée par trois valeurs :

- L , la luminance, exprimée en pourcentage (0 pour le noir à 100 pour le blanc)
 - a et b deux gammes de couleur allant respectivement du vert au rouge et du bleu au jaune avec des valeurs allant de -120 à +120. Le mode La^*b^* couvre ainsi l'intégralité du spectre visible par l'œil humain et le représente de manière uniforme. Il permet donc de décrire l'ensemble des couleurs visibles indépendamment de toute technologie graphique.
- L'espace La^*b^* est optimisé pour "l'uniformité perceptuelle". En effet, un petit changement ($\Delta L, \Delta a, \Delta b$) donne le même ordre de différence perceptuelle n'importe où dans l'espace La^*b^* .

3.3 L'attention visuelle humaine

Afin de proposer une approche de la vision basée sur les mécanismes attentionnels il faut essayer d'appréhender l'attention humaine. L'attention est l'action de concentrer son esprit sur un ou plusieurs éléments particuliers. Cela renvoie au fait que nous ne traitons pas avec le même degré de profondeur l'ensemble des informations en provenance de notre environnement. Ainsi, le déplacement du regard ne se fait pas de manière aléatoire sur l'image mais suit les traits saillants (figure 3.3), privilégiant certains aspects de la scène. Le regard reste en général environ un quart de seconde sur un point, la scrutation à proprement parler ne durant que 50 ms.

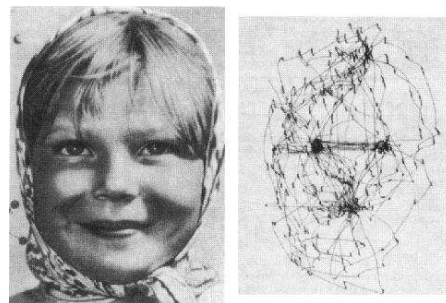


FIG. 3.3 – Mode d'exploration oculaire une image. Dans le cas de portraits se sont les yeux et le nez qui sont particulièrement visés [17].

Deux grands types de processus sont à l'oeuvre dans le système perceptif :

- les processus " bottom-up " : ils sont automatiques, et renvoient au fait que nous pouvons extraire de notre environnement, ou plutôt sélectionner, (via nos sens) différentes informations qui vont être agencées vers notre cerveau pour y être décodées (figure 3.4 a)).

- les processus " top-down " : ce sont des processus contrôlés. Dans certains cas c'est le cerveau lui-même qui envoie directement l'information vers les systèmes sensoriels. En effet les perceptions peuvent être influencées par ce que s'attend à voir ou a en mémoire une personne. Ainsi l'approche top-down est une approche principalement dirigée par les concepts. C'est plutôt la connaissance a priori qui va guider le processus de vision (3.4 b)).

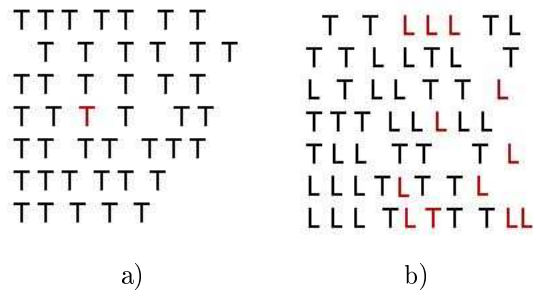


FIG. 3.4 – Trouvez le T rouge a) impossible de ne pas le voir : processus bottom-up ; b) Plus difficile qu'à coté car cela demande un processus actif mettant en jeu diverses "stratégies" perceptives : processus top-down.

Ces deux processus sont indispensables pour pouvoir interpréter des scènes en temps réel. Le mécanisme qui nous intéresse ici est le " bottom-up ". En effet, afin de réduire la complexité de la reconnaissance des formes, nous souhaitons pouvoir extraire d'une image ses informations pertinentes, sans connaissances préalables.

3.4 Un modèle d'attention visuelle basé sur le processus humain

Comme pour la vision humaine, l'attention visuelle représente un outil fondamental pour la vision par ordinateur. De nombreuses études ont été réalisées depuis une vingtaine d'années pour trouver un modèle informatique d'attention visuelle. Un des modèles à partir duquel de nombreux travaux sont issus est le modèle de contrôle de l'attention bottom-up de Koch et Ullman [12].

Ce modèle commence par l'extraction d'un certain nombre de caractéristiques de la scène comme la couleur et l'orientation. Les caractéristiques extraites donnent naissance à des cartes "d'évidence", qui elles donnent les points importants d'une scène pour chaque caractéristique. Enfin, ces cartes d'évidence sont combinées en une carte, dite "carte de saillance", qui encode l'intensité du stimulus ou la saillance pour toute position de la scène visuelle.

La carte de saillance reçoit des entrées du processus visuel primaire et permet une stratégie de contrôle efficace. L'attention balaye simplement la carte de saillance dans un ordre décroissant de saillance. Cette méthode est totalement guidée par les données, et ne prend pas en compte une connaissance préliminaire de la scène. Il s'agit donc d'un mécanisme "bottom-up".

La figure [12] décrit ce système. L'image est encodée par les neurones, au travers de la détection de caractéristiques pré-attentionnelles, en cartes de contrastes pour chacune des caractéristiques. Au sein de chaque carte de caractéristiques les neurones rivalisent spatialement pour la saillance. Les cartes de caractéristiques sont combinées ensuite pour obtenir la carte de saillance. La figure montre que la composante top-down peut aussi intervenir.

Le chapitre suivant décrit un modèle basé sur cette théorie [12].

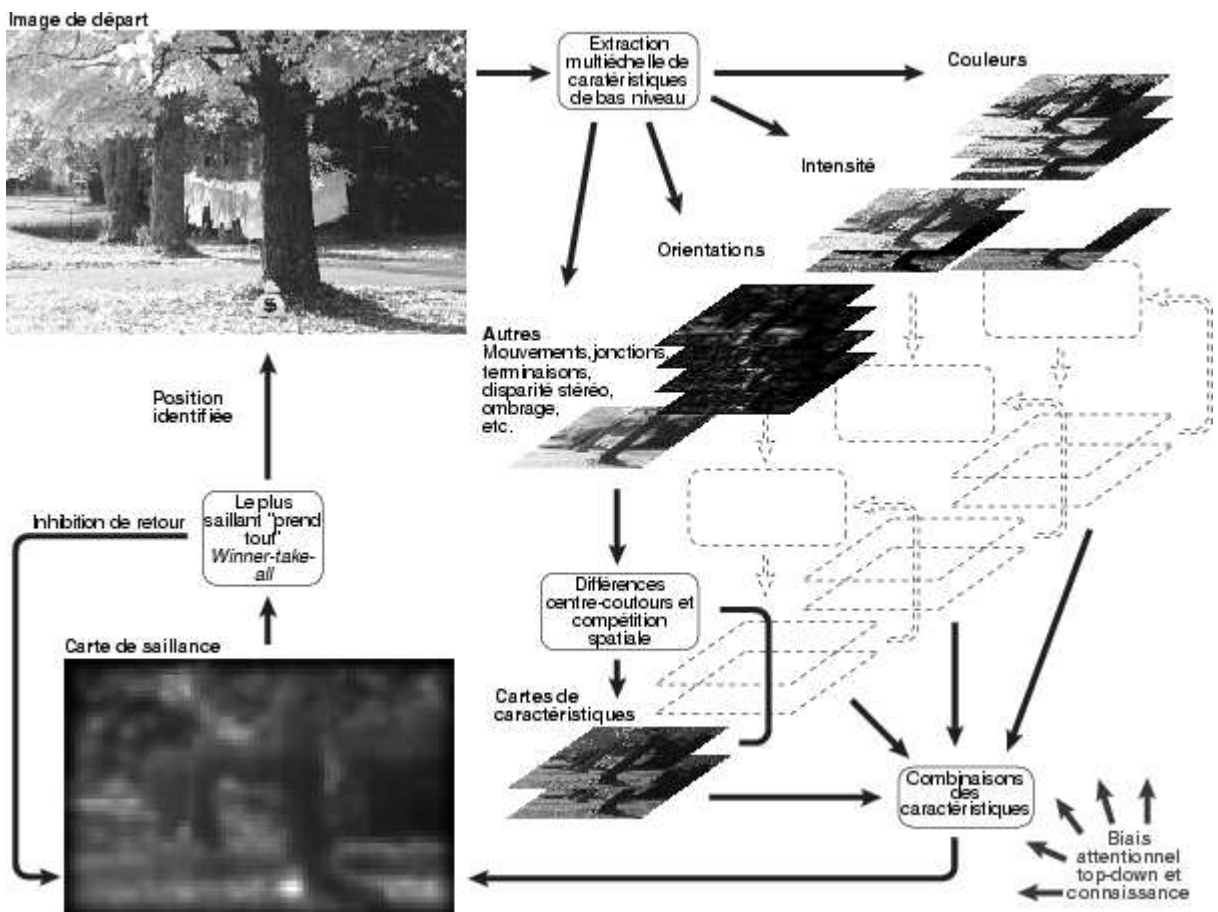


FIG. 3.5 – Système d'attention visuelle bottom-up, d'après [12]

Chapitre 4

Modèle d'attention visuelle basée sur la saillance

Ce stage de DEA a principalement porté sur l'étude et l'amélioration du modèle développé par Itti, Koch et Niebur [13], qui transforme l'image de départ en une carte exprimant l'intensité de la saillance en chaque point : la carte de saillance. Dans ce chapitre, la construction de ce modèle est expliquée. Elle est suivie d'une rapide évaluation.

4.1 Description du modèle

Comme avec la théorie de contrôle de l'attention bottom-up de Koch et Ullman [12], le modèle comporte trois étapes principales (figure 4.1). Tout d'abord les cartes de caractéristiques sont extraites, puis les cartes d'évidence, et pour finir la carte de saillance.

4.1.1 Extraction des cartes de caractéristiques

L'extraction de certaines caractéristiques d'une image se fait en calculant des cartes, dites de caractéristiques, qui représentent l'image de départ suivant une ou plusieurs caractéristiques bien définies. Ainsi, on obtient une représentation multi-caractéristiques de la scène. Chacune de ces cartes est calculée par un ensemble d'opérations, définies sous le terme "**centre - région contournante**". Ce nom a été donné pour la raison suivante : les neurones visuels sont plus sensibles à une petite région de l'espace visuel (le centre), alors que le signal présent autour (région contournante) inhibe la réponse neuronale (par exemple cellule ON-OFF). Par exemple, si une ligne verticale est entourée par des lignes horizontales, alors la réponse en cet endroit sera plus grande que si elle était entourée par des lignes verticales. L'opération "centre - région contournante" comporte deux étapes :

- Construction d'une représentation discrète multirésolution de l'image
- Calcul de la différence entre les niveaux fins et les niveaux grossiers de cette représentation multirésolution

La représentation discrète multirésolution d'une image est donnée par une série d'images de plus en plus petites, chacune étant la réduction de la précédente. Ce sous-échantillonnage peut introduire des problèmes d'aliasing. Ainsi, il est indispensable d'appliquer un filtre passe-bas, au moment du sous-échantillonnage, afin d'éviter que le spectre ne se replie sur lui-même. C'est un filtre gaussien qui est utilisé pour ce lissage, et la représentation multirésolution devient alors une pyramide gaussienne [2], définie comme suit :

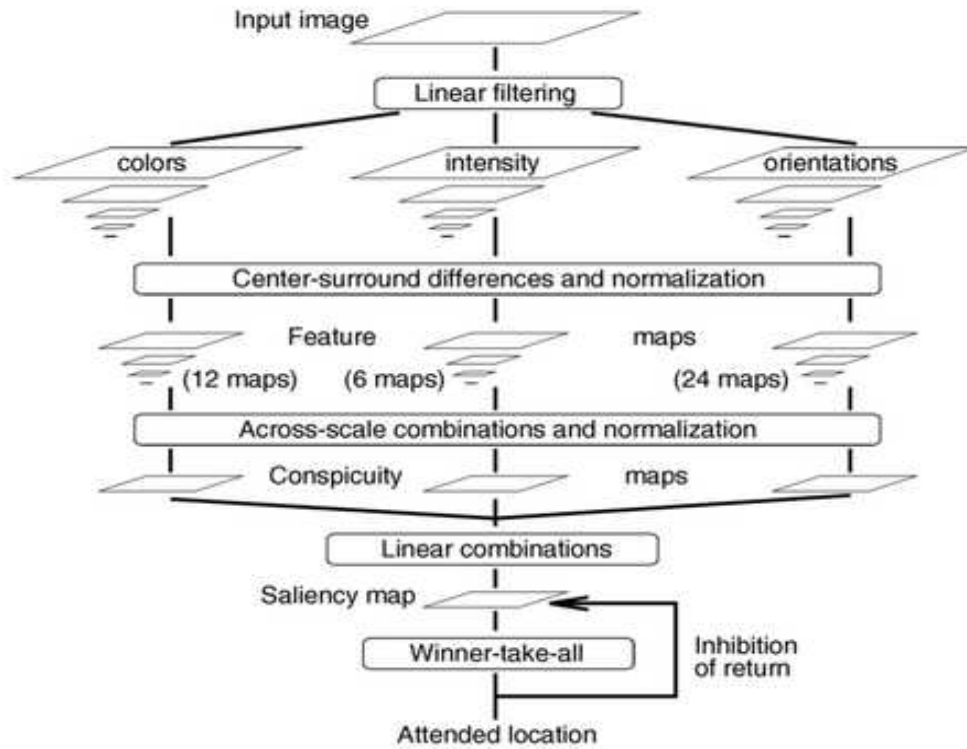


FIG. 4.1 – Architecture du modèle défini dans [13]

$$G_0(x, y) = I \quad , \quad G_{i+1}(x, y) = \frac{1}{2\pi\sigma} e^{-\frac{x^2+y^2}{2\sigma^2}} * G_i(x, y) \quad , \quad (4.1)$$

G_{i+1} étant une image plus petite que G_i . Chaque pixel de G_{i+1} est obtenu en calculant la moyenne des pixels autour dans G_i pondérés par une gaussienne. Si le rapport entre les deux images est 2 alors la pyramide est dite dyadique (figure 4.2 b)). C'est ce qui est utilisé dans le système présenté ici.

La différence "centre - région contournante" (figure 4.2 c)) est ensuite implémentée en calculant la différence entre les niveaux grossiers et ceux précis de la pyramide gaussienne. Dans le système décrit ici, le centre est un pixel au niveau $c \in \{2, 3, 4\}$, et la région contournante le pixel correspondant au niveau $s = c + \delta$, avec $\delta \in \{3, 4\}$. On ne se sert que des niveaux de 2 à 8. En effet, les niveaux 0 et 1 sont trop précis et pas assez lisses, tandis que les niveaux plus grands que 8 ou 10 ne sont plus assez significatifs. L'opérateur "centre - région contournante" sera noté " Θ " par la suite. Il est obtenu par une interpolation au niveau le plus fin et par une soustraction pixel par pixel des deux images. Il s'agit en fait d'une pyramide laplacienne de gaussiennes (Equation 4.2)).

$$L_i(x, y) = G_i(x, y) - Expand(G_j(x, y)) \quad , \quad (4.2)$$

avec Expand une interpolation bilinéaire. Les laplaciennes de gaussienne capturent raisonnablement bien le "chapeau mexicain" de la cellule rétinienne ganglionnaire.

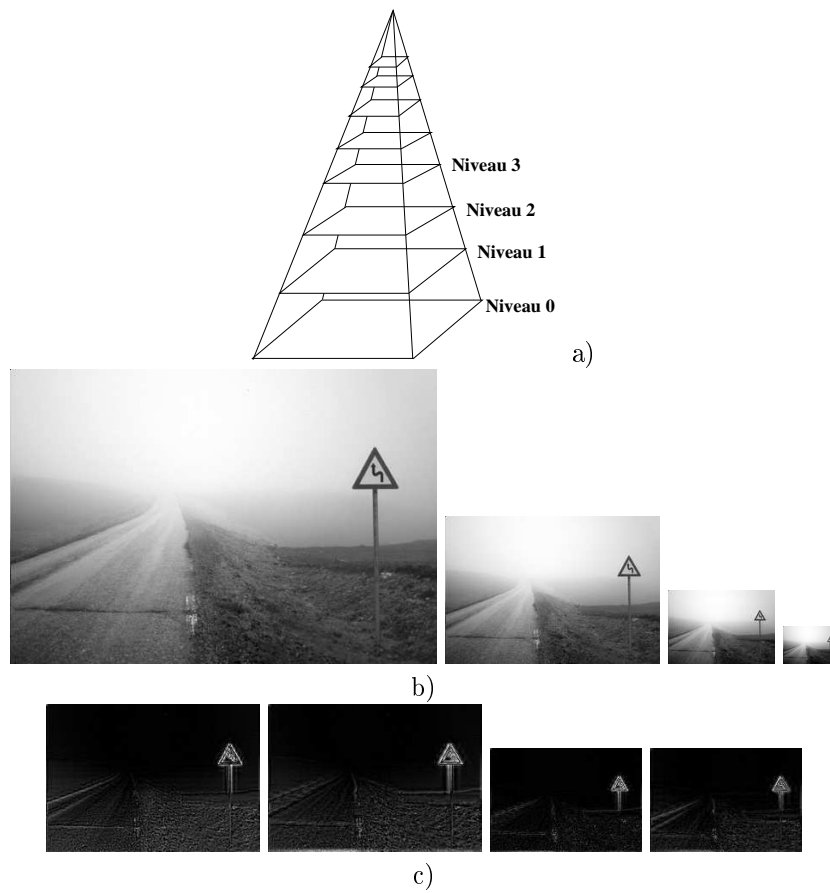


FIG. 4.2 – a) Structure pyramidale; b) Exemple d'une pyramide dyadique; c) Résultat obtenu avec l'opérateur "centre - région contournante" (pyramide laplacienne de gaussiennes)

Cette différence "centre - région contournante" est utilisée pour trois caractéristiques :

– **L'intensité :**

Soient r, g, b les canaux rouge, vert et bleu de l'image d'entrée. L'image d'intensité est obtenue avec l'équation $I=(r+g+b)/3$. A partir de I , une pyramide gaussienne $I(\sigma)$ avec $\sigma \in [0..8]$ est créée. $I(\sigma)$ est alors utilisée pour construire le premier ensemble de cartes de caractéristiques correspondant au contraste. Chez les mammifères, le contraste d'intensité est détecté par des neurones sensibles aux centres foncés sur régions contournantes claires ou aux centres clairs sur régions contournantes foncées. L'opérateur de différence "centre - région contournante" prend directement en compte les deux cas. Pour $c \in \{2, 3, 4\}$ et $s = c + \delta$, avec $\delta \in \{3, 4\}$, on obtient six cartes de caractéristiques $\mathcal{I}(c, s)$:

$$\mathcal{I}(c, s) = |I(c) \ominus I(s)| \tag{4.3}$$

Ce n'est que le module du signal qui est gardé afin de ne tenir compte que de la taille absolue du contraste d'intensité.

– **La couleur :**

Afin de découpler la teinte de l'intensité, les trois canaux r , g et b sont normalisés par l'intensité I . Quatre nouveaux canaux sont ensuite créés, R pour le rouge, G pour le vert, B pour le bleu et Y pour le jaune (équation 4.4). Les quatre pyramides gaussiennes $R(\sigma)$, $G(\sigma)$, $B(\sigma)$ et $Y(\sigma)$ sont finalement construites.

$$R = r - \frac{g+b}{2}, G = g - \frac{r+b}{2}, B = b - \frac{r+g}{2}, Y = \frac{r+g}{2} - \frac{|r-g|}{2} - b \quad (4.4)$$

Les cartes de caractéristiques obtenues à partir de ces quatre pyramides gaussiennes représentent la double opposition de couleurs du système visuel humain (rouge vs vert et bleu vs jaune). On obtient douze cartes de caractéristiques pour la couleur : $\mathcal{RG}(c, s)$ pour l'opposition rouge/vert et vert/rouge, et $\mathcal{BY}(c, s)$ pour l'opposition bleu/jaune et jaune bleu.

$$\mathcal{RG}(c, s) = |(R(c) - G(c)) \Theta (G(s) - R(s))| \quad (4.5)$$

$$\mathcal{BY}(c, s) = |(B(c) - Y(c)) \Theta (Y(s) - B(s))| \quad (4.6)$$

– **L'orientation :**

Les cartes de caractéristiques pour l'orientation sont obtenues à partir de l'image d'intensité I en utilisant des pyramides de Gabor $O(\sigma, \theta)$, où σ représente le niveau dans la pyramide et θ l'orientation choisi. La pyramide est calculée en convoluant chaque image de la pyramide gaussienne d'intensité avec un filtre de Gabor (Equation 4.7). Les auteurs ont choisi d'utiliser quatre orientations $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$.

$$O(\sigma, \theta) = Ga(\theta, \sigma) * I \quad (4.7)$$

On obtient 24 cartes de caractéristiques $O(c, s, \theta)$:

$$O(c, s, \theta) = |O(c, \theta) \Theta O(s, \theta)| \quad (4.8)$$

4.1.2 Extraction des cartes d'évidence

La carte de saillance donne les points saillants de l'image, chacun ayant une activité plus ou moins forte selon l'importance du point. Elle se calcule en combinant les cartes de caractéristiques obtenues précédemment. Une des difficultés pour combiner les 42 cartes de saillance est qu'elles représentent des données a priori non comparables, à des échelles différentes. Des objets saillants dans seulement quelques cartes peuvent être masqués par le bruit ou par d'autres objets moins saillants des autres cartes. Les auteurs proposent une méthode de normalisation pour résoudre ce problème, en introduisant l'opérateur de normalisation $\mathcal{N}(\cdot)$. Cet opérateur favorise les cartes contenant un petit nombre de pics de forte intensité, et supprime globalement celles qui possèdent un très grand nombre de pics de valeurs très proches (figure 4.3). Voici la méthode de calcul de $\mathcal{N}(\cdot)$ sur une carte de caractéristiques :

1. Normaliser les valeurs de la carte dans l'intervalle fixe $[0..M]$.
2. Trouver les coordonnées du maximum global dont l'activité vaut M , et calculer la moyenne \bar{m} de tous ses autres maxima locaux.
3. Multiplier globalement la carte par $(M - \bar{m})^2$.

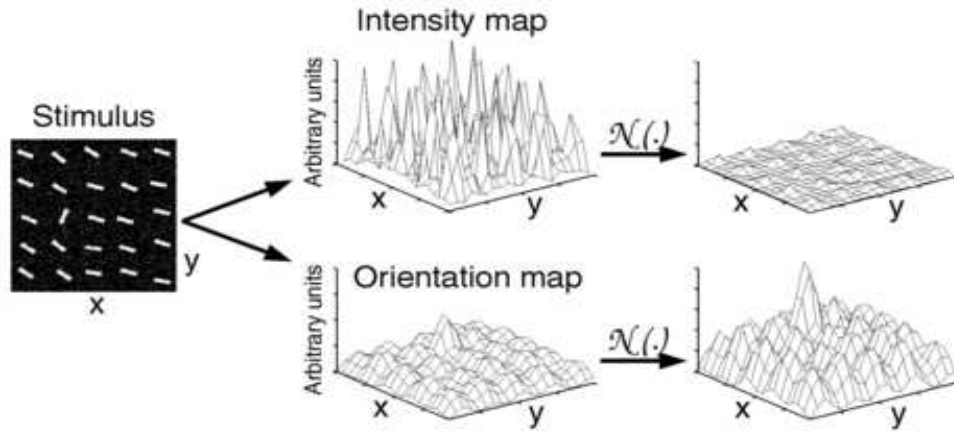


FIG. 4.3 – Opérateur de normalisation.

Comparer l'activité maximale sur toute l'image avec la moyenne sur tous les maximums locaux mesure la différence entre l'activité de la zone la plus active et la moyenne des autres zones. Quand, cette différence est grande, la zone la plus active ressort. Quand, par contre la différence est faible, la carte ne contient rien de particulier à détecter, et est quasi-supprimée. L'opérateur $\mathcal{N}(\cdot)$ est appliqué à toutes les cartes de caractéristiques.

Pour pouvoir combiner les cartes de caractéristiques normalisées malgré les différentes échelles, elles sont toutes ramenées au niveau 4. Trois cartes "d'évidence" sont alors créées : $\bar{\mathcal{I}}$ pour l'intensité (équation 4.9), $\bar{\mathcal{C}}$ pour la couleur (équation 4.10) et $\bar{\mathcal{O}}$ pour l'orientation (équation 4.11). L'opérateur " \oplus " calcule cette réduction au niveau 4 des cartes et les somme point par point.

$$\bar{\mathcal{I}} = \bigoplus_{c=2}^4 \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{I}(c, s)) \quad (4.9)$$

$$\bar{\mathcal{C}} = \bigoplus_{c=2}^4 \bigoplus_{s=c+3}^{c+4} [\mathcal{N}(\mathcal{RG}(c, s)) + \mathcal{N}(\mathcal{BY}(c, s))] \quad (4.10)$$

$$\bar{\mathcal{O}} = \sum_{\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}} \mathcal{N} \left(\bigoplus_{c=2}^4 \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{O}(c, s, \theta)) \right) \quad (4.11)$$

4.1.3 Extraction de la carte de saillance et des points saillants

La carte de saillance peut maintenant être obtenues en faisant la moyenne des trois cartes (équation 4.12).

$$\mathcal{S} = \frac{1}{3} (\mathcal{N}(\bar{\mathcal{I}}) + \mathcal{N}(\bar{\mathcal{C}}) + \mathcal{N}(\bar{\mathcal{O}})) \quad (4.12)$$

A partir de cette carte de saillance, les points d'attention visuelle peuvent être extraits. Les auteurs utilisent une méthode "Winner Take All". Le pixel de valeur maximum définit le point le plus saillant de l'image, celui vers lequel l'attention du regard se porte. Comme avec la vision humaine, on peut considérer qu'une zone autour de ce point a aussi été "vue" au cours de cette première attention. Cette zone est donc remplacée par un disque noir et ne sera plus visitée. Une nouvelle image est obtenue sur laquelle on répète de nouveau l'opération. Le résultat final est une liste de coordonnées des points saillants.

4.2 Évaluation de la méthode

Afin de pouvoir évaluer quantitativement la méthode décrite, il aurait fallu pouvoir la comparer avec les points d'attention visuelle humaine sur un ensemble d'images tests. Cependant, nous n'avons pas eu la possibilité de collecter des données sur des humains pendant ce stage. Une méthode permettant de comparer les coordonnées humaines aux coordonnées artificielles est décrite dans [15]. Les conclusions de cet article ont validé le modèle informatique sur des images tests principalement de type panneaux de la route. Néanmoins, le nombre de tests effectué par les auteurs étant encore largement insuffisant, il ne leur a pas été possible de tirer une conclusion définitive quant à la possibilité que ce système informatique soit une bonne copie du système d'attention visuelle humaine.

Nous avons donc essayé de donner une évaluation uniquement basée sur les coordonnées des points saillants obtenus, et sur l'application visée : la détection de classes d'objets. Ce qui nous intéresse ici est d'obtenir les zones de l'image permettant de classer les objets présents. Par exemple, nous souhaiterions plus tard appliquer la méthode à la détection de visage, pour laquelle il est important que les points saillants correspondent aux yeux, au nez, à la bouche... Les divers tests effectués sur la méthode présentée ici ont été souvent très décevants sur la classe visage, et les résultats obtenus (voir chapitre 6 pour des résultats) ne nous ont pas semblé exploitables pour une quelconque phase d'apprentissage et de reconnaissance. Ainsi, dans la plupart des cas seuls les yeux ont été détectés et même parfois aucune zone intéressante de l'image n'a été trouvée. Par contre, les tests effectués sur des images simples contenant des panneaux de la route sont apparus très concluants. Cela est facilement explicable car les panneaux de la route ont des couleurs qui ressortent beaucoup ainsi que des orientations bien définies ce qui les met bien en valeur. Vous trouverez quelques résultats obtenus avec le système décrit précédemment dans le chapitre 6, figure 6.2. Il semble que le type de performance obtenu par le modèle présenté dans ce chapitre, ne dépend que d'un seul facteur : Seuls les objets dont au moins une des caractéristiques se dégagent bien du reste de l'image auront une forte saillance. C'est le cas pour les panneaux de la route où les couleurs et les orientations utilisées se différencient bien du fond. Cela pose des problèmes pour les visages si le contraste entre les cheveux et le fond est trop important.

En conclusion, les performances obtenues par le système d'attention visuelle basé sur la saillance valident assez bien l'idée qu'une unique carte de saillance, créée à partir de plusieurs caractéristiques de l'image, peut guider l'attention visuelle. Un des avantages d'une telle méthode est qu'elle est assez facilement parallélisable, ce qui pourrait permettre un calcul temps réel. Cependant, sur de nombreux objets, les tests effectués ne sont pas assez satisfaisants, en particulier pour une application telle que la détection de classes d'objet. Lorsqu'aucune des trois caractéristiques d'un objet ne ressort assez, il n'apparaît pas saillant. Le système nécessite donc des modifications dont quelques possibilités sont présentées dans le chapitre suivant.

Chapitre 5

Analyse et améliorations du modèle

Le but du stage de DEA était d'étudier et d'améliorer le système décrit au chapitre précédent, en particulier en le rendant multi-échelle. Un nouveau système ayant pour application principale la détection d'objets devait en être déduit. La première partie du travail a donc consisté à comprendre et implémenter la méthode d'Itti. Il a ensuite été possible de l'analyser et de lui apporter des modifications. Ce chapitre détaille les différentes améliorations apportées. Pour pouvoir quantifier les résultats, il aurait fallu être capable de les comparer avec ce que donne l'attention humaine, ce qui n'a pas été possible. Nous nous contentons donc ici de comparer, pour chaque modification, les points saillants obtenus avec la méthode initiale avec ceux de la méthode corrigée. Quelques résultats des modifications les plus importantes sont présentés. Comme dans de nombreux problèmes de vision, certaines modifications dépendent du type d'image à traiter. Ainsi, quand cela est possible, nous avons essayé de conclure sur la meilleure méthode à utiliser pour tel ou tel type d'images.

Pour toutes les images présentées les points de saillance les plus forts correspondent aux cercles les plus clairs. Plus le cercle est foncé, moins il attire le regard.

5.1 Système multi-échelle

Le système décrit par Itti ne prend en compte qu'une seule résolution pour le calcul des points saillants de l'image. En effet, plusieurs niveaux de détails de chaque pyramide laplacienne sont utilisés pour calculer les cartes d'évidence et la carte de saillance, mais ces cartes sont à une échelle fixe (niveau 4 des pyramides). Cependant, les cartes de caractéristiques montrent que les points les plus saillants à une résolution ne sont pas forcément les plus saillants à d'autres. Par exemple, une petite zone attractive à un niveau de détail élevé pourra ne pas être trouvée en n'interpolant les images qu'au niveau 4 des pyramides, d'autant plus que la pyramide gaussienne utilisée est dyadique.

Afin de prendre en compte différents niveaux de détails pour l'obtention des points saillants, nous avons rendu le système décrit au chapitre précédent multi-échelle. Les cartes de caractéristiques sont calculées pour différentes valeurs de " c' " telles que $c \in \{c', c' + 1, c' + 2\}$ (figure 5.1). Ainsi, plusieurs cartes d'évidence, et donc plusieurs cartes de saillance, sont obtenues pour chaque caractéristique, en interpolant les images au niveau $c'+2$ et en les soustrayant pixel par pixel (opérateur \oplus) :

$$\bar{\mathcal{I}}(c') = \bigoplus_{c=c'+2}^{c'+4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{I}(c, s)) \quad (5.1)$$

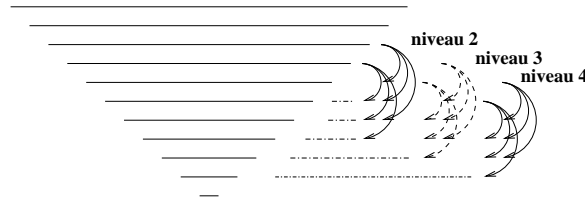


FIG. 5.1 – Mise en place du système multi-échelle.

$$\bar{\mathcal{C}}(c') = \bigoplus_{c=c'+2}^{c'+4} \bigoplus_{s=c+3}^{c+4} [\mathcal{N}(\mathcal{RG}(c, s)) + \mathcal{N}(\mathcal{BY}(c, s))] \quad (5.2)$$

$$\bar{\mathcal{O}}(c') = \sum_{\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}} \mathcal{N} \left(\bigoplus_{c=c'+2}^{c'+4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{O}(c, s, \theta)) \right) \quad (5.3)$$

$$\mathcal{S}(c') = \frac{1}{3} (\mathcal{N}(\bar{\mathcal{I}}(c')) + \mathcal{N}(\bar{\mathcal{C}}(c')) + \mathcal{N}(\bar{\mathcal{O}}(c'))) \quad (5.4)$$

Le mécanisme "Winner Take All" permet ensuite de trouver un ensemble de points saillants sur chacune de ces cartes de saillance. Pour combiner ces différents ensembles, acquis à différentes résolutions, on parcourt la pyramide de cartes de saillance de haut en bas (le niveau le plus haut correspondant à l'image la plus grande), et on regarde combien de correspondance dans les autres images possèdent chaque point saillant (figure 5.2). Comme les ensembles ont été acquis pour des résolutions différentes, on interpole toutes les coordonnées à l'image la plus grande. On admet que les points dans différentes cartes de saillance se correspondent si leurs coordonnées interpolées ne sont pas distantes de plus de cinq ou six pixels. Finalement, on considère que les points les plus saillants globalement sont ceux ayant le plus de correspondances dans les autres cartes, c'est à dire ceux qui sont les plus saillants quelle que soit la résolution. Lorsque le nombre de correspondances est identique pour plusieurs points saillants, le plus saillant est le point dont la moyenne de l'intensité sur l'ensemble des cartes de saillance est la plus grande.

La figure 5.3 présente un des nombreux résultats obtenus. Sur cette figure, on remarque qu'en utilisant le système initial, aucune zone du visage n'est détectée. Avec seulement une résolution supplémentaire, les deux yeux sont détectés, et en rajoutant encore une de plus, la bouche est trouvée. Il est difficile de dire si ces parties du visage sont également des zones attractives pour le regard humain. Néanmoins, on peut conclure qu'à des fins de reconnaissance des formes, le système multi-échelle peut nettement améliorer les résultats. En effet, il est impossible de dire, en utilisant une méthode d'attention visuelle, qu'un objet appartient à la classe visage si les yeux et la bouche ne sont pas détectés.

Il est important de noter que même si parfois les résultats sont restés inchangés, le modèle ne les a jamais dégradés. Ce système multi-échelle est particulièrement utile pour des images contenant de nombreuses zones saillantes dans l'image. Il permet de faire ressortir celles qui sont les plus attractives à différentes résolutions. Les pyramides gaussiennes étant dyadiques, il n'est pas possible d'utiliser plus des 4 résolutions pour ce système. En effet au-dessus du niveau 10 des pyramides gaussiennes, les images deviennent non significatives. Nous verrons un peu plus loin dans ce chapitre qu'il est possible d'utiliser une pyramide non-dyadique qui permettra de prendre en compte plus de résolutions pour le système multi-échelle final.

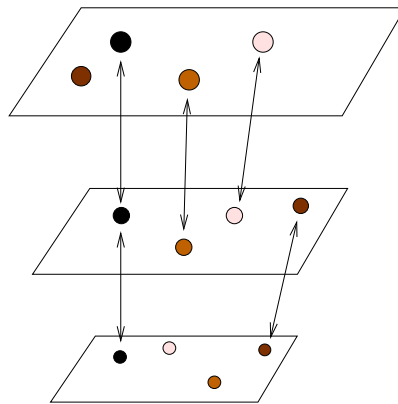


FIG. 5.2 – Correspondance entre des points saillants à différentes échelles.



FIG. 5.3 – Résultats obtenus avec le système multi-échelles a) système initial; b) Utilisation de 2 échelles $c'=2,3$; b) Utilisation de 3 échelles $c'=2,3,4$; b) Utilisation de 4 échelles $c'=2,3,4,5$.

5.2 Cartes de caractéristiques

5.2.1 Espaces de couleurs

Comme il a été décrit dans le chapitre 3, les couleurs peuvent être décrites par différents espaces. Nous nous sommes interrogés sur le meilleur à utiliser dans le cadre de ce projet. Itti a choisi d'utiliser le mode $RGB_{normalise}$, mais est-il vraiment meilleur que l'espace RGB? De plus, la CIE a mis en place le système La^*b^* , qui prend plus en compte les facteurs physiologiques de la perception de la couleur par l'oeil humain (section 3.2).

Les couleurs complémentaires (rouge, vert, bleu) associées produisent le meilleur contraste car, combinées les unes aux autres, elles donnent le blanc. Ainsi, pour des images où c'est principalement le contraste qui fait ressortir les zones intéressantes, l'espace RGB serait le meilleur à utiliser. Par exemple, pour les yeux et le nez d'un portrait, le contraste et l'orientation sont les caractéristiques les plus importantes.

Les panneaux de la route, eux, sont quasiment uniquement composés des couleurs rouge, vert, bleu, jaune, noir ou blanc. Pour bien faire ressortir ces couleurs du reste de l'image, il est intéressant de séparer la teinte de l'intensité, qui varie beaucoup avec l'illumination. Cet espace de couleurs permet de bien mettre en valeur les objets dont la teinte se distingue clairement du fond. Ainsi, il faudrait utiliser le système $RGB_{normalise}$ pour la détection des panneaux de la route. Une grande majorité des images prises dans la nature contient de grandes zones de couleurs quasi uniformes et assez pures (souvent vert, bleu...), et de contraste d'intensité assez varié. Ce sont alors les objets qui se différencient de part leurs teintes qui attirent le regard en premier. En effet, il est clair qu'un objet de couleur verte dans l'herbe sera difficilement repérable, quelle que soit son orientation ou sa luminance. Pour ce type d'image, l'espace $RGB_{normalise}$ paraît donc être le plus adapté.

Lorsque l'ensemble des couleurs composant une image est très varié, l'espace qui convient le mieux est La^*b^* . En effet, non seulement il différencie bien la teinte de l'intensité mais il couvre aussi l'intégralité du spectre visible et le représente de manière uniforme. Les images de piétons prises en ville sont justement généralement très colorées. Les points saillants seraient donc meilleurs avec cet espace.

Afin de valider ces remarques, nous avons comparé les points saillants obtenus pour chacun de ces trois espaces, sur un grand nombre d'images. Quelques-uns sont montrés et détaillés en annexe A. Pour la majorité des images, les résultats obtenus illustrent bien les analyses précédentes. Par exemple, nous avons vérifié qu'il est préférable d'utiliser l'espace $RGB_{normalise}$ lorsqu'on traite des images du type panneaux de la route (panneaux toujours trouvés), RGB lorsqu'il s'agit de portraits (yeux toujours détectés). Les conclusions obtenues sur l'espace de couleur à utiliser suivant le type d'image en entrée sont résumés dans le tableau suivant.

Type d'images	Meilleur système de couleurs
Panneaux de la route	RGB normalisé
Portraits	RGB
Portraits d'animaux	La^*b^*
Nature	RGB normalisé
Piétons	La^*b^*

TAB. 5.1 – Meilleurs systèmes de couleur suivant le type d'image testée.

Pour une grande partie des images testées, l'espace de couleurs a véritablement influé sur la position des points saillants. Il serait donc intéressant que notre système final prenne en compte le type d'image en entrée afin de pouvoir choisir l'espace de couleur qui convient.

5.2.2 Nombre d'orientations

Les primates sont sensibles à une douzaine d'orientations, les humains à environ 20. Le système décrit dans le chapitre précédent n'en prend en compte que 4. En effet, seules les orientations 0, 45, 90 et 135° ont été détectées jusqu'à présent. Cependant, de nombreux objets de notre environnement présentent des formes complexes, contenant un nombre plus grand d'orientations. Nous nous sommes donc interrogés sur la possible influence d'une telle perte d'information.

Rappel :

Il y a dualité entre domaine spatial et fréquentiel : La largeur de la gaussienne d'un filtre de Gabor dans le domaine fréquentiel, et sa largeur dans le domaine spatial sont inversement proportionnelles. Ainsi si la bande passante dans le domaine spatial est grande alors elle est faible dans le domaine fréquentiel. Un filtre de Gabor est le produit d'une sinusoïde, complexe par une enveloppe gaussienne 2D (équation 5.5).

$$h(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \cdot e^{2\pi i(Ux+Vy)} \quad , \quad (5.5)$$

avec $x' = x.\cos(\theta) + y.\sin(\theta)$ et $y' = y.\cos(\theta) - x.\sin(\theta)$, σ la largeur de la gaussienne. $F^2 = U^2 + V^2$ définie la fréquence centrale, et $\theta = \arctan \frac{V}{U}$ l'orientation. Sa transformée de Fourier s'écrit :

$$H(u, v) = e^{-2\pi^2\sigma^2[(u'-U')^2+(v'-V')^2]} \quad , \quad (5.6)$$

avec $u' = u.\cos(\theta) + v.\sin(\theta)$ et $v' = v.\cos(\theta) - u.\sin(\theta)$. Les coordonnées (U',V') subissent la même rotation que le centre (U,V). Ainsi, H(u,v) est un filtre passe-bande gaussien dont le petit axe est orienté d'un angle Ω par rapport à l'abscisse u, tandis que h(x,y) est une sinusoïde complexe modulée par une fonction gaussienne 2D, et orienté d'un angle Ω par rapport à l'abscisse x. La direction de représentation d'une ligne dans le domaine spatial et celle dans le domaine fréquentiel sont donc orthogonales. La largeur de la bande passante est donnée par :

$$B = \log\left[\frac{\pi F\sigma + \alpha}{\pi F\sigma - \alpha}\right] \quad , \quad (5.7)$$

avec $\alpha = \sqrt{\frac{\ln 2}{2}}$.

Les résultats d'une convolution par un filtre de Gabor dépendent de la bande passante des filtres. En effet, si la bande passante est large alors des contours qui ne sont pas complètement orientés dans le sens du filtre pourront quand même être détectés car le résultat de la convolution ne sera pas nul (figure 5.4). Par contre si la bande passante est très fine alors une convolution par un filtre orienté à 0° ne détectera que les contours très proches de 0°. Ainsi, un nombre d'orientation plus grand que 4 n'est utile que si la bande passante des filtres est petite. Le coût de la convolution par un filtre de Gabor étant vraiment conséquent il est préférable que le nombre d'orientation utilisé soit faible.

Nous avons donc choisi une bande passante assez élevée et vérifié si n'utiliser que quatre orientations fait varier ou non sur les résultats, en comparant les résultats ma méthode initiale avec huit

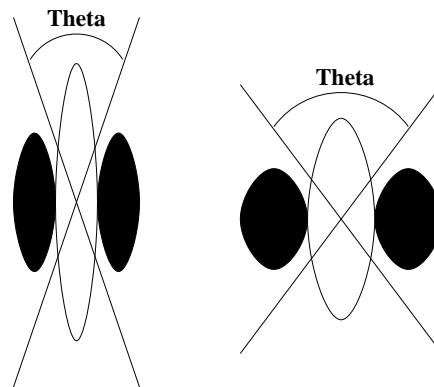


FIG. 5.4 – Taille de l'angle définissant le nombre d'orientations pouvant être détectées par le filtre de Gabor suivant la largeur de la bande passante

orientations ($\theta \in \{180 * i/8, i \in [0..7]\}$) puis douze ($\theta \in \{180 * i/12, i \in [0..11]\}$) (voir Annexe B pour des exemples de résultats). Comme souhaité, les coordonnées des points saillants n'ont quasiment pas changés pour un nombre d'orientation plus important. On peut alors se demander si deux, ou même une seule, orientations ne suffiraient pas. Cela revient à ne prendre vraiment en compte que les contours quasi verticaux et horizontaux de l'image, ce qui n'a pas grand intérêt puisqu'il est très rare que des objets du monde réel se limite à ces deux orientations, ou alors, à augmenter encore la bande passante, ce qui rend trop floue la carte de caractéristique d'orientation et ne fait pas assez ressortir les points saillants. Les tests ont bien montrés que seulement deux orientations, même en choisissant une bande passante plus grande dégrade les résultats. Ainsi, il semble que 4 orientations avec une bande passante assez élevée soit le meilleur compromis entre vitesse de calcul et qualité des résultats.

5.2.3 Pyramide gaussienne

Les pyramides gaussiennes calculées dans la méthode initiale sont dyadiques et les calculs ne sont effectués qu'à partir du niveau 2, les niveaux 0 et 1 n'étant pas assez lisses. Ceci entraîne deux défauts majeurs. D'une part, le changement entre deux niveaux de ces pyramides est brutal puisque le rapport entre les tailles de deux images est deux. D'autre part, le niveau 2 est une image 4 fois plus petites que l'image originale donc de résolution assez faible, ce qui peut faire perdre un certain nombre de détails dans les images. Afin de résoudre ces deux problèmes, l'idée a été d'utiliser des pyramides non dyadiques pour lesquelles la différence entre deux niveaux est plus faible, en rajoutant de un à trois niveaux intermédiaires entre chaque niveau des pyramides dyadiques initiales. Le rapport entre deux niveaux de la nouvelle pyramide est alors $2^{\frac{1}{i}}$, avec $i \in \{1, 2, 3, 4\}$.

L'algorithme de calcul des ces niveaux intermédiaires est le suivant : Tout d'abord la pyramide dyadique est calculée, puis chaque niveau intermédiaire est construit à partir de l'image qui le précède dans la pyramide dyadique (figure 5.5). Cette méthode diminue le temps de calcul et évite la transmission d'erreurs de niveau en niveau.

Pour le calcul des pyramides laplaciennes, nous avons les mêmes valeurs qu'initialement, c'est à dire : $s = c + \delta$, avec $\delta \in \{3, 4\}$, qui sont tous des niveaux assez lisses. Ainsi, l'image au niveau 4 de la pyramide laplacienne est plus grande qu'avec la méthode originale, et contient donc plus de détails. Ceci est particulièrement utile pour des portraits où de nombreux points saillants de tailles assez faibles sont présents. Pour la plupart des tests effectués sur des images de visages du type photomaton, les deux yeux et la bouche ont été détectés avec quatre niveaux intermédiaires, alors que cela était plus rare avec une pyramide dyadique. Au contraire, pour des images prises dans la nature, avec les objets saillants assez loin et gros, les points d'attention sont souvent plus grossiers et présentent un

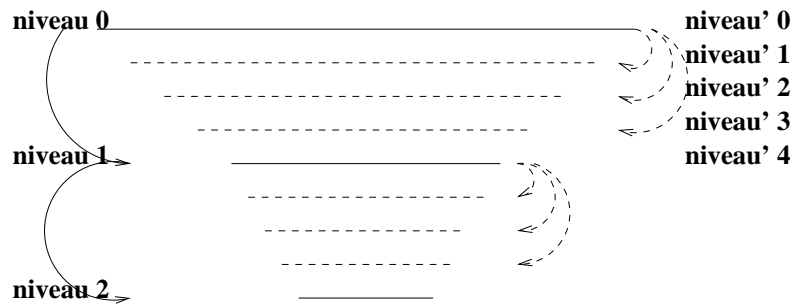


FIG. 5.5 – Construction des niveaux intermédiaires à partir d’une pyramide dyadique.

contraste important avec le fond, ce qui les rend très visible à faible résolution. Dans ce cas une pyramide dyadique est suffisante. En effet, sur les images testées, en rajoutant des niveaux intermédiaires les zones saillantes intéressantes sont encore détectées, mais d’autres points sont aussi rajoutés (pas toujours nécessaires comme une petite zone de ciel bleu au milieu de branches d’arbre).

La figure 5.6 présente un des résultats obtenus. On peut voir qu’en n’utilisant qu’une pyramide dyadique, seul le nez est détecté, tandis qu’avec quatre niveaux intermédiaires les yeux sont également trouvés, et qui plus est, ces trois zones (2 yeux et nez) sont les premières focalisées. Encore une fois, pour une application à la détection d’objets, les yeux, le nez et les oreilles sont les zones les plus importantes à trouver. Il apparaît donc que quatre niveaux intermédiaires est la meilleure solution. Cela s’explique ainsi : l’image présentée contient de nombreux détails. Or les zones intéressantes à détecter sont assez petites et moins contrastées en intensité et couleur que, par exemple, les poils blancs sur le support noir en bas de l’image. En utilisant une carte de saillance à un niveau de détail trop faible, ce sont les grandes zones de l’image de départ où des contours forts sont présents qui sont les plus saillantes. En effet, le blob du laplacien est trop grand par rapport à la taille de la zone saillante que l’on aurait souhaité trouvé (oeil).

Le temps de calcul étant d’autant plus long que le nombre de niveaux intermédiaires est grand (calculs faits sur des images plus grandes), on préfère souvent travailler avec des pyramides dyadiques quand cela est possible. Cependant, pour ne pas rajouter au système final le choix du nombre intermédiaire à utiliser suivant le type d’image en entrée, nous avons décidé d’utiliser quatre niveaux intermédiaires pour toutes les images. Cela n’affecte pas les résultats car, si les zones intéressantes à détecter sont trouvées avec une pyramide dyadique, elles le sont aussi avec quatre niveaux intermédiaires. De plus, il est très difficile de tirer une conclusion sur le nombre de niveaux à utiliser suivant le type d’image à traiter, ce nombre dépendant plus de la taille des zones saillantes souhaitées et de la force des contours que du type d’image lui-même.

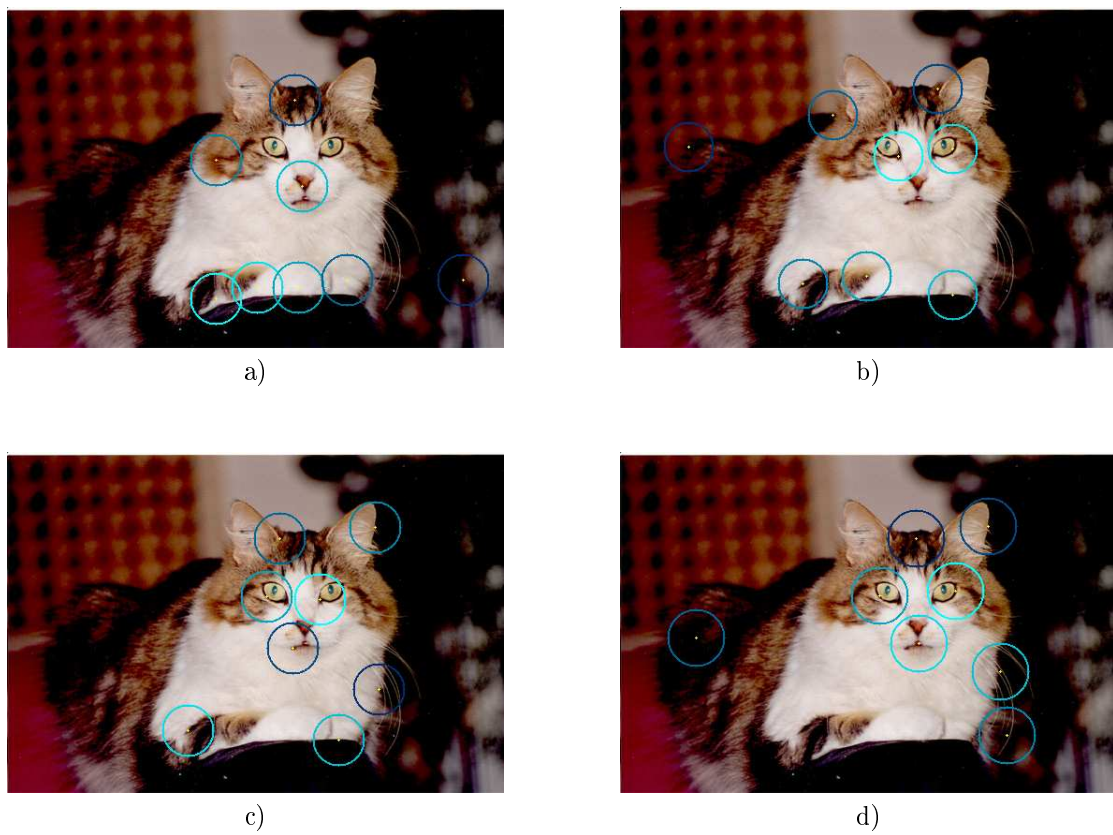


FIG. 5.6 – Résultats obtenus avec a) une pyramide dyadique, b) 1 image intermédiaire, c) 2 images intermédiaires, d) 3 images intermédiaires.

5.3 Carte de saillance

5.3.1 Combinaison des cartes d'évidence

Le système décrit dans le chapitre précédent utilise trois caractéristiques différentes : le contraste d'intensité, la couleur et l'orientation. La carte de saillance est tout simplement définie comme étant la moyenne des trois cartes d'évidence. Nous nous sommes demandés si la présence de ces trois caractéristiques était vraiment indispensable, et si l'une d'entre elles ne devait pas avoir plus d'importance que les autres. Cette sous-section étudie successivement l'influence du contraste d'intensité, de la couleur puis de l'orientation. Dans chaque cas, nous avons retiré ou donné deux fois plus d'importance à la carte d'évidence correspondante pour le calcul de la carte de saillance. Quelques résultats des tests réalisés sont montrés en Annexe C.

Influence du contraste d'intensité :

L'étude de l'influence des cartes d'évidence commence donc par l'étude du contraste d'intensité. Cette caractéristique peut paraître indispensable. En effet un objet noir sur blanc attire le regard, quelle que soit son orientation. Pour les humains, ce sont les cellules ON-OFF de la rétine, se représentant très bien par les pyramides laplaciennes, qui joue ce rôle. Les points saillants dus au contraste sont en fait les contours forts de l'image transformée en niveau de gris.

Les tests ont montré qu'en l'absence de cette caractéristique, certains points saillants comme les yeux dans un visage ne sont pas trouvés. En fait, les yeux se caractérisent principalement par un fort changement de contraste et un changement d'orientation. En l'absence de la caractéristique de contraste, les couleurs et orientations prennent plus d'importance et d'autres points sont détectés. Cependant pour des images où le contraste est faible, retirer cette caractéristique améliore les résultats. On peut donc supposer que le signal a été sur-normalisé.

Au contraire, donner deux fois plus d'importance au contraste qu'aux deux autres caractéristiques ne changent quasiment pas les résultats. Ne pouvant pas comparer les coordonnées des points saillants obtenus avec celles qu'aurait trouvé un humain, il est très difficile de dire si la méthode est améliorée avec plus d'importance pour le contraste, les variations étant tellement minimes.

Influence de la couleur :

On peut penser que certaines zones d'une image ressortent grâce à un changement de couleur, sans qu'il n'y ait de gros changements de contraste d'intensité ni d'orientation. Ainsi la couleur serait une caractéristique très importante, voire la plus importante. Nous avons quand même souhaités observer son influence sur un certain nombre d'images (Annexe C.2). Ne pas considérer la caractéristique de couleur revient en fait à travailler sur une image en niveau de gris.

Sur les tests réalisés, les résultats sont apparus en général véritablement moins bons sans la caractéristique de couleur qu'avec. Encore une fois, dans le cas d'images peu contrastées en couleur, les résultats sont améliorés car la caractéristique d'orientation influe plus. Elle ne sera donc enlevée du système final que dans le cas d'images en niveau de gris. Lui donner plus d'importance dégrade également les résultats. Cela signifie que l'on ne peut pas considérer cette caractéristique comme plus importante que les deux autres.

Un test supplémentaire a été réalisé pour les couleurs. Comme expliqué dans la section 3.1, les cônes S correspondants au bleu sont nettement moins présents dans la rétine que les cônes M et L. Nous avons donc observé les résultats en donnant plus d'importance à la carte de caractéristique de l'opposition vert/rouge, rouge/vert qu'à celle de l'opposition bleu/jaune, jaune/bleu. Malgré ce que l'on aurait pu penser, ce changement n'a eu qu'une importance très minime, et plutôt néfaste, sur les

résultats. Les deux oppositions rouge/vert et bleu/jaune seront donc présentes dans le système final, et à niveau égal.

Influence de l'orientation :

Comme pour les couleurs, et le contraste d'intensité, nous nous sommes intéressés à l'influence de l'utilisation de la caractéristique orientation sur les résultats. Les différentes images testées ont presque toutes donné de meilleurs résultats en présence de la caractéristique d'orientation. Comme on pouvait s'y attendre d'après les remarques faites pour les deux autres caractéristiques, l'orientation est d'autant plus importante que les couleurs sont uniformes dans l'image. Le deuxième test a été surprenant. En effet, il est apparu que pour la plupart des images testées, donner deux fois plus d'importance à l'orientation donnent des positions bien meilleures pour les points saillants. Il est difficile de donner une justification sûre à ce phénomène. Cependant, une des raisons peut être que les deux autres caractéristiques mises ensemble amplifient trop l'importance du contraste d'intensité, même si le signal a été normalisé, voir sur-normalisé. Les changements d'orientations ne sont alors pas assez significatifs par rapport aux fortes variations de contraste.

Ainsi pour le système global final, l'orientation aura plus d'importance que les deux autres caractéristiques.

Conclusion : Les analyses précédentes peuvent se résumer ainsi :

- Ne pas négliger les caractéristiques d'intensité ou de couleur améliore les résultats sauf dans le cas d'images peu contrastées.
- La présence de l'orientation améliore les résultats surtout si les images sont peu contrastées.
- Donner plus d'importance au contraste ne change pas les résultats
- Donner plus d'importance aux couleurs dégrade les résultats
- Donner plus d'importance à l'orientation améliore nettement les résultats

5.3.2 Normalisation

La normalisation utilisée dans [13] est simple et a un faible coût de calcul. Cependant, elle présente des défauts que Itti a décrit dans sa thèse [11]. Tout d'abord la méthode n'est pas plausible biologiquement car elle nécessite le calcul d'un maximum global. Or, il a été montré que les neurones du cortex visuel ne sont connectés que localement. Ensuite, cette normalisation favorise une unique position (celle d'activité maximum). Il serait préférable que chaque carte de caractéristique représente un petit ensemble de points d'activités assez grandes. Enfin, la normalisation n'est pas robuste au bruit qui peut masquer certains maxima locaux.

Itti dans [11] propose une seconde méthode de normalisation, basée sur les connexions cortico-corticales du système visuel humain. Des études ont montré que les interactions entre le centre et la région contournante sont dominées par une composante inhibitoire de la région contournante vers le centre. Des plus, cette inhibition apparaît plus forte à une certaine distance du centre, et faible à une distance trop proche ou trop éloignée du centre. Ainsi il semble que ces interactions peuvent être modélisées par une différence de gaussiennes (figure 5.7).

L'implémentation de cette idée se fait de la manière suivante. Chaque carte de caractéristique est tout d'abord normalisée sur l'intervalle $[0..1]$, puis additionnée alors chaque carte de caractéristique après convolution par un filtre de différence de gaussiennes (équation 5.8). Ensuite, les résultats négatifs sont mis à zéro à chaque itération. Les cartes de caractéristiques sont sujettes à une dizaine d'itérations du processus décrit par l'équation 5.9, où \mathcal{M} est la carte de caractéristique et C_{inh} une constante d'inhibition. C_{inh} permet de supprimer des régions où l'inhibition et l'excitation sont quasi exactement équilibrées.

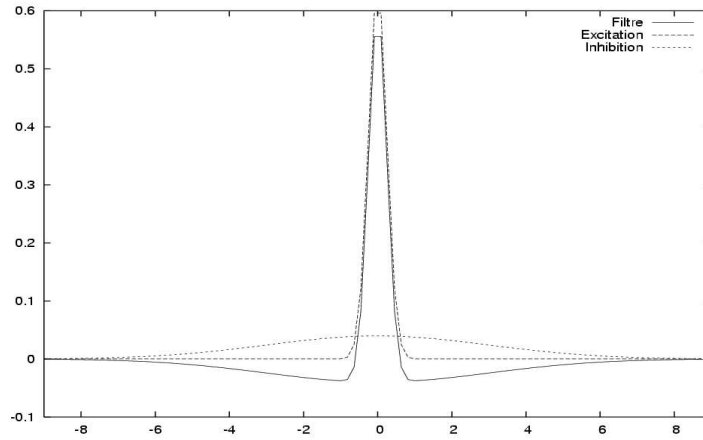


FIG. 5.7 – Filtre utilisé pour la normalisation itérative

$$DOG(x, y) = \frac{c_e x^2}{2\pi\sigma_e x^2} e^{-\frac{x^2+y^2}{2\sigma_e x^2}} - \frac{c_i nh^2}{2\pi\sigma_i nh^2} e^{-\frac{x^2+y^2}{2\sigma_i nh^2}} \tag{5.8}$$

$$\mathcal{M} \leftarrow |\mathcal{M} + \mathcal{M} * DOG - C_{inh}|_{\geq 0} \tag{5.9}$$

La figure 5.8 montre les variations sur la carte de saillance après 0, 2, 4 et 8 itérations. Sur cette figure, les zones uniformes de l’image sont successivement altérées sur la carte de saillance, devenant nulles après une dizaine d’itérations. Ainsi, aucun point de saillance ne sera détecté dans les zones uniformes de l’image. En effet, la normalisation par différence de gaussiennes permet d’éliminer itérativement les moyennes locales et de conserver seulement les fortes discontinuités.

Nous avons montré ce résultat sur une carte de saillance, mais ce sont chacune des cartes d’évidence qui sont sujettes à 10 itérations de cette normalisation par filtres gaussiens. Le choix du nombre d’itérations est quelque peu arbitraire. Avec un nombre infini d’itérations, chaque carte non nulle convergera vers un seul pic, constituant une représentation pauvre de la scène. Avec seulement quelques itérations, la compétition spatiale est faible et insuffisante (figure 5.8 b)), et la normalisation DOG n’est finalement pas très utile.

En conclusion, la normalisation itérative par filtres gaussiens améliore les résultats, en mettant bien en valeur les maxima locaux de l’image. Elle est particulièrement efficace dans le cas d’images contenant de grandes zones assez uniformes. Le défaut de cette méthode est son temps de calcul.

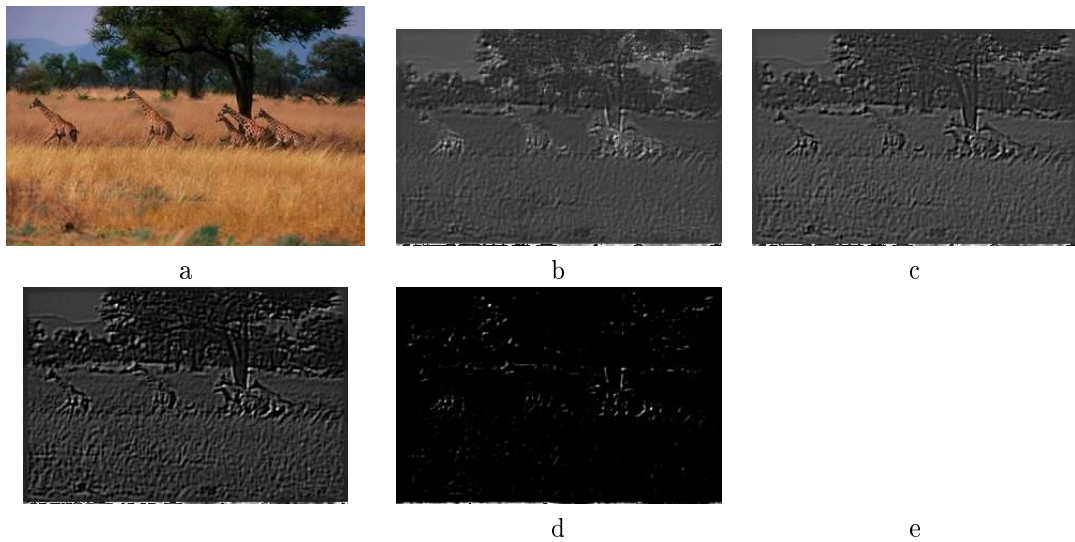


FIG. 5.8 – Résultat de la normalisation par différence gaussienne sur la carte de saillance a) image originale; b) 0 itération; c) 2 itérations d) 4 itérations; e) 8 itérations

Chapitre 6

Système d'attention visuelle multi-échelle

A partir des améliorations et conclusions du chapitre précédent, un nouveau système d'attention visuelle, basé sur celui décrit au chapitre 4, peut être construit. Les nouveautés par rapport au système initial sont les suivantes :

- Le système est multi-échelle. Il prend en compte différentes résolutions de l'image d'entrée pour le calcul des cartes d'évidence.
- Les pyramides gaussiennes sont non dyadiques, avec quatre niveaux intermédiaires, ce qui permet un changement moins brutal entre différentes résolutions et l'obtention des points saillants à des échelles plus élevées.
- Le mode de couleur est choisi suivant le type d'image à traiter. Par défaut, le mode choisi est $RGB_{normalise}$.
- L'orientation a deux fois plus d'importance que les deux autres caractéristiques lors du calcul des la carte de saillance.
- La normalisation est faite par un filtre de différence de gaussiennes, comme conseillé dans [11].

La structure de ce nouveau système est montrée figure 6.1.

La combinaison de toutes les améliorations étudiées au chapitre précédent a permis d'améliorer nettement les résultats sur une grande partie des images testées. C'est le cas images peu contrastées, pour lesquelles les modifications influant le plus sont l'utilisation de la normalisation par filtre de différence de gaussiennes, et l'importance deux fois plus grandes de la caractéristique d'orientation. Pour des images très détaillées ou lorsque les zones saillantes à trouver sont assez petites et que des contours forts sont présents, ce sont la méthode multi-échelle et les quatre niveaux intermédiaires de la pyramide gaussienne qui importent le plus. Lorsque les couleurs sont très variées, c'est l'utilisation de l'espace La^*b^* , tandis que si c'est le contraste qui prédomine pour les zones de saillance, c'est le mode RGB.

Quelques résultats significatifs des améliorations sont présentés sur la figure 6.2. Comme on peut le voir, ils sont tous de qualités égales ou supérieures à la méthode initiale. Nous rappelons au lecteur que les premiers points saillants trouvés par le système correspondent au cercles les plus clairs. Plus le cercle est foncé, moins il est saillant. Dans certains tests réalisés, ce ne sont pas les coordonnées des points saillants qui ont été modifiées mais l'ordre dans lequel ils sont trouvés. Le principal défaut de notre méthode par rapport au modèle initial est son temps de calcul. En effet, utiliser quatre niveaux intermédiaires, une méthode multi-échelle et la normalisation itérative par filte gaussiens augmente

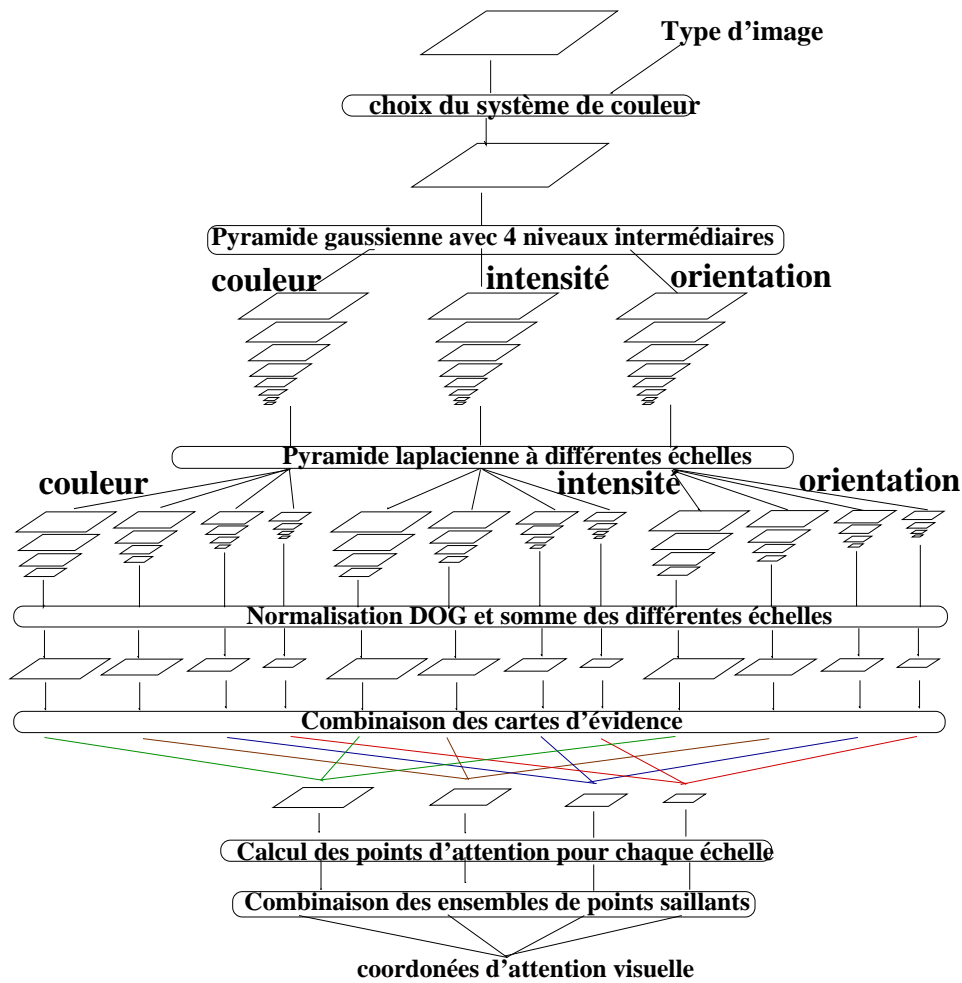


FIG. 6.1 – Architecture du nouveau modèle

de quelques minutes le coût de calcul. Cependant, pour l'instant, la qualité des résultats est importante que cette complexité.

Comme il n'a pas été possible de comparer les résultats de notre système avec ce qu'aurait détecté un humain, il ne nous est pas possible d'affirmer qu'il se rapproche plus du système qui dirige les fixations visuelles humaines que le système initial décrit par Itti dans [13]. Cependant, cela ne nous paraît pas si important car ce qui nous intéresse dans le cadre de ce projet est d'appliquer la méthode à la reconnaissance des formes. La conclusion quant à la validité de notre système doit donc porter sur sa possible application à la classification d'objets. Nous étudions ci-après le cas de la classe visages, le temps ne nous ayant pas permis de nous pencher sérieusement sur la possibilité de détecter d'autres classes d'objets.

Il est clair que notre système nous approche plus d'une possible détection qu'avec la méthode initiale. En effet, au moins deux zones importantes comme un oeil et la bouche ont été détectées sur tous les tests réalisés. Cependant, ces deux zones ne suffiront pas à la mise en place d'un descripteur pour la classe "visage". Le descripteur devra donc être créé pour des images où seul un visage est présent et où le contraste entre les cheveux, ou le cou, et le fond est faible. Comme on peut le voir sur la figure 6.3, toutes les zones caractéristiques d'un visage sont détectées avec cette condition. Ainsi un détecteur de la classe visage pourrait d'ores et déjà être mis en place mais son application à



FIG. 6.2 – a) Résultats obtenus avec le système décrit au chapitre 4; b) Résultats avec le système d'attention visuelle multi-échelle

la détection de visages sur des images plus compliquées apparaît encore très difficile. Notre système nécessite encore d'être amélioré pour parvenir aux performances escomptées.

En conclusion, notre système présente pour un grand nombre d'images de bien meilleurs résultats que la méthode initiale. Il n'est cependant pas encore directement utilisable pour la reconnaissance des formes, même si les résultats laissent penser qu'il le sera après encore quelques améliorations.

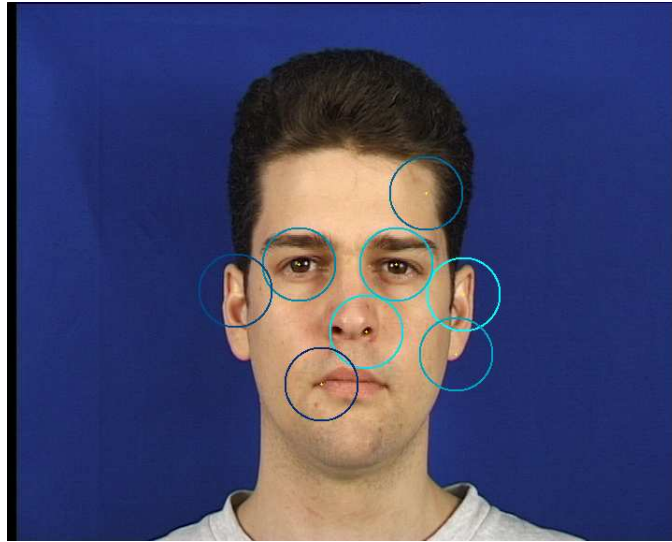


FIG. 6.3 – Résultat obtenu sur un visage dont les cheveux présentent un contraste assez faible avec le fond.

Chapitre 7

Conclusion et Perspectives

En partant des observations et améliorations du système d'attention visuelle multi-échelle décrit dans [13], nous avons mis au point un nouveau système. Contrairement à la méthode initiale, le nouveau système est multi-échelle et les pyramides gaussiennes utilisées ne sont plus dyadiques. Cela permet de détecter les points saillants à partir de cartes de saillance à différentes résolutions, certaines contenant plus de détails que l'unique carte du modèle initial. De plus, l'espace de couleur varie maintenant suivant le type d'image en entrée, et plus d'importance est donnée à la caractéristique d'orientation.

Les résultats obtenus sur une vingtaine d'images variées sont très encourageants, leurs qualités étant en général nettement meilleurs qu'avec la méthode initiale. En particulier, notre système s'adapte à différents types d'images, ne se limitant plus aux cas où l'objet saillant est assez grand et ressort vraiment du reste de l'image de part l'une de ses caractéristiques (intensité, teinte ou orientation). Il est tout de même dommage qu'aucune comparaison avec des données obtenues sur des humains n'est pu être entreprise. L'application visée étant la reconnaissance des formes, on peut tout de même d'ores et déjà valider en partie la méthode pour la détection de visages : une grande partie de l'ensemble des points saillants permettant de définir un visage est en général trouvée, et un descripteur de cette classe visage pourrait dès maintenant être construit si les images d'apprentissage ne présentent pas un trop grand contraste entre les cheveux et le fond.

Cependant, ce système d'attention visuelle présente encore quelques défauts. Notamment, ce n'est pas l'ensemble complet des points saillants d'un objet qui est détecté mais seulement une grande partie, à laquelle s'ajoute des zones indésirables pour la phase de reconnaissance. Le système a donc encore besoin d'être modifié. Les perspectives d'amélioration sont nombreuses, et nous en décrivons brièvement certaines dans la suite. A la fin du chapitre est donnée une rapide introduction concernant l'application de la méthode à la détection de classe d'objets.

7.1 Ajout des symétries

Tout d'abord malgré toutes les analyses et modifications apportées sur les trois caractéristiques utilisées jusqu'à présent, tous les points saillants intéressants d'une image ne sont pas trouvés. On peut donc penser que ces trois caractéristiques ne sont pas suffisantes. Locher et Nodine [4] ont montré que les symétries attirent l'œil. Reisfeld et al. [7] ont mis en place un algorithme de détection des points d'attention de symétrie locale pour les images de niveau de gris. G. Heidemann [9] a ensuite étendu leur méthode à des images couleur. Les résultats présentés dans [9] sont très prometteurs. Ainsi, nous poursuivons notre étude en rajoutant comme quatrième caractéristique les symétries dans le calcul de la carte de saillance. Une telle caractéristique pourrait s'avérer très intéressante pour la détection de classes d'objets comme les visages. En effet dans certains tests, seul un œil ou une oreille a été obtenu. De nombreux objets sont symétriques et les applications seraient donc nombreuses.

7.2 Zone de recherche du point saillant suivant

Un des défauts de la méthode utilisée est qu'elle cherche les points saillants dans toute l'image. Or, des études [6] sur des singes attestent que chaque point de saillance ne se trouve pas dans une région quelconque de l'image. Pour les singes, le premier point de saillance est dans une petite zone, dont la taille dépend du nombre d'objets présents, au centre de l'image. Chaque point saillant suivant est à une distance assez faible du point précédent. Il serait peut-être intéressant de tenir compte de ces observations pour notre système. Cela permettrait de définir un ordre de parcours de l'image plus plausible biologiquement, qui débiterait proche du centre de l'image.

7.3 Ajout du processus top-down

Comme il a été précisé dans le chapitre Etat de l'art, deux grands types de processus, le bottom-up et le top-down, sont à l'oeuvre dans le mécanisme d'attention visuelle humaine. Le système décrit dans ce rapport ne tient compte que du processus bottom-up. De ce fait, les premiers points saillants obtenus ne correspondent pas toujours à ceux recherchés pour pouvoir reconnaître un objet précis dans la scène. Par exemple, dans le cas d'images contenant des visages, ce ne sont pas forcément les yeux, nez, bouche ou oreilles qui sont détectés en premiers. Ceci empêche de mettre en place des détecteurs efficaces pour une éventuelle phase d'apprentissage. Il apparaît donc que cette composante bottom-up ne suffit pas l'application désirée. Rajouter le processus top-down revient à indiquer au système la forme des caractéristiques à détecter pour un type d'objet. Les zones de l'image correspondant à ses caractéristiques ont alors une intensité plus élevée dans la carte de saillance qu'avec le processus bottom-up seul. L'ajout du processus top-down devrait permettre une grande avancé vers la phase d'apprentissage.

7.4 Utilisation du contexte

Une autre amélioration du système possible serait d'ajouter le contexte afin de définir le type d'image en entrée. Ainsi, le choix de l'espace de couleur pourrait être fait directement par le système. Pour améliorer le processus top-down, il serait de plus possible d'apprendre au système quelles caractéristiques sont à privilégier connaissant les objets importants généralement présents dans le contexte étudié.

7.5 Application à la détection d'objets

Le but final du projet débuté au cours de ce stage est d'utiliser les séquences de points saillants obtenues pour une tâche de reconnaissance des formes. Il s'agit de mettre en place une méthode d'apprentissage machine permettant d'identifier les sous séquences correspondant au parcours de saillance de la zone de l'image contenant l'objet souhaité. Sur un ensemble d'images d'apprentissage pour une classe d'objet donnée, on pourrait par exemple évaluer la probabilité de la classe "la zones saillante appartient l'objet" par rapport à la classe "la zone saillante n'appartient pas à l'objet". Chaque zone saillante serait alors un "sous-"descripteur, et pour la détection, on pourrait combiner les probabilités de tous ces "sous-"descripteurs d'un objet. Par exemple pour un visage, chaque élément saillant (oeil, nez, bouche...) représenterait un descripteur et la probabilité que l'ensemble de ces éléments appartienne à la classe "objet" définirait un descripteur global de l'objet à détecter. Cette perspective ne sera mise en oeuvre que quand les résultats obtenus par le système laisseront penser qu'elle marchera pour différentes classes d'images.

Bibliographie

- [1] A. L. Rothenstein et J. K. Tsotsos A. Zaharescu. Towards a biologically plausible active visual search model. 2004.
- [2] P. J. Burt and E. H. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, April 1983.
- [3] J.G Daugman. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, 20, 1980.
- [4] P.J Locher et C.F Nodine. Symmetry catches the eye. *Eye Movements : From Physiology to Cognition*, 1987.
- [5] L. Itti et C.Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3), 2001.
- [6] B. C. Motter et J. W. Holsapple. The guidance of eye movements during active visual search. *Vision Research*, 38, 1998.
- [7] D. Reisfeld H. Wolfson et Y. Yeshurun. Context-free attentional operators : the generalized symmetry transform. *International Journal of Computer Vision*, 14, 1995.
- [8] D. Gabor. Theory of communication. *J. of the Institute of Electrical Engineers*, 1946.
- [9] G. Heidemann. Focus of attention from local color symmetries. *IEEE Transactions On Pattern Analysis ans Machine Intelligence*, 26, July 2004.
- [10] E. Hering. Zur lehre vom lichtsinn : Sechs mittheilungen an die kaiserl. *Akademie der Wissenschaften in Wien*, 1878.
- [11] L. Itti. Phd : Models of bottom-up and top-down visual attention. *California Institute of Technology*, January 2000.
- [12] C. Koch and S. Ullman. Shifts in selective visual attention : Towards the underlying neuronal circuitry. *Human Neurobiology*, 4, 1985.
- [13] C. Koch L. Itti and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, November 1998.
- [14] S. Marcelja. Mathematical description of the responses of simple cortical cells. *J. Opt. Soc. Am.*, 70/11, 1980.
- [15] H. Hüdli et R. Müri N. Ouerhani, R. von Wartburg. Empirical validation of the saliency-based model of visual attention. *Electronic Letters on Computer Vision and Image Analysis*, 3(1), 2004.
- [16] Padgham and Saunders. The perception of light and colour. *Academic Press, London*, 1975.
- [17] A. Yarbus. Eye movements and vision. *Plenum Press, New York*, 1967.
- [18] Young. On the theory of light and colors. *Philosophical Transactions of the Royal Society*, 1802.

Annexe A

Résultats obtenus avec différents espaces de couleur

Dans cette annexe venant en complément de la sous-section 5.2.1, les résultats obtenus sur deux images avec différents espaces de couleurs sont présentés. Cette section a expliqué pourquoi l'espace RGB est probablement le meilleur pour détecter des yeux ou le nez d'un visage. En effet, les couleurs complémentaires produisant le meilleur contraste. Or, il est clair que sur l'image ci-dessous contenant un visage, les yeux noirs sur la peau claire ressortent grâce à cette variation du contraste d'intensité. Il en est de même pour le nez dont la teinte est identique au reste du visage mais dont l'orientation et le contraste d'intensité varient. Avec le système $RGB_{normalise}$, les yeux sont trouvés et le coin de la bouche, mais pas le nez. Pour une phase de reconnaissance, les points saillants obtenus avec l'espace RGB sont plus utilisables que ceux avec $RGB_{normalise}$. Avec l'espace La^*b^* seul un oeil est détecté.

Pour l'image du chien, dans le premier cas la langue n'est pas détectée car la teinte des poils bruns et de la langue rose est en fait très proche. Or le système $RGB_{normalise}$ met en valeur les objets dont la teinte se distingue du fond, et pour lesquels la variation du contraste d'intensité n'est pas trop grande. Le système RGB favorisant le contraste, deux points sont trouvés sur le contour entre les oreilles et le fond et trois sur des contours entre les poils blancs et les poils bruns, points pas vraiment nécessaires pour une phase de reconnaissance. Ainsi pour cette image séparer la chrominance de la luminance donne des points plus intéressants. C'est en fait l'espace La^*b^* qui permettra le mieux d'utiliser les points saillants obtenus pour la détection d'objets. .

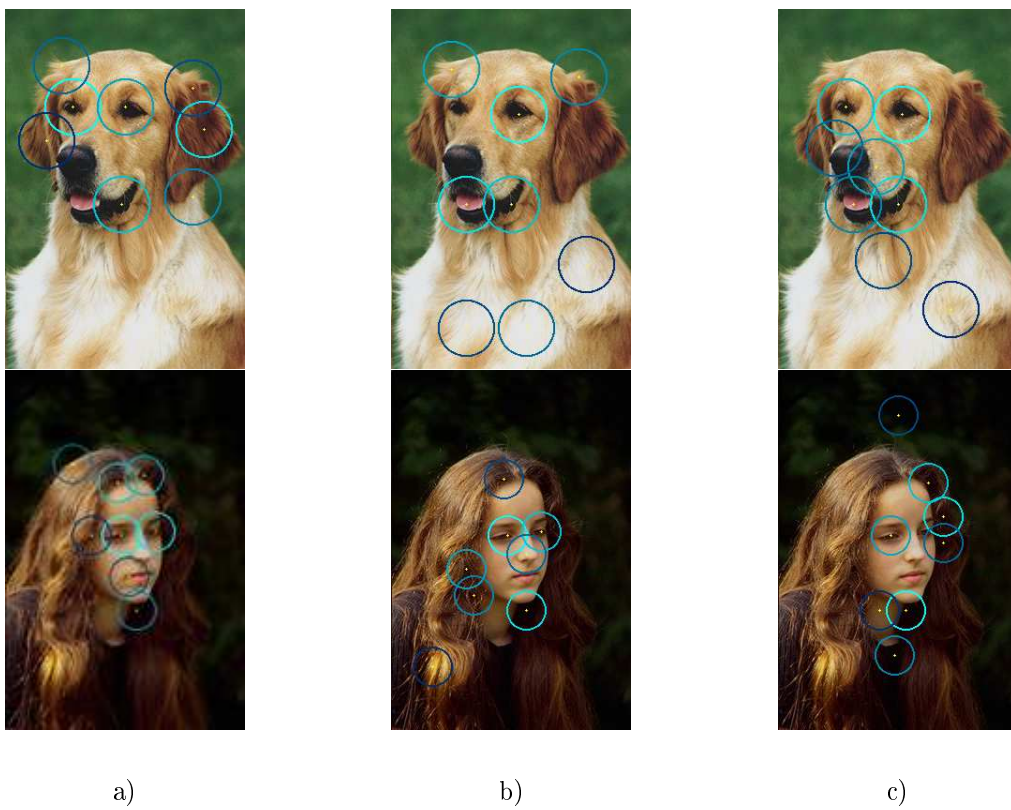


FIG. A.1 – Résultats obtenus en partant de la méthode [13], avec différents systèmes de couleur. a)méthode décrite dans l'article (système de couleur : RGB normalisé) ; b)RGB c)Lab.

Annexe B

Résultats obtenus avec différentes orientations

Cette annexe présente les résultats obtenus sur deux images en utilisant un nombre différent d'orientations. Elle montre que la différence obtenue pour les points saillants à différentes orientations est très minime. Les trois images présentées restent inchangées et donnent pas toutes les girafes, se que l'on aurait souhaité. Pour le visage, les résultats sont assez bons, mais pas suffisant pour une éventuelle phase de reconnaissance, le nez, la bouche et les yeux, n'étant pas tous trouvés. Avec 8 et 12 orientations, le menton n'est plus saillant. Il est remplacé par un point supplémentaire dans les cheveux, ayant probablement une orientation non mise en valeur directement par les filtres de Gabor de 4 orientations. En effet, le peu de changements pour les deux tests avec différentes orientations s'explique par le fait que la bande passante du filtre de Gabor utilisé est assez élevée. Cependant, elle n'est peut-être pas tout à fait assez grande pour vraiment prendre en compte toutes les orientations.

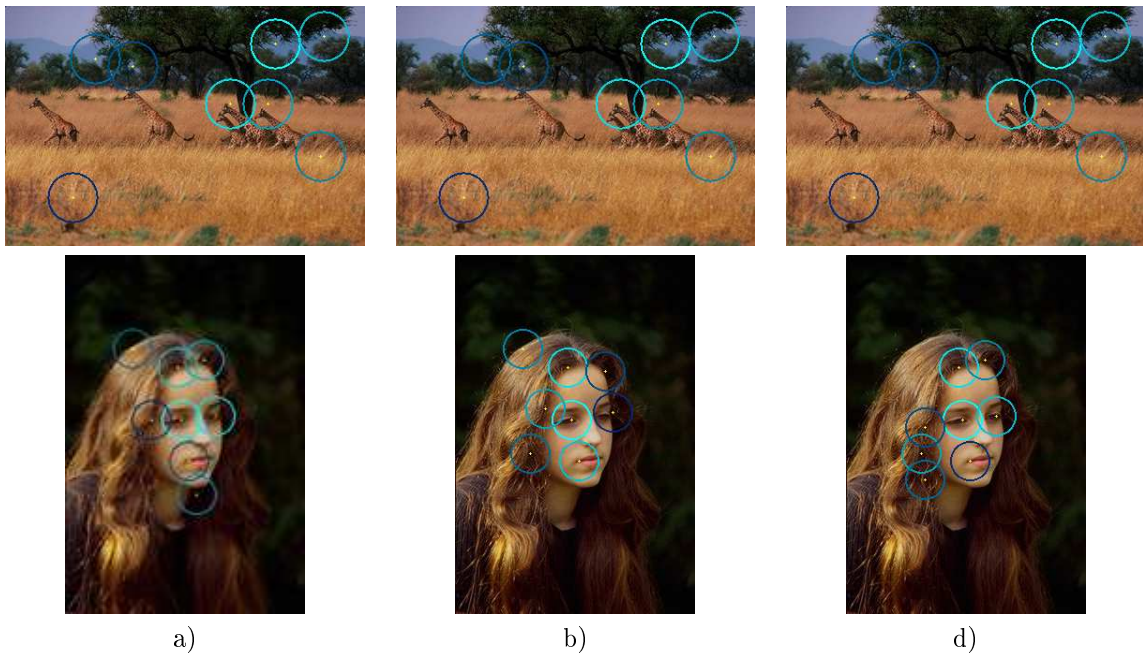


FIG. B.1 – Résultats obtenus en partant de la méthode [13] , avec plusieurs différentes orientations. a) 4 orientations, comme dans l'article; b) 8 orientations; c) 12 orientations.

Annexe C

Influence des cartes d'évidence

Dans ce chapitre sont présentés quelques résultats obtenus concernant l'influence des cartes d'évidence pour le calcul de la carte de saillance (voir sous section 5.3.1). Pour chacune des trois caractéristiques, cette annexe montre les résultats obtenus sur deux images en ne prenant plus en compte cette caractéristique et en lui donnant plus d'importance.

C.1 Résultats de l'influence du contraste d'intensité

Les images ci-dessous montrent deux résultats significatifs obtenus sur l'influence du contraste. Pour le visage, il est clair, que sans la présence de la caractéristique du contraste d'intensité, les résultats sont mauvais. En effet, un grand nombre de points saillants se retrouvent en haut de l'image où aucun objet intéressant n'est présent. Les zones se caractérisant par un fort contraste comme le contour du visage ou les yeux ne sont plus détectés. Les coordonnées des points saillants ne sont quasi pas changés par rapport à la méthode initiale en rajoutant deux fois plus d'importance à cette caractéristique. En fait, seul un point saillant a vraiment bougé. Il s'agit du point initialement présent sur le contour séparant les cheveux du fond. Cela est facilement compréhensible car le contraste à cet endroit est faible par rapport au contraste entre le visage et les cheveux. Or en amplifiant l'influence du contraste, il est normal que soient trouvées des zones plus contrastées.

Une grande partie de la deuxième image n'est pas contrastée : les girafes ont une chrominance et une luminance proches du fond. Ainsi, en enlevant la caractéristique de contraste, plus d'importance est donnée aux orientations et une girafe de plus est détectée. En lui donnant plus d'importance, les coordonnées des points restent totalement inchangés. Seul l'ordre dans lequel sont trouvés ces points est légèrement modifié, mais pas assez de manière significative pour conclure que plus d'importance améliore les résultats.

Ainsi, ces résultats confirment qu'enlever la caractéristique de contraste du système n'améliore les résultats que pour des images peu contrastées en couleur et intensité. Pour des images contrastées, les résultats sont beaucoup dégradés. Lui donner plus d'importance ne modifie pas les résultats de manière significative.



FIG. C.1 – a) Résultats obtenus en ne considérant que les caractéristiques de couleur et d'orientations ; b) Résultats obtenus avec la méthode Itti ; c) Résultats en donnant deux fois plus d'importance au contraste d'intensité.

C.2 Résultats de l'influence de la couleur

Voyons maintenant l'influence de la caractéristique de couleur sur quelques résultats. Les mêmes images ont été étudiées que pour les autres influences. De nouveau, pour l'image de droite, enlever la caractéristique de couleur améliore les résultats puisque plus de girafes (objets qui nous intéressent ici) sont toutes trouvées. Le résultat sans la couleur est aussi légèrement meilleur que celui sans le contraste d'intensité. En fait, les girafes ont quasiment la même teinte que le fond. Le contraste d'intensité est faible mais légèrement plus élevé que ce contraste de couleur, et ce sont encore les orientations qui priment.

Il est de nouveau difficile de conclure quant à l'influence sur cette image d'une valorisation de la carte d'évidence de couleur. Les points saillants ont légèrement bougé, mais il est impossible de dire si les nouvelles zones trouvées sont meilleures ou non. Dans ce cas, il aurait fallu pouvoir comparer le résultat avec les points saillants qu'aurait donnés un humain.

Le visage est aussi une image comportant peu de variations de teinte. Ainsi enlever cette caractéristique permet de trouver les yeux et la bouche, zones présentant un contraste important avec la peau, et un changement d'orientation brutal. Pour les mêmes raisons, favoriser la caractéristique de couleur dégrade le résultat sur le visage. Ce n'est en effet quasiment que le contour le séparant des cheveux qui est maintenant détecté.



FIG. C.2 – a) Résultats obtenus en ne considérant que les caractéristiques de contraste et d'orientations; b) Résultats obtenus avec la méthode Itti; c) Résultats obtenus en donnant deux fois plus d'importance aux couleurs qu'au contraste et à l'orientation, c) Résultats obtenus en donnant deux fois plus d'importance aux couleurs

C.3 Résultats de l'influence de l'orientation

Dans ce dernier annexe, nous illustrons l'influence de l'orientation. Comme nous l'avons vu dans les deux précédentes annexes, l'orientation est importante pour les deux images présentées, car elles sont peu contrastées en couleur ou en intensité. Pour le visage, la bouche n'est plus détectée sans cette caractéristique. Dans la deuxième image, seule une girafe est maintenant trouvée, les autres points donnant des zones sans grand intérêt. Ainsi, pour ce type d'images, la caractéristique d'orientation est indispensable. Par contre, lui donner plus d'importance permet de trouver le nez du visage, et toutes les girafes de la deuxième image. Cela illustre bien des conclusions données sur l'orientation dans la section 5.3.1.



FIG. C.3 – a) Résultats obtenus en ne considérant que les caractéristiques de contraste et de couleurs ; b) Résultats obtenus avec la méthode Itti ; c) Résultats obtenus en donnant deux fois plus d'importance à l'orientation qu'au contraste et aux couleurs.