Mécanismes d'Attention Visuelle sur Rétine Programmable

Taha Ridene[1] et Antoine Manzanera[2]

[1] École Nationale Supérieure des Mines de Paris,
Centre de Robotique (CAOR),
60 Bd Saint-Michel, 75272 Paris cedex 06, France
Tél : int+ 33 1 40 51 93 51, Fax : int+ 33 1 43 26 10 51
taha.ridene@ensmp.fr

[2] École Nationale Supérieure de Technologies Avancées,
Laboratoire d'électronique et informatique,
32 Bd Victor, 75739 Paris Cedex 15, France
Tél : int+ 33 1 45 52 44 42, Fax : int+ 33 1 45 52 83 27
antoine.manzanera@ensta.fr

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 l'environnement dans lequel il opère. Le rôle principal de ce mécanisme est d'accélérer le processus de vision, en réduisant sensiblement la quantité d'informations visuelles qui sera traitée par les tâches de plus haut niveau. Cependant, les méthodes mises en œuvre dans la littérature sont généralement très coûteuses en opérations de bas/moyen niveau. Les rétines numériques programmables, grâce à leur parallélisme massif, sont très bien adaptées pour de telles tâches. L'implantation d'une approche attentionnelle bio inspirée sur une rétine programmable, implique néanmoins la prise en compte de contextes architectural et algorithmique radicalement différents. Dans cet article, nous présentons les étapes d'implantation d'un modèle informatique d'*Attention Visuelle* basé sur les cartes de saillance, sur la rétine programmable *PVLSAR34*. La démarche algorithmique est donc massivement parallèle, et les traitements sont fortement fondés sur des notions de *multi-échelle*.

Mots clés : Attention Visuelle, rétine programmable, cartes de saillance, traitement multi-échelle.

1 Introduction

Bien que la quantité d'information présente dans notre environnement soit énorme, le système visuel humain a la capacité de gérer et d'interpréter très efficacement les images perçues, grâce à des mécanismes permettant de représenter l'environnement sous une forme compacte, en supprimant tout information redondante et en se focalisant sur les éléments les plus saillants[1]. Ces *Mécanismes d'Attention Visuelle* guident les mouvements d'œil pour placer la fovéa sur les parties les plus saillantes de la scène explorée. Deux stratégies d'exploration ont été révélées : Une stratégie descendante (*top-down*) qui repose sur une connaissance a priori, et une stratégie ascendante (*bottom-up*) faisant référence à l'*Attention Visuelle* involontaire.

[1] Un élément visuellement saillant, c'est un élément qui ressort prioritairement lors de la perception visuelle d'une scène, au point de prendre une importance cognitive particulière.

Les différentes études réalisées sur l'*Attention Visuelle* dans des domaines divers comme la psychologie, la psychophysique ou la neurophysiologie, ont été les principales source d'inspiration des chercheurs en vision artificielle. Pour la modélisation informatique d'*Attention Visuelle*, divers modèles se sont succédés [10] [8] [13], se basant essentiellement sur la théorie d'intégration des primitives [14]. La vocation de ces modèles informatiques est, en réduisant l'information visuelle à son "essence", de réduire le coût computationnel de la tâche de vision. Néanmoins, la multiplicité des primitives, et le caractère multi-échelle des traitements implique une charge de calcul sur le bas-niveau qui peut être considérable, en particulier dans la stratégie *bottom-up*.

Récemment, des travaux ont amené à la maturité des rétines artificielles, rapprochant les imageurs des rétines biologiques, en intégrant une puissance de calcul directement au niveau des capteurs. De nombreux modèles analogiques ou numériques ont été développés selon ce principe, on se référera à [1] pour une étude comparative détaillée. La rétine programmable *PVLSAR34* développée à *l'ENSTA* en 2004 (voir FIG. 1 **(a)**), est une machine massivement parallèle de 40 000 cellules interconnectées selon une grille 2d 200x200 en topologie *4-connexe*, chacune d'elle étant constituée d'un capteur photosensible et d'un processeur, doté d'une mémoire d'environ 50 bits. Chaque couple de processeurs adjacents partage une partie de leur mémoire, ce qui permet de communiquer des données entre pixels voisins. Au niveau du jeu d'instruction, chaque processeur peut effectuer les opérations booléennes élémentaires *("ET", "NON", "OU", etc.)*. Le mode de parallélisme est purement SIMD[2].

En intégrant le calcul de manière massivement parallèle directement au niveau du capteur, la rétine artificielle réduit très fortement le flux de données dans le système de vision, et donc le coût computationnnel associé aux tâches de bas-niveau. Il est donc très tentant d'associer le concept de rétine programmable et les *Mécanismes d'Attention Visuelle*. C'est ce que nous faisons dans cet article, en développant un modèle informatique cellulaire *bottom-up* d'*Attention Visuelle*, basé sur les cartes de saillance [8]. Nous présentons le modèle dans le section 2, puis nous discuterons les résultats dans la section 3.

2 Modèle cellulaire d'Attention Visuelle

2.1 Description du modèle global

Le modèle suit celui de l'attention *bottom-up* présentée dans [10], et comporte trois étapes principales. Tout d'abord l'extraction des cartes de caractéristiques, ensuite la génération des cartes d'évidence, et enfin l'intégration des cartes d'évidence dans une carte de saillance.
Les primitives que nous avons choisi d'extraire sont *l'intensité* et *l'orientation* pour une intégration d'une carte de saillance statique, et pour intégrer une carte de saillance dynamique, un module d'estimation de *mouvement* a été développé. L'absence de la *couleur* est justifiée par le fait que *PVLSAR34* est une rétine à niveaux de gris.
La figure 1 **(b)** illustre le modèle global de saillance considéré. Nous détaillerons dans ce qui suit chaque composante de ce modèle.

2.2 La représentation multi-échelle

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 primitives pré attentives.

[2] Single Instruction Multiple Data : Toute la grille est commandée par un séquenceur externe qui envoie une séquence d'instructions à la rétine selon une fréquence fixe, et à chaque pas de calcul, tous les 40 000 processeurs exécutent exactement la même instruction. Un programme rétinien est donc entièrement défini par la séquence d'instructions envoyée par le séquenceur.

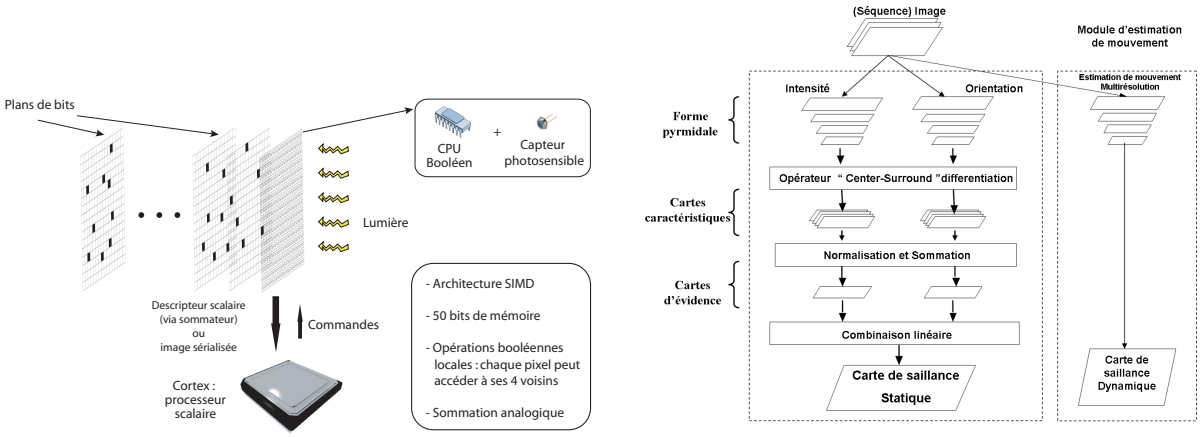

Fig. 1. (a) Un système à base de rétine programmable : la rétine peut-être vue comme une grille de couples microprocesseur et capteur photosensible. Les processeurs exécutant tous la même instruction, les opérations mémoires reviennent à manipuler des plans de bits (50 dans la dernière génération). Les opérations disponibles se réduisent à des décalages de plans de bits et à des opérations booléennes entre plans. Le cortex peut décider des instructions à envoyer et traite les données transformées par la rétine. **(b)** Modélisation informatique *bottom-up* d'*Attention Visuelle* basée sur les cartes de saillance.

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 *Center Surround*[3] comportant deux étapes :

1. Construction d'une représentation multi-échelle de la primitive ;
2. Calcul de la différence entre les niveaux fins et les niveaux grossiers de cette représentation.

L'étape 1 a comme résultat la forme sur laquelle se base tout les traitements du modèle cellulaire d'*Attention Visuelle*, les études déjà menées en analyse multi-échelle reposent sur deux étapes (Filtrage gaussien ; Sous-échantillonnage) donnant naissance à des formes pyramidales [4] [3] [6], les modèles attentionnels [8] [13] utilisent des pyramides dyadiques[4].

Dans notre cas la sous résolution n'a pas lieu d'être puisqu'on opère en mode SIMD. La structure multi-échelle que nous utilisons est celle du *cube gaussien* [11] [12]. L'implantation cellulaire du filtre gaussien a été réalisée sous la forme de moyennes locales itérées sur des voisinages de taille croissante. L'optimisation SIMD consiste à regrouper les sommes de manière dyadique et à utiliser des facteurs de normalisation égales à des puissances de 2 (voir FIG.2**(a)**).

Le nombre de niveaux de notre représentation multi résolution est fixé à 6, en prenant en compte l'image originale de niveau 0. Dans notre modèle, le centre est un pixel appartenant au niveau $c \in \{0, 1, 2\}$ du cube gaussien, et la région contournante le pixel correspondant au niveau $s = c + \delta$ avec $\delta \in \{2, 3\}$.

[3] Opérateur Center Surround « 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). Dans le cas d'une ligne horizontale entourée par des lignes verticales, la réponse en cet endroit sera plus grande que si elle était entourée par des lignes horizontales.

[4] Une pyramide est dite dyadique si le passage d'un niveau (i) à un niveau $(i+1)$ se fait avec le rapport 1/4.

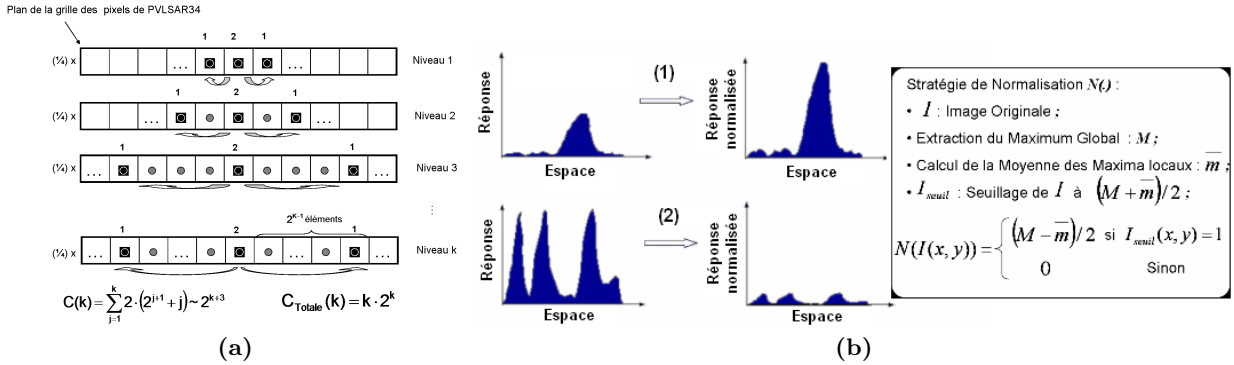

Fig. 2. (a) Simulation du filtre gaussien par itération de moyennes locales dans le sens horizontale (De même pour le sens verticale), opérant en mode SIMD, cette simulation est applicable pour les 40 000 éléments de la grille et dans les deux sens. Le gain en calcul dyadique peut être vu dans le cas ou nous élargissons le voisinage à additionner. Pour chaque niveau un nombre d'itérations est fixé. **(b)** Des objets saillants dans seulement quelques cartes peuvent être masqués par le bruit ou par d'autres objets moins saillants des autres cartes, le principe de normalisation est d'augmenter la réponse dans le cas (1) et de la diminuer dans le cas (2), l'implantation sur *PVLSAR34* est faite selon l'algorithme de *N(.)*.

2.3 Les primitives visuelles

L'opérateur « Center Surround » est appliqué aux trois primitives pré-attentives développées.

- L'intensité : Nous appliquons l'opérateur « Center Surround » en prenant comme entrée en première étape le cube gaussien.
- L'orientation : Les travaux précédents utilisent en général les filtres de Gabor[5] pour calculer cette primitive [8] [7] [13], nous avons choisi une approche basée sur des mesures de l'argument du gradient [2], qui est plus adaptée en calcul cellulaire.
- Le mouvement : Une mesure de flot optique est réalisée sur le cube gaussien d'entrée, et puis l'étape de différenciation est appliquée sur cette mesure.

2.4 Génération et combinaison des cartes d'évidence

La carte de saillance est calculée en combinant les différentes cartes caractéristiques. Une des difficultés pour cette combinaison est que ces cartes représentent des données a priori non comparables à des échelles différentes. Pour pallier à cette difficulté une étape de normalisation est appliquée (voir FIG.2 **(b)**). Deux stratégies principales ont été utilisées dans l'état de l'art [8] [9], l'application directe de ces deux stratégies pose des problèmes liés à la nécessité d'une grande dynamique de représentation (nombre flottant), notre méthode a suivi la même logique mais avec une implantation différente au niveau de l'extraction de $\mathbf{M}$ et $\overline{\mathbf{m}}$, l'algorithme est exposé dans la figure 2 **(b)**.

Nous intégrons ensuite pour chaque primitive les différentes cartes normalisées en une carte unique dite d'évidence, donnant la réponse, en terme de degrés de saillance des éléments de la scène observée, par rapport à la primitive considérée.

La carte de saillance finale est obtenue en combinant la carte de saillance dynamique résultat du traitement de la primitive mouvement et de la carte de saillance statique résultat de la combinaison linéaire des deux cartes d'évidence d'intensité et d'orientation.

[5] Un filtre de Gabor est une fonction sinus à laquelle on ajoute une enveloppe gaussienne.

3 Résultats et Discussion

La modélisation cellulaire de notre modèle attentionnel a donné naissance à des structures multi-échelles optimisées en calcul des *cubes gaussien* et de *DoG* bien adaptés à *PVLSAR34* (voir FIG.3), cet implantation pourrait être importée sur des architectures parallèles équivalentes.

L'evaluation du travail s'est faite pour l'instant en terme de complexité, les tests sur les primitives intensité et orientation de la carte de saillance statique sont exposés dans les tableaux 1 et 2. L'évaluation de la primitive mouvement est encore en cours.

Nous remarquons que nous approchons le temps-réel à un facteur de 3/4 pour la carte de saillance à base d'intensité. En fait l'implantation actuelle nous permet de traiter l'équivalent de 17 images/s, et en analysant le tableau 1 nous remarquons que la tâche la plus complexe est celle de la normalisation (78% du temps CPU). En augmentant le nombre de primitives cette complexité devient encore plus grande et la contrainte temps-réel n'est plus respectée (voir Tab. 2). Cela est dû au fait qu'à cette étape nous étions amenés à réaliser des opérations analogiques (ex : somme globale) en dehors de la rétine à travers le *cortex*, et avec l'implantation actuelle nous multiplions à chaque normalisation le nombre de ces opérations, ce qui pénalise l'ensemble du processus par un transfert coûteux réalisé entre rétine et cortex.

Les résultats que nous avons eus bien que partiellement satisfaisants, pourront être améliorés pour atteindre le temps-réel. Nous imaginons une amélioration par l'implantation d'une autre méthode de normalisation moins coûteuse en temps CPU.

(a) (b)

Fig. 3. (a) : Image Originale. (b) : Différence entre le Niveau 1 du cube gaussien avec $\sigma 1 =' 1.4017'$ et le niveau 3 avec $\sigma 2 =' 4.9940'$.

Tâche	Nbr cycles	Temps CPU (ms)
Cube Gaussien	49714	9.47
Différentiation	8622	2.775
Normalisation	117762	48.37
Intégration	9135	1.295
Coût total	185233	61.91

Tab. 1. Coût de la carte d'évidence d'intensité.

Tâche	Nbr cycles	Temps CPU (ms)
Cube Gaussien	49714	9.47
4 cubes d'orientation	27468	3.735
Différentiation	206928	10.97
Normalisation	471048	189.3
Intégration	42021	7.374
Coût total	797179	220.849

Tab. 2. Coût de la carte d'évidence d'orientation.

4 Conclusion

Les performances des applications, temps-réel en traitement d'images sont susceptibles d'être améliorées via des mécanismes d'accélération du processus de vision appelés *Mécanismes d'Attention Visuelles*. Dans ce cadre, et se basant sur une architecture parallèle, la rétine *PVLSAR34*, nous avons implanté un modèle cellulaire *bottom-up* d'*Attention Visuelle* basé sur les primitives : intensité, orientation et mouvement. L'implantation nous a permis de développer une approche basée sur les calculs cellulaires et les traitements multi-échelles et l'évaluation du modèle s'est faite par rapport au critère de la complexité en temps CPU et en cycles Rétines. L'approche attentionnelle étudiée pourrait être eprouvèe par des évaluations dans un cadre applicatif. Nous imaginons l'intégration de ce modèle attentionnel pour un algorithme de *tracking* déja réalisé sur *PVLSAR34*.

Des améliorations peuvent être appliquées sur deux niveaux. D'un côté nous pourrons imaginer, au niveau architectural l'intégration des calculs analogiques se faisant jusqu'à maintenant en dehors de *PVLSAR34* : pour l'instant seules des opérations locales ont pu être implémentées, mais une nouvelle génération de rétines [5] apportera des primitives régionales, par l'utilisation du concept d'asynchronisme, ce qui facilitera la tâche pour des traitements plus poussés. D'un autre côté, en se comparant au système visuel humain, le bon fonctionnement d'un processus attentionnel a souvent besoin de la connaissance a priori, une perspective de notre travail pourrait être le développement d'une approche attentionnelle hybride par l'ajout d'un module *top-down* au modèle *bottom-up* présenté dans ce papier.

Références

1. T.M. Bernard and F. Paillet. Output methods for an associative operation of programmable artificial retinas. In *IEEE/RJS Int. Conf. on Intelligent Robots and Systems*, pages 752–757, September 1997.
2. N. Burrus and T. Bernard. Adaptive vision leveraging digital retinas : Extracting meaningful segments. In *ACIVS*, 2006.
3. P.J. Burt. The pyramid as a structure for efficient computation. *Multiresolution Image Processing and Analysis*, pages 6–35, 1984.
4. P.J. Burt and E.H. Adelson. The laplacian pyramid as a compact image code. In *IEEE Transactions on Communications*, volume 31, pages 532–540, 1983.
5. V. Gies. *Increasing Interconnection Network Connectivity for Reducing Operator Complexity in Asynchronous Vision Systems.* PhD thesis, Universitité Paris-Sud XI, Décembre 2005.
6. R. Hummel. The scale-space formulation of pyramid data structures. *Parallel Computer Vision*, pages 107–123, 1987.
7. L. Itti and P. Baldi. A principled approach to detecting surprising events in vide. pages 1063–69, 2005.
8. L. Itti and C. Koch. A model of saliency-based visual attention for rapid scene analysis. *Trans. Pattern Anal. Mach. Intell*, 20(11) :1254–59, 1998.
9. L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10) :1489–1506, 2000.
10. C. Koch and S. Ullman. Shifts in selective visual attention : towards the underlying neural circuitry. *Human Neurobiology*, 4 :219–227, 1985.
11. P. Meer. Simulation of constant size multiresolution representations on image pyramids. *Pattern Recognition Letters*, 8 :229–236, 1988.
12. P. Meer. Stochastic image pyramids. *Computer Vision, Graphics, and Image Processing*, 45 :269–294, 1989.
13. N. Ouerhani. *Visual Attention : From Bio-Inspired Modeling to Real-Time Implementation.* PhD thesis, Université de Neuchatel Facultée des Sciences, 2003.
14. A.M. Treisman and K. Gelade. A feature-integration theory of attention. *Cognit Psychol*, 12(1) :97–136, 1980.