

AIX-MARSEILLE UNIVERSITÉ

MÉMOIRE DE MASTER

---

# Computational modelling of visual object localization in the magnocellular pathway

---

*Auteur:*

Pierre ALBIGÈS

*Superviseur:*

Emmanuel DAUCÉ

*Une thèse présentée à*

ECOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTÉ

*en vue de l'obtention du diplôme de*

MASTER DE NEUROSCIENCES, SPÉCIALITÉ INTÉGRATIVES ET COGNITIVES

*et réalisée au sein de*

Institut de Neurosciences des Systèmes

*Durant la période : 04/12/2017 - 02/03/2018*



AIX-MARSEILLE UNIVERSITÉ

## *Résumé*

Faculté des Sciences, département de Biologie

Master de Neurosciences

Master de Neurosciences, spécialité Intégratives et Cognitives

### **Computational modelling of visual object localization in the magnocellular pathway**

by Pierre ALBIGÈS

Malgré des décennies de développement et une place prépondérante au sein de la recherche en intelligence artificielle, la vision par ordinateur conserve toujours des défauts majeurs comparée à la vision naturelle en terme de performance et d'adaptabilité à l'environnement. L'un de ces défauts est à la puissance de calcul faramineuse nécessaire pour obtenir une vision artificielle haut performance.

Afin de proposer une solution à ce problème, nous avons construit un agent markovien de vision artificielle capable de détecter la position d'une cible dans son environnement visuel et de réaliser une action (saccade oculaire) en conséquence, afin d'optimiser spécifiquement le traitement des informations provenant de cette cible. L'originalité de notre approche tient en l'application d'un filtre rétinien directement sur l'image perçue par le modèle, avant que soit réalisé tout traitement d'information. Ce filtre, disponible sous deux modèles (Wavelets et LogPolar), permet de reproduire l'acuité rétinienne (maximale en son centre puis diminuant progressivement avec l'excentricité) et donc de réduire fortement la puissance de calcul nécessaire pour percevoir l'environnement visuel.



# Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction</b>                     | <b>1</b>  |
| 1.1      | Vision naturelle . . . . .              | 1         |
| 1.2      | Vision artificielle . . . . .           | 3         |
| <b>2</b> | <b>Matériel et méthodes</b>             | <b>5</b>  |
| 2.1      | Support physique et numérique . . . . . | 5         |
| 2.2      | Modèle POMDP . . . . .                  | 5         |
| 2.3      | Champs rétinien . . . . .               | 6         |
| 2.4      | Apprentissage supervisé . . . . .       | 7         |
| <b>3</b> | <b>Résultats</b>                        | <b>9</b>  |
| 3.1      | Apprentissage supervisé . . . . .       | 9         |
| 3.2      | Prédiction de la position . . . . .     | 9         |
| 3.3      | Prédiction de la catégorie . . . . .    | 10        |
| <b>4</b> | <b>Discussion et perspectives</b>       | <b>11</b> |
|          | <b>Bibliography</b>                     | <b>15</b> |
| <b>A</b> | <b>Figures</b>                          | <b>17</b> |
| <b>B</b> | <b>Code source</b>                      | <b>27</b> |



# 1 Introduction

## 1.1 Vision naturelle

Tous les êtres vivants utilisent la vision à un degré ou à un autre, et pour de nombreuses espèces -y compris la notre- elle est même la modalité perceptive principale. Elle est alors primordiale pour appréhender l'environnement et interagir avec celui-ci, que ce soit dans une optique de survie de l'individu ou dans la construction de relations sociales.<sup>11</sup>

Chez les vertébrés la vision débute à la surface de la rétine, où les cellules photovoltaïques (cônes et bâtonnets) réalisent la transduction des signaux lumineux qui les atteignent en signaux électriques, transmissibles aux réseaux nerveux en aval.<sup>11</sup>

Les cônes et les bâtonnets sont différenciables par un certain nombre de caractéristiques, notamment leur sensibilité aux longueurs d'ondes lumineuses et leur distribution au sein de la rétine. Ces différences permettent à la rétine de rester fonctionnelle et de continuer à nous fournir des informations pertinentes dans de nombreux contextes, y compris lorsque la luminance est très faible.<sup>11</sup>

Le champs visuel peut être divisé en deux parties. La vision centrale (environ 2° chez l'humain) est anatomiquement soutenue par la fovea, une région rétinienne comprenant uniquement des cônes. On y observe l'acuité visuelle la plus importante (la région présente les champs récepteurs les plus petits de la rétine) et une forte perception des couleurs.<sup>11</sup>

La vision périphérique quant à elle comprends majoritairement des bâtonnets, présente une faible acuité et une perception des couleurs très faible (voire nulle). Elle est par contre très sensible à des variations de luminance et de fréquence spatiale (donc aux mouvements).<sup>11</sup>

Plus précisément, l'acuité visuelle diminue avec l'excentricité par rapport à la fovéa. Autrement dit, les champs récepteurs visuels grandissent avec cette excentricité. Cette modulation de l'acuité permet de limiter le flux d'informations devant être traité par le système visuel en aval en ne conservant que

les informations jugées pertinentes par le système (passant d'un flux arrivant à la rétine estimé à  $10^8$  bits/s à une sortie estimée à  $10^2$  bits/s, soit une réduction de plus de 99%). C'est ce qui a été théorisé par Barlow en 1961 sous le principe du codage efficace.<sup>8,12,11,7</sup>

Lors de l'exploration de son environnement visuel, un agent va pouvoir détecter des stimuli dans sa vision périphérique mais ne va pas y présenter assez d'acuité pour en réaliser une description précise. En conséquence, l'agent va réaliser des saccades oculaires (mouvements brefs (20-60ms) des globes oculaires) afin de placer l'image de la cible (ou tout du moins sa position prédite dans l'espace) au niveau de la fovéa, permettant ainsi de traiter les informations en provenant avec la plus grande précision possible.<sup>8,11</sup>

L'activité rétinienne est ainsi transmise le long des voies nerveuses visuelles jusqu'au cortex visuel, où sera réalisé la majorité du traitement des informations -notamment haut niveau- qu'elle code.<sup>11</sup> Entre la rétine et le cortex existe un certain nombre d'étapes, mais tout au long de ces voies la distribution rétinienne de l'information (la rétinotopie, figure A.1) est conservée.<sup>11</sup>

Dans leurs travaux de 1962 et 1977, Hubel et Wiesel émettent l'hypothèse des courants visuels, définissant trois voies nerveuses naissant dans le **corps genouillé latéral** (LGN) et projetant sur le **cortex visuel primaire** (V1) : les voies **magnocellulaire** (M), **parvocellulaire** (P) et **koniocellulaire** (K). Chacune d'entre elles transporte des influx nerveux codant pour des caractéristiques différentes des stimuli visuels.<sup>2,11</sup>

L'activité des cellules M, par exemple, ne distingue pas les couleurs mais est sensible à des différences fines de luminance, de contraste et de fréquence temporelle. Ces caractéristiques semblent notamment lier la voie M au traitement de la luminance et des mouvements.<sup>2,11</sup>

Cette multiplicité de voies visuelles est conservée au delà de V1, où l'on décrit deux courants sortants : les voies **ventrale** (transportant et traitant majoritairement pour des informations provenant de la voie P) et **dorsale** (transportant et traitant majoritairement pour des informations provenant de la voie M).<sup>11,4,6</sup>

La voie ventrale communique ainsi principalement avec les aires cérébrales du lobe temporal, l'activité de son réseau étant primordiale pour la reconnaissance et l'identification des objets visuels. La voie dorsale quant à elle communique principalement avec les aires du lobe pariétal, l'activité de ce réseau étant primordiale pour le traitement des relations spatiales entre les objets visuels ainsi que pour le



guidage attentionnel et visuomoteur vers eux.<sup>11,4,6</sup>

Parmi ce réseau dorsal, on trouve l'**aire intrapariétale latérale** (LIP), qui reçoit en partie des informations directement depuis V1 et V2 (contournant donc le traitement d'aires en amont, dont l'aire MT) codant pour des stimuli dans le champ visuel périphérique. Des travaux ont d'ailleurs relié l'activité des neurones du LIP à la représentation spatiale des objets visuels et à la planification des saccades oculaires.<sup>11</sup>

La planification et l'exécution des saccades oculaires implique le fonctionnement d'un réseau nerveux complexe et dont on ne connaît pas encore entièrement l'état, comprenant des régions cérébrales allant des aires visuelles associatives (comprenant le LIP) jusqu'au tronc cérébral, en passant par le thalamus (figure A.2).<sup>12</sup>

## 1.2 Vision artificielle

La vision représentant notre modalité perceptive principale et les aires dédiées à traiter ses informations occupant une part significative de notre cortex cérébral (jusqu'à 50% chez certains primates), l'étudier permet de mieux comprendre le fonctionnement général de notre système nerveux.<sup>12</sup>

De nombreux domaines d'étude s'intéressent donc au fonctionnement du système visuel, de ses aspects les plus moléculaires jusqu'aux fonctions les plus intégrées. Parmi eux, les neuromathématiques se basent sur les données expérimentales (anatomique, physiologique et comportementale) pour émettre des modèles mathématiques sur le fonctionnement d'une partie ou de l'ensemble de la modalité visuelle. Idéalement, ces théories doivent pouvoir expliquer son activité dans l'ensemble des contextes observables, mais aussi pouvoir prédire son comportement dans de nouveaux contextes.<sup>12</sup>

Mais ces modèles ne sont pas une finalité en soi dans la compréhension du système. Ils peuvent permettre d'identifier dans les théories de son fonctionnement des défauts ou des zones obscures à notre compréhension et donc diriger les études expérimentales vers ces points.<sup>12</sup>

L'identification de ces points et la démonstration de ces théories peut passer par le domaine des neurosciences computationnelles, qui va tenter d'appliquer ces modèles mathématiques dans des modélisations du système nerveux. Au delà de la possible validation ou invalidation des modèles, ces intégrations permettent de résoudre des problèmes d'ingénierie (puissance de calcul disponible, vitesse de traitement, adaptabilité à l'environnement, ...) en s'inspirant des systèmes biologiques, très optimisés,

et donc de créer des systèmes artificiels neuromimétiques plus performants et utilisables dans des systèmes embarqués ou des interfaces cerveau-machine.

Dans cette étude, nous avons tenté de construire un modèle simple de localisation de cible visuelle dans un champs rétinien (imposant une vision centrale où l'acuité est maximale et une vision périphérique où l'acuité diminue avec l'excentricité). Le fonctionnement du modèle fait écho à une vision simplifiée du fonctionnement du système visuel (figure A.3).<sup>12</sup>

Le modèle doit être capable de détecter dans sa vision périphérique une cible visuelle aux caractéristiques simples (représentée par un stimulus provenant le base de données MNIST), de prédire précisément sa position et de réaliser une saccade oculaire afin de la placer dans sa fovea, ce qui lui permet alors de l'identifier avec une certitude élevée.

Pour cela, plusieurs méthodes (et sous-méthodes) sont utilisables. Les **modèles de saillance** tentent de décrire une image en fonction des régions (ou pixels) qui présentent la plus grande probabilité de fournir des informations pertinentes. Généralement, après compétition entre chacune de ces régions, la gagnante est considérée comme la plus saillante et donc celle qui devrait attirer les saccades. Une fois cette région visitée, le système l'inhibe dans sa représentation et visite la région gagnante suivante, et ainsi de suite. Ces modèles permettent donc de créer des histogrammes de probabilité de fixations oculaires et semblent correctement modéliser l'activité de certaines régions cérébrales (pulvinar, colliculus supérieur, sillon intraparietal), mais présentent certaines limites, dont l'absence de ciblage d'objets partiellement ou entièrement cachés.<sup>1,7</sup>

Les **modèles de contrôle**, tentent quant à eux de prédire quelles règles le système devra suivre pour répondre au mieux à une tâche. A chaque saccade, de nouvelles informations sur l'environnement sont collectées et permettent de changer l'opinion de l'agent sur le monde.<sup>1</sup>

Notre modèle s'inspirera librement de chacune de ces deux catégories pour son fonctionnement sans être l'adaptation exclusive de l'une d'entre elles.

## 2 Matériel et méthodes

### 2.1 Support physique et numérique

L'ensemble des simulations ont eut lieu sur un ordinateur portable hébergeant une machine virtuelle (caractéristiques rassemblées dans le tableau A.1).

Les modélisations ont été réalisées à l'aide du langage de programmation **Python** (version 3.6.4) renforcé de la librairie **TensorFlow** (version 1.4) et de l'interface graphique **Jupyter**.

La base de données **MNIST** a été utilisée pour l'apprentissage et l'évaluation du modèle. Elle contient 70.000 images de chiffres manuscrits (60.000 pour l'entraînement, 10.000 pour l'évaluation) centrés et dont la taille a été normalisée. Chaque image est accompagnée d'un label décrivant quel chiffre elle contient.

### 2.2 Modèle POMDP

Le problème de recherche d'information dans un contexte d'exploration de l'environnement visuel peut être formulé comme un **processus de décision Markovien partiellement observable** (POMDP).<sup>1</sup>

Dans un POMDP (figure A.4), l'agent perçoit partiellement l'**état de l'environnement**  $S$  à un temps  $t$  (dans ce travail, l'environnement visuel) et peut réaliser des **actions**  $A$  (ici des saccades oculaires) qui peuvent avoir des conséquences sur l'environnement et sa perception  $O$  (l'environnement visuel perçut au travers du champs rétinien). L'agent va ainsi construire un **état de croyance**  $B$  (ici les prédictions de position ou de catégorie du stimulus) en fonction des observations et des actions réalisées jusqu'ici.<sup>1</sup>

Un tel système doit satisfaire la **propriété de Markov**, qui décrit que la distribution de probabilité des futurs états ne dépends que de l'état précédent et pas de toute la séquence d'états en amont. Cette propriété retire donc de l'agent la notion de mémoire à long terme (ici correspondant à tout ce qui a pu se dérouler avant  $t_{-1}$ ) des états précédents de l'environnement, mais aussi des actions et des observations réalisées dans le passé.

Ainsi lors de l'évolution du système dans le temps, on considère que l'état suivant de l'environnement est uniquement influencé par son état actuel et l'action (éventuelle) réalisée par l'agent (équation 2.1).<sup>1</sup>

$$p(s_{t+1}|s_{1:t}, a_{1:t}, o_{1:t}) = p(s_{t+1}|s_t, a_t) \quad (2.1)$$

De même, les observations actuelles de l'agent ne dépendent que de l'état actuel de l'environnement et de l'action (éventuelle) qu'il réalise (équation 2.2).<sup>1</sup>

$$p(o_t|s_{1:t}, a_{1:t}) = p(o_t|s_t, a_t) \quad (2.2)$$

$$B_t^i = p(S_t = i|A_{1:t}, O_{1:t}) \quad (2.3)$$

## 2.3 Champs rétinien

Avant d'être utilisée par le modèle, l'image provenant de la base MNIST subit un certain nombre de transformations.

Chaque image présente originellement 28x28 pixels auxquels correspondent des niveaux de gris parmi 255 valeurs possibles (permettant un codage sur seulement 8bits par pixel). Cette image est placée au hasard (mais de façon normocentrée) sur une image au fond blanc de 128x128 pixels afin de réduire sa taille par rapport à l'environnement visuel qui la contient (figure A.5).

A cette image est ensuite appliqué un *filtre Wavelets* ou un *filtre LogPolar*, deux méthodes permettant d'obtenir un champs rétinien, c'est à dire un filtre visuel dont l'acuité est maximale en son centre (simulant la fovea) et diminuant avec l'excentricité (simulant la vision périphérique).

Dans les deux cas, la transformation mathématique imposée par le filtre est calculée à l'avance, permettant de l'appliquer à chaque nouvel environnement visuel et après chaque saccade oculaire (donc à chaque nouvelle observation), tout en économisant au maximum la puissance de calcul disponible.<sup>8</sup>

Le **filtre Wavelets** consiste en l'application sur l'image d'une grille de résolution, définissant des anneaux concentriques de taille croissante et contenant des superpixels dont la taille augmente avec l'excentricité de l'anneau. Chacun des superpixels possédant un niveau de gris correspondant à la moyenne des valeurs des pixels qu'il contient, la résolution de l'image ainsi traitée diminue à chaque nouvel anneau, et donc par palier.<sup>8</sup> Les effets de ce filtre sont visibles sur la figure A.6.

Le **filtre LogPolar** quant à lui est basé sur une approche neuromimétique, visant à reproduire la forme et l'organisation réelle des champs récepteurs présents dans le système visuel biologique. Pour suivre cette approche, il est constitué d'un ensemble de filtres LogGabors (figure A.7), qui sont des fonctions mathématiques (plus précisément, ils sont construits en multipliant une fonction gaussienne à une exponentielle complexe) présentant un certain nombre d'avantages pour la modélisation des champs récepteurs naturels.<sup>3</sup>

Nous les utilisons donc pour modéliser la distribution des champs récepteurs rétiniens, mais il est aussi possible de facilement les adapter (en modifiant seulement quelques paramètres) à la modélisation des aires visuelles primaires et associatives et peut donc être ré-utilisé dans le cadre d'un modèle visuel multi-couches neuromimétique, notamment pour la modélisation de la voie visuelle ventrale.<sup>4</sup>

Une représentation graphique du filtre utilisé dans ces travaux et son effet sur l'un de nos stimuli visuels sont visibles sur les figures A.8 et A.9.

## 2.4 Apprentissage supervisé

Afin d'obtenir un modèle à la fois performant et adaptable, nous l'avons soumis à un apprentissage supervisé sous la forme d'une **régression linéaire multivariée** optimisée par **descente de gradient**. Cette méthode de *machine learning* a été préférée notamment pour la simplicité de sa mise en place, mais pourra être remplacée par une méthode plus complète lorsque le modèle sera dans un stade de développement plus avancé.

Dans cette méthode, nous calculons une hypothèse  $h_\theta$  (équation 2.4) sur la répartition des stimuli en multipliant chacune des valeurs  $x$  de l'entrée (l'image après transformation par le filtre rétinien) à un poids  $\theta$  qui lui est spécifique, puis en ajoutant éventuellement un biais  $b$ .

$$h_\theta(x) = \theta^T x + b \quad (2.4)$$

Ces poids sont ensuite optimisés par descente de gradient (équation 2.5), où ils sont comparés aux labels  $y$  des entrées (correspondant aux informations que l'on veut apprendre, par exemple les coordonnées  $(x, y)$  de la cible) pour un nombre d'exemples et d'itérations fixées. Le paramètre d'apprentissage  $\alpha$  influence très fortement l'entraînement et sa valeur doit être adaptée pour éviter un sous- ou un sur-apprentissage (révélant respectivement une valeur trop faible ou trop importante).

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m (h_\theta(x^i) - y^i) x_j^i \quad (2.5)$$

En parallèle peut être calculé le coût  $J(\theta)$  (équation 2.6), dont l'évolution au cours de l'entraînement est un indicateur de l'efficacité de l'apprentissage. Sa valeur devant décroître au cours du temps, l'optimisation du modèle peut se faire en tentant de la réduire au maximum.

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_\theta(x^i) - y^i)^2 \quad (2.6)$$

Nous construisons ainsi un modèle de machine learning contenant deux couches et dont le comportement suit un modèle POMDP. Pour cela, nous réalisons l'entraînement avec en entrée deux vecteurs: l'entrée  $x$ , produit d'une transformation mathématiques de l'image après application de l'un des filtres rétinien, ainsi que le label  $y$  comprenant (suivant la couche que l'on entraîne) soit les coordonnées  $(x, y)$  de la cible, soit sa catégorie (le chiffre que l'image contient, sous la forme d'un vecteur one-hot). La première couche (*détecteur*), dont le développement est l'objectif principal de ce travail, est entraînée à prédire la position du stimulus dans l'image perçue au travers du champs rétinien, permettant de réaliser une saccade jusqu'aux coordonnées prédites et donc d'approcher la cible de la vision fovéale. Chaque saccade permet donc d'optimiser la détection du signal.<sup>5</sup> La seconde (*classifieur*) est entraînée à prédire la catégorie du stimulus dans l'image perçue au travers du champs rétinien et permet d'arrêter l'exploration de l'image lorsque sa prédiction présente une certitude assez élevée.

## 3 Résultats

Cette partie contient des résultats préliminaires, le modèle étant encore en cours de construction et d’optimisation lors de l’écriture de ce rapport.

### 3.1 Apprentissage supervisé

L’étape de benchmarking du paramètre d’apprentissage  $\alpha$  (équation 2.5) permet de rendre compte de son importance sur l’efficacité de l’apprentissage et du modèle.

Dans le cadre du filtre *Wavelets* et lorsque seulement le *détecteur* est entraîné, on peut ainsi observer qu’une valeur  $\alpha \geq 0.4$  entraîne un sur-apprentissage très important (le coût augmente fortement au cours de l’entraînement, figures A.10 et A.11), tandis que  $\alpha = 0.3$  semble représenter une valeur optimale pour l’apprentissage (figure A.12).

Par contre, lorsque *détecteur* et *classifieur* sont tous deux entraînés, deux jeux de poids indépendants doivent être optimisés par l’apprentissage. Chaque couche possède ainsi son propre paramètre  $\alpha$  (respectivement  $\alpha_{detect}$  et  $\alpha_{classif}$ ) et l’on peut donc calculer leurs coûts indépendamment (figure A.13).

### 3.2 Prédiction de la position

Après avoir été entraîné, le modèle semble capable de détecter la cible dans son environnement et de prédire précisément sa position (figure A.14 et A.15). Il semble aussi capable d’utiliser cette prédiction pour réaliser une saccade aux coordonnées prédites de la cible visuelle, ce qui modifie en conséquence sa perception de l’environnement.

Mais une seule saccade n’est pas toujours suffisante pour atteindre la cible (figure A.16) et le nombre de saccades nécessaires augmente avec la distance initiale de la cible du centre de fixation (figure A.17). Cette relation pourrait provenir de la diminution de l’acuité avec l’excentricité dans la champs visuel

(provoquée par le champs rétinien), entraînant une diminution de la précision des prédictions (figure A.18).

### 3.3 Prédiction de la catégorie



## 4 Discussion et perspectives

Encore une fois, nous préférons insister sur le fait qu'au moment de l'écriture de ce rapport, le modèle est toujours en cours de développement et donc que les idées émises dans ce chapitre ainsi que le précédent restent des hypothèses qui devront nécessairement être confirmées lors de travaux ultérieurs.

Notre modèle semble ainsi capable de suivre le fonctionnement d'un modèle POMDP en réalisant à tour de rôle une observation de son environnement et une action ayant pour objectif d'améliorer la perception de cet environnement.

L'agent va donc réaliser une série de saccades jusqu'à réussir à placer la cible visuelle au niveau de sa fovéa, où sa description (ici sa catégorisation) pourra être réalisée avec la plus grande acuité possible, et donc avec le plus grand taux de réussite. Ce comportement "atteindre-et-décrire" instauré par le fonctionnement POMDP semble cohérent avec ce que l'on peut observer dans les systèmes biologiques.<sup>11,9</sup>

En complément à ces observations sur le comportement générale du modèle, nous avons pu observer que plus une cible est éloignée de la fovéa de l'agent lorsqu'elle est présentée initialement, moins ses prédictions seront précises et en conséquences plus nombreuses seront les saccades destinées à atteindre la position de la cible. Encore une fois, ce comportement semble cohérent avec ce que l'on observe dans les systèmes biologiques. A noter toutefois que le profil de performances du modèle, notamment concernant l'évolution de la taille de ses erreurs de prédiction avec l'excentricité de la cible, correspond à de faibles performances biologiques. Plus exactement, le modèle semble présenter un profil de performances similaire au système biologique lorsque la cible est présentée peu de temps (150ms), mais lorsqu'elle est présentée plus longtemps (1s) les performances du système biologique ne semblent pas être influencées par l'excentricité de la cible. Des comparaisons quantitatives seront ici nécessaires pour confirmer ou infirmer ces similarités.<sup>10</sup>

Plusieurs étapes ont d'ores et déjà été identifiées afin de rendre le modèle à la fois plus performant et plus proche d'une certaine réalité neurologique.

La première consistera en la modification de la prédiction du modèle (à l'heure actuelle, la prédiction correspond à deux coordonnées "certaines" où la cible devrait être présente) en prédiction probabiliste. Cette transition permettrait de traiter la perception du modèle comme une carte de probabilité (ou de chaleur) où chaque point de l'espace est relié à une probabilité de contenir la cible. Ainsi la prédiction ne sera pas réalisée sur un point précis de l'espace mais sur un ensemble de points dont l'écart-type devrait augmenter avec l'excentricité (d'après les résultats observés sur la figure A.18). De plus, cette carte de probabilité se mettant à jour à chaque nouvelle saccade (puisque une nouvelle observation de l'environnement est alors réalisée), le problème de recherche de la localisation précise de la cible devrait se résoudre de lui-même en explorant tour à tour chacune des localisations les plus probables.<sup>1,9</sup>

Une seconde étape sera de reconstruire entièrement les méthodes d'apprentissage afin de non seulement augmenter fortement les performances du *classifieur*, pour l'instant loin d'approcher les performances standards des modèles *machine learning* de vision artificielle, mais aussi de tenter d'augmenter au maximum les performances du *détecteur* pour tenter d'atteindre les performances biologiques. L'une des solutions envisageables pour ce dernier point consiste en la transition d'un apprentissage indépendant des deux couches, comme c'est le cas actuellement, vers un modèle d'apprentissage utilisant la sortie (sous la forme d'une carte de probabilité) du *classifieur* pour guider l'entraînement (c'est à dire s'en servir comme d'un label) du *détecteur*. Cette optique permettrait non seulement d'améliorer les performances du modèle, mais aussi d'augmenter son autonomie vis à vis de l'utilisateur humain en produisant lui-même une partie de ses labels d'apprentissage.

Lorsque le modèle présentera des performances jugées suffisantes, l'étape suivante sera probablement de le mettre à l'épreuve dans des conditions pour lesquelles il n'aura pas été programmé afin de tester la robustesse de la généralisation obtenue grâce à l'apprentissage automatisé. Pour cela, l'un des moyens à notre disposition est d'altérer la qualité de l'image utilisée en entrée via l'introduction de bruit. Idéalement, ce bruit devra être construit pour suivre des règles écologiques, c'est à dire qu'il devra modéliser le bruit que l'on peut retrouver lors d'une perception naturelle de l'environnement visuel. Deux méthodes sont pour le moment envisagées : introduire des pseudo-stimuli créés en "mélangeant" des stimuli utilisés pour l'apprentissage (permettant d'obtenir des régions comprenant les mêmes variances de niveaux de gris que lors de la présence d'un stimulus réel mais sans présenter d'information pertinente) et/ou créer un bruit corrélé par transformation mathématiques.<sup>9</sup>

---



# Bibliography

- [1] Nicholas J Butko and Javier R Movellan. “Infomax control of eye movements”. In: *Autonomous Mental Development, IEEE Transactions on* 2.2 (2010), pp. 91–107.
- [2] Rachel N. Denison et al. “Functional mapping of the magnocellular and parvocellular subdivisions of human LGN”. In: *NeuroImage* 102.P2 (2014), pp. 358–369. ISSN: 10959572. DOI: [10.1016/j.neuroimage.2014.07.019](https://doi.org/10.1016/j.neuroimage.2014.07.019). arXiv: [15334406](https://arxiv.org/abs/15334406).
- [3] Sylvain Fischer et al. “Self-invertible 2D log-Gabor wavelets”. In: *International Journal of Computer Vision* 75.2 (2007), pp. 231–246. DOI: [10.1007/s11263-006-0026-8](https://doi.org/10.1007/s11263-006-0026-8). URL: <http://invibe.net/LaurentPerrinet/Publications/Fischer07cv?action=AttachFile\&do=view\&target=Fischer07cv.pdf>.
- [4] Jeremy Freeman and Eero P. Simoncelli. “Metamers of the ventral stream”. In: *Nature Neuroscience* 14.9 (2011), pp. 1195–1204. ISSN: 10976256. DOI: [10.1038/nn.2889](https://doi.org/10.1038/nn.2889).
- [5] Karl Friston et al. “Perceptions as hypotheses: Saccades as experiments”. In: *Frontiers in Psychology* 3.MAY (2012), pp. 1–20. ISSN: 16641078. DOI: [10.3389/fpsyg.2012.00151](https://doi.org/10.3389/fpsyg.2012.00151). arXiv: [NIHMS150003](https://arxiv.org/abs/NIHMS150003).
- [6] Melvyn A. Goodale and David A. Westwood. “An evolving view of duplex vision: Separate but interacting cortical pathways for perception and action”. In: *Current Opinion in Neurobiology* 14.2 (2004), pp. 203–211. ISSN: 09594388. DOI: [10.1016/j.conb.2004.03.002](https://doi.org/10.1016/j.conb.2004.03.002).
- [7] Laurent Itti and Christof Koch. “A saliency-based search mechanism for overt and covert shifts of visual attention”. In: *Vision Research* 40.10-12 (2000), pp. 1489–1506. ISSN: 0042-6989. DOI: [10.1016/S0042-6989\(99\)00163-7](https://doi.org/10.1016/S0042-6989(99)00163-7).
- [8] Philip Kortum and Wilson S. Geisler. “Implementation of a foveated image coding system for image bandwidth reduction”. In: *SPIE Proceedings* 2657 (1996), pp. 350–360. ISSN: 0277786X. DOI: [10.1117/12.238732](https://doi.org/10.1117/12.238732).
- [9] J Najemnik and Wilson S. Geisler. “Optimal eye movement strategies in visual search”. In: *Nature reviews. Neuroscience* 434 (2005), pp. 387–391. URL: <http://dx.doi.org/10.1038/nature03390>.

- [10] M K Uddin, Y Ninose, and S Nakamizo. "Accuracy and precision of spatial localization with and without saccadic eye movements: A test of the two-process model". In: *Psychologia* 47 (2004), pp. 28–34. ISSN: 00332852. DOI: [10.2117/psysoc.2004.28](https://doi.org/10.2117/psysoc.2004.28).
- [11] John S. Werner and Leo M. Chalupa, eds. *The new visual neurosciences*. MIT Press. 2014, p. 1675. ISBN: 9780262019163.
- [12] Li Zhaoping. *Understanding vision : theory, models and data*. Oxford Uni. 2014, p. 383. ISBN: 9780199564668

# A Figures

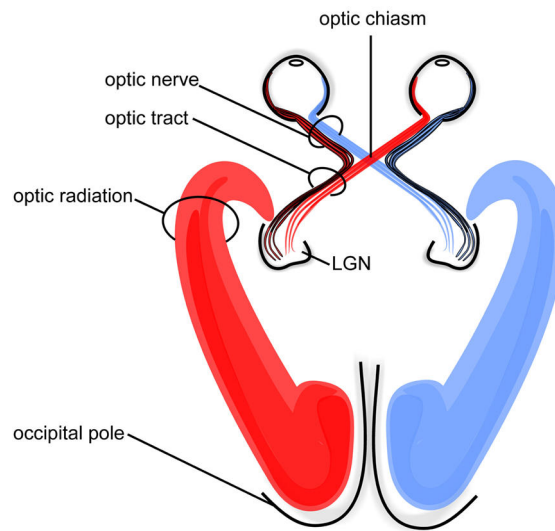


FIGURE A.1: Schéma des voies visuelles précorticales humaines (adapté de Hofer S. et al., 2010 via Wikimedia Commons [CC BY 3.0])

|  | Identifiant        | Système d'exploitation | Processeur                             | Mémoire vive | Carte graphique         |
|--|--------------------|------------------------|--|--------------|-------------------------|
| Machine physique                           | ASUS ROG G75VW     | Windows 7 64-bit SP1   | Intel Core i7-3610QM<br>2,30GHz (8CPU) | 8 GB (DDR3)  | NVIDIA GeForce GTX670M  |
| Machine virtuelle<br>(ressources allouées) | VirtualBox v.5.2.6 | Ubuntu 16.04           | 4 CPU, 90% des ressources              | 5298 Mo      | Support GPU non-utilisé |

TABLE A.1: Matériel physique et numérique utilisé pour réaliser les modélisations

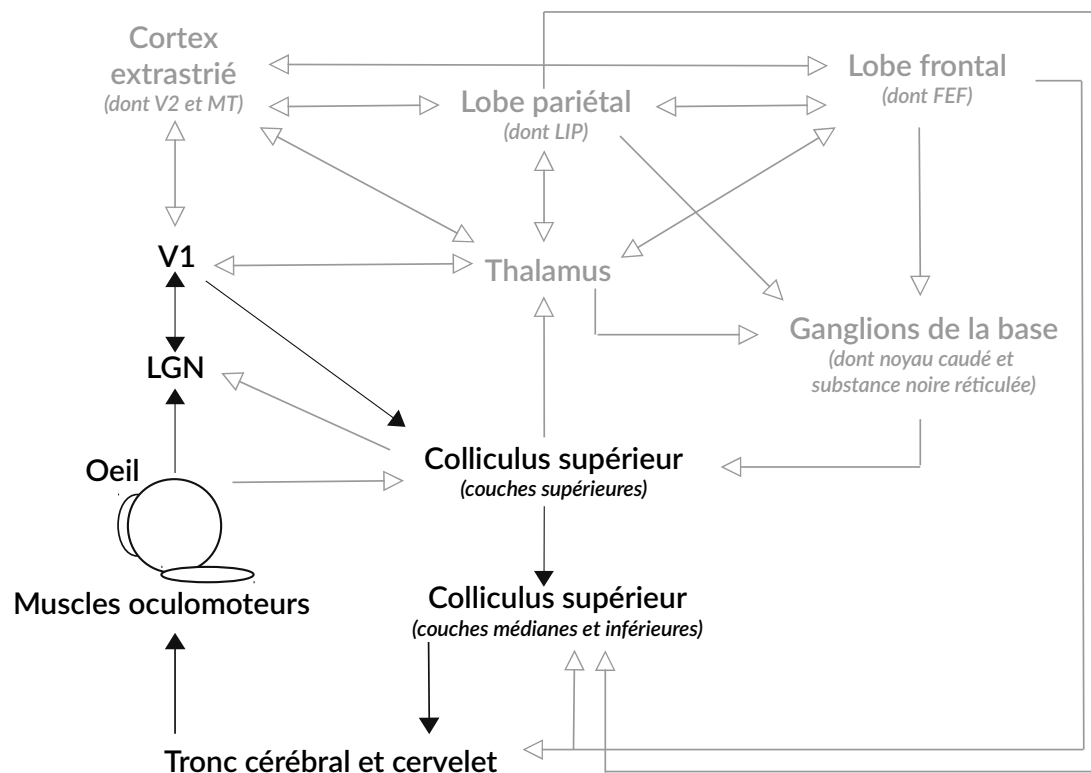


FIGURE A.2: Schéma simplifié du réseau nerveux impliqué dans la planification et l'exécution des saccades oculaires. Les parties noires correspondent aux composantes dont notre modèle tente de modéliser les fonctions, tandis que celles grises correspondent au reste du réseau (adapté de [12])

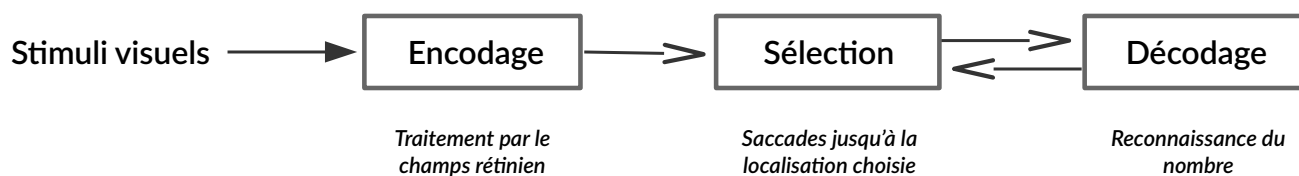


FIGURE A.3: Schéma simplifié du fonctionnement du système visuel avec son équivalence dans le modèle (adapté de [12])



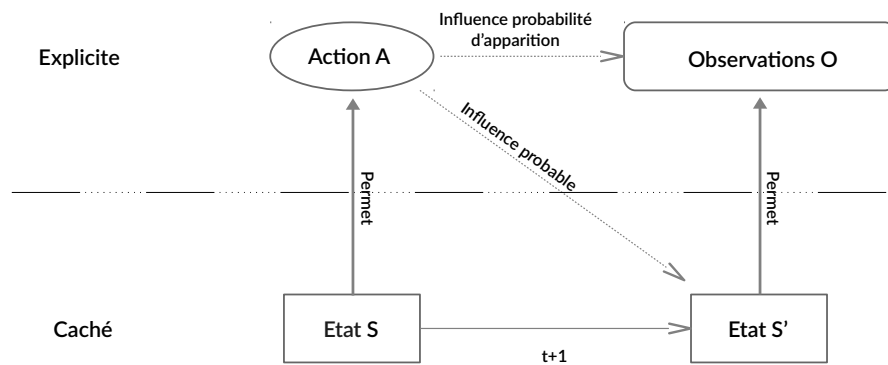


FIGURE A.4: Schéma des interactions entre l'agent et son environnement au cours du temps dans un modèle POMDP

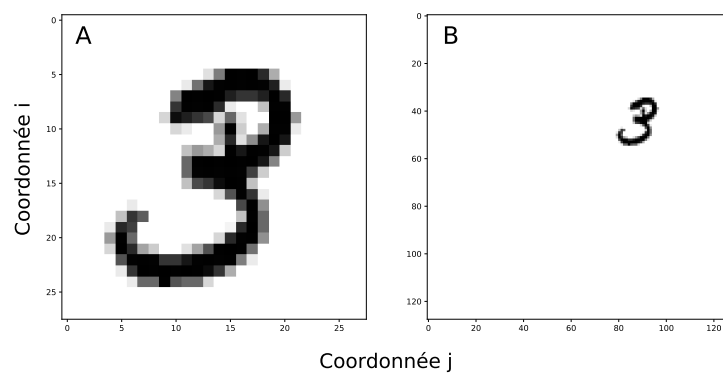


FIGURE A.5: **A.** Image originale tirée de MNIST ; **B.** Image après transformation géométrique et placé aux coordonnées  $(i = -20, j = 25)$

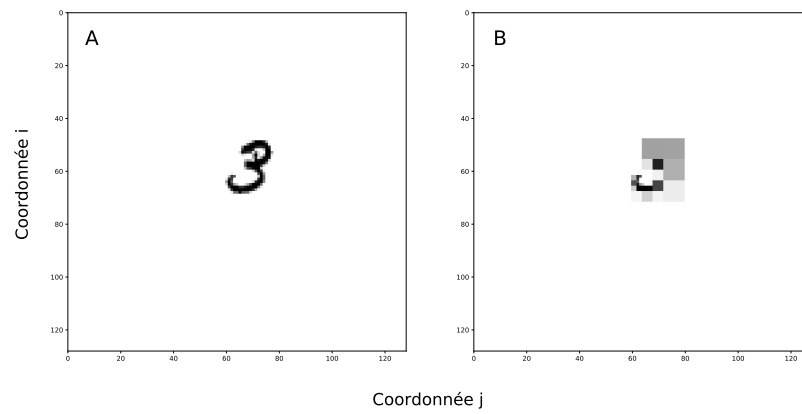


FIGURE A.6: **A.** Image avant application d'un filtre et placé aux coordonnées  $(i = -6, j = 6)$   
; **B.** Image après transformation par le filtre *Wavelets*

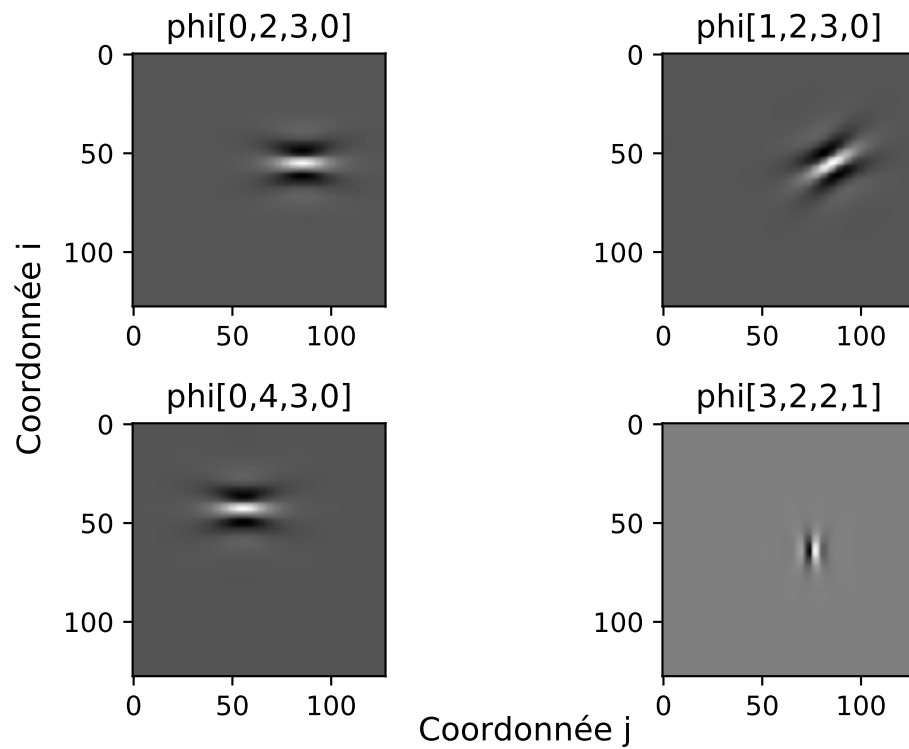


FIGURE A.7: Exemples de sous-unités LogGabor constituant le filtre *LogPolar*, avec leurs paramètres respectifs.

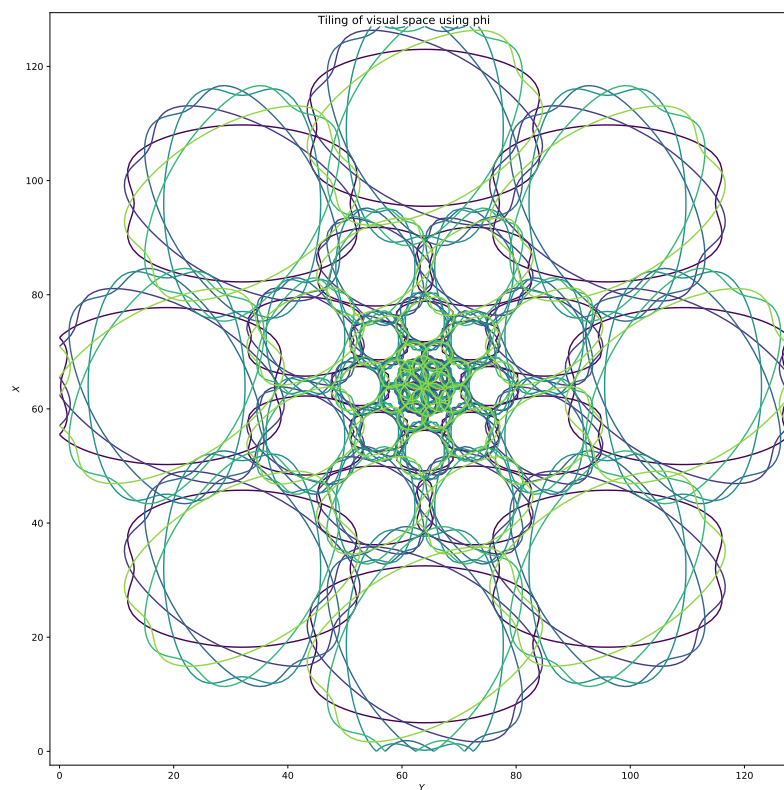


FIGURE A.8: Représentation graphique du filtre *LogPolar* ( $N_{\theta} = 6, N_{\text{orient}} = 8, N_{\text{scale}} = 5, N_{\text{phase}} = 2$ )

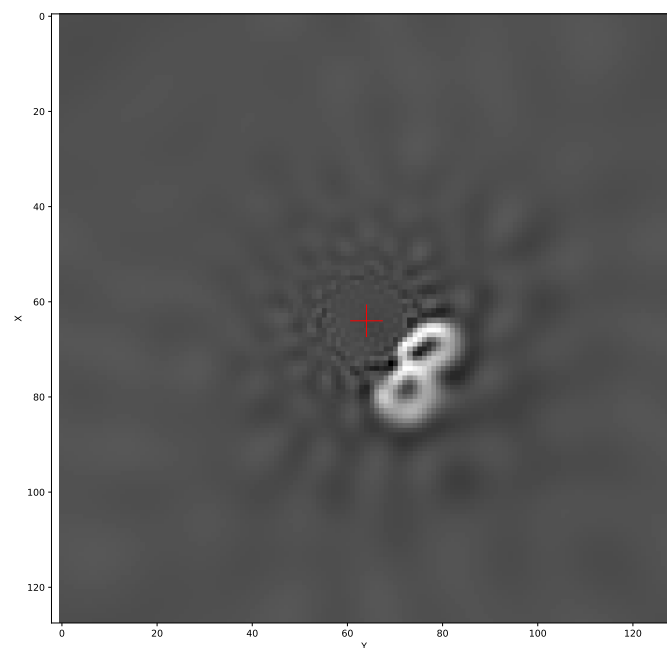


FIGURE A.9: Image après transformation par le filtre *LogPolar* ( $N_{\theta} = 6, N_{\text{orient}} = 8, N_{\text{scale}} = 5, N_{\text{phase}} = 2$ )

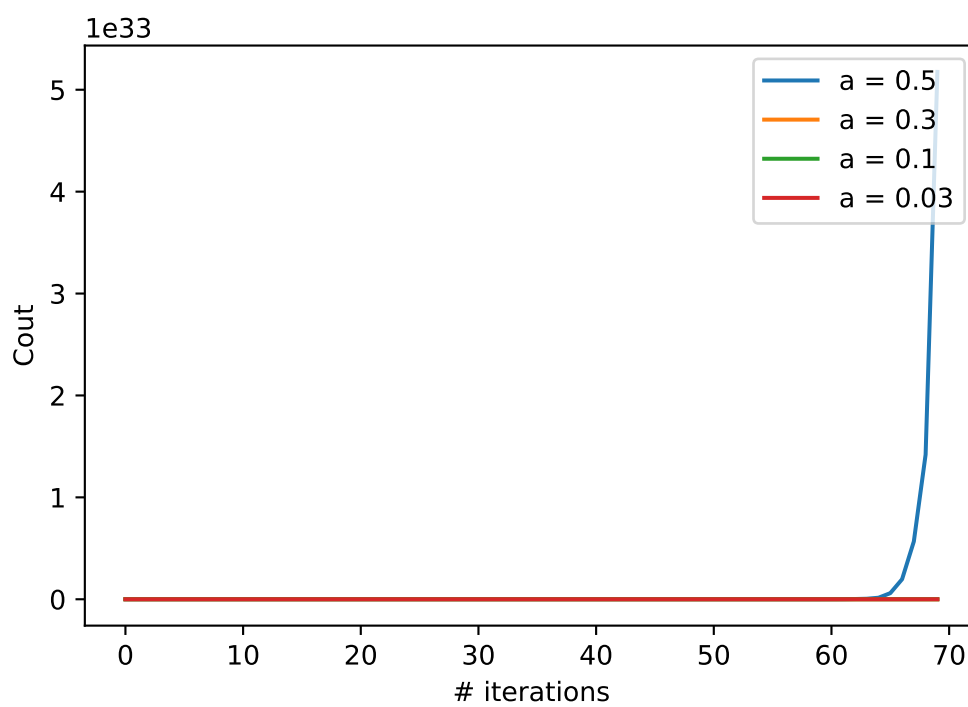
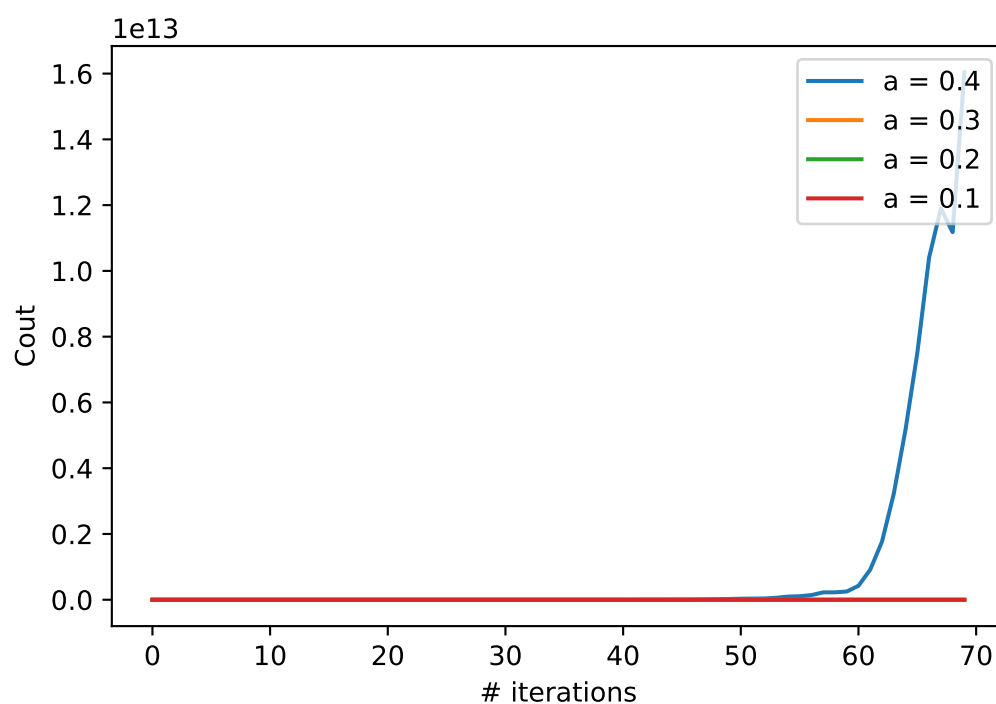
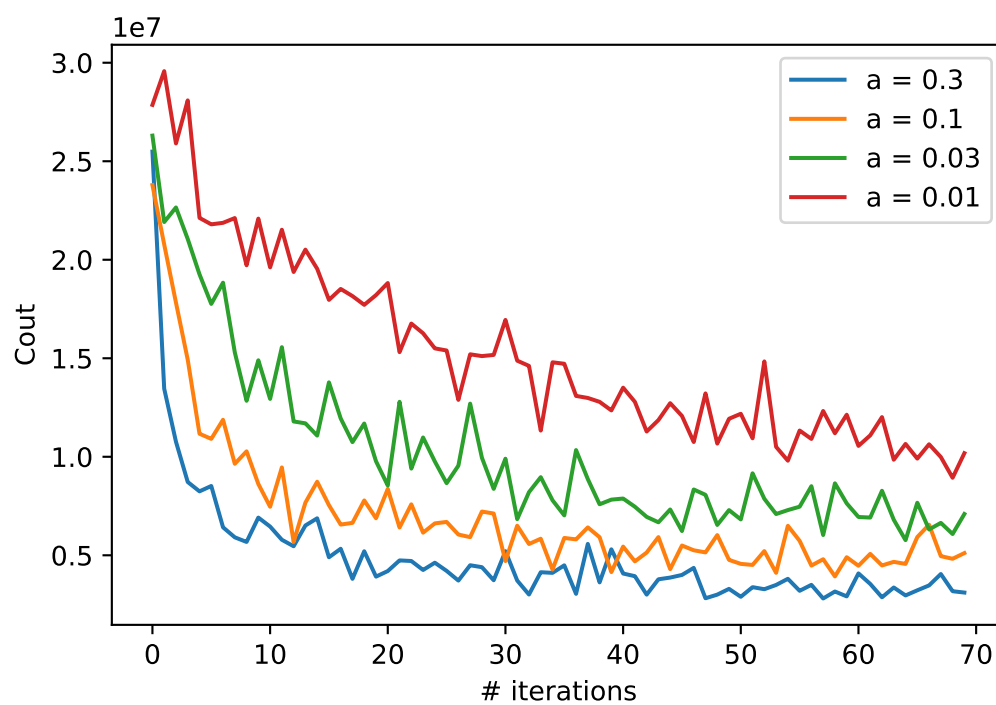


FIGURE A.10: Effet du paramètre alpha sur l'apprentissage dans le cadre d'un filtre *Wavelets*

FIGURE A.11: Effet du paramètre alpha sur l'apprentissage dans le cadre d'un filtre *Wavelets*FIGURE A.12: Effet du paramètre alpha sur l'apprentissage dans le cadre d'un filtre *Wavelets*

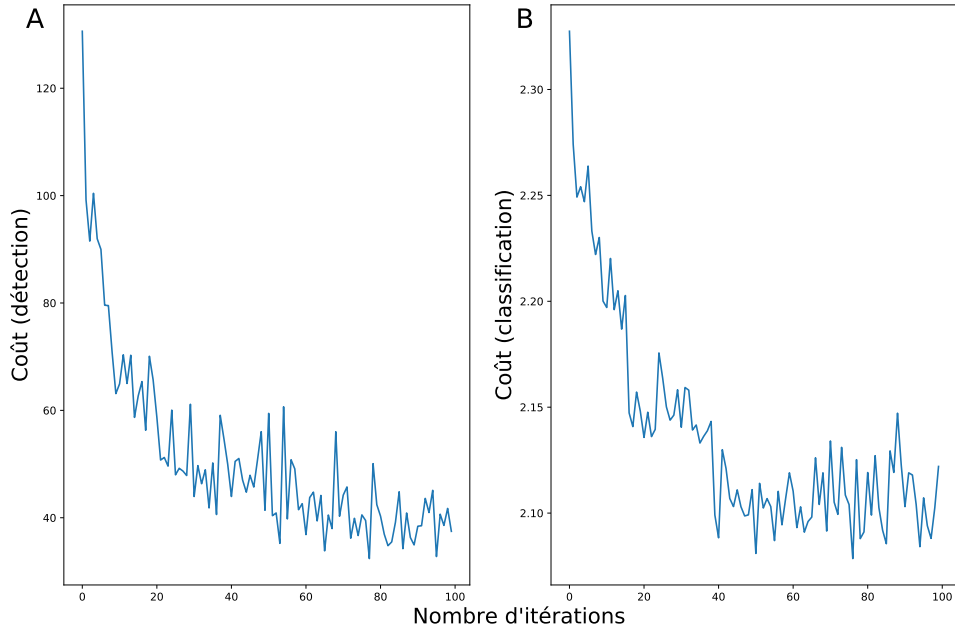


FIGURE A.13: Réduction du coût des couches *décteur* et *classifieur* lors de l'apprentissage, dans le cadre d'un filtre *LogPolar* (taille de la base d'apprentissage : 1000, nombre d'itérations : 100,  $\alpha_{detect} = 0.0015$ ,  $\alpha_{classif} = 0.3$ )

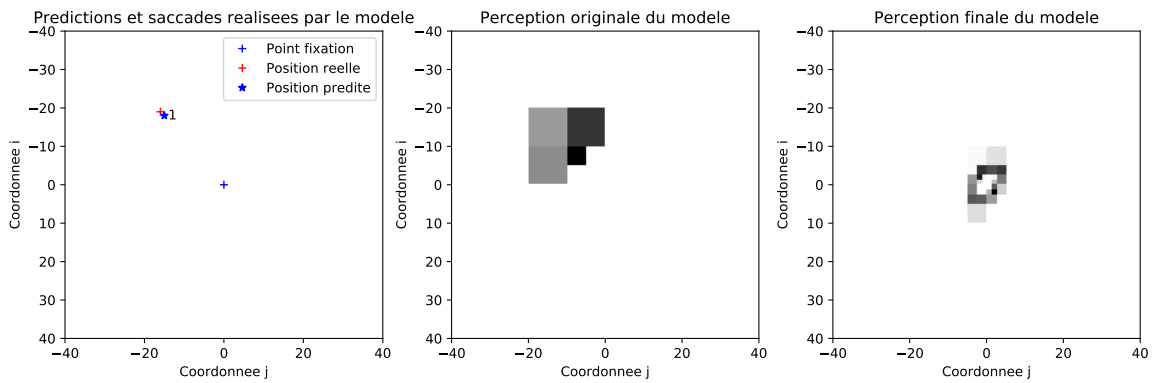


FIGURE A.14: Exemple de perception et comportement saccadique du modèle entraîné, dans le cadre d'un filtre *Wavelets*

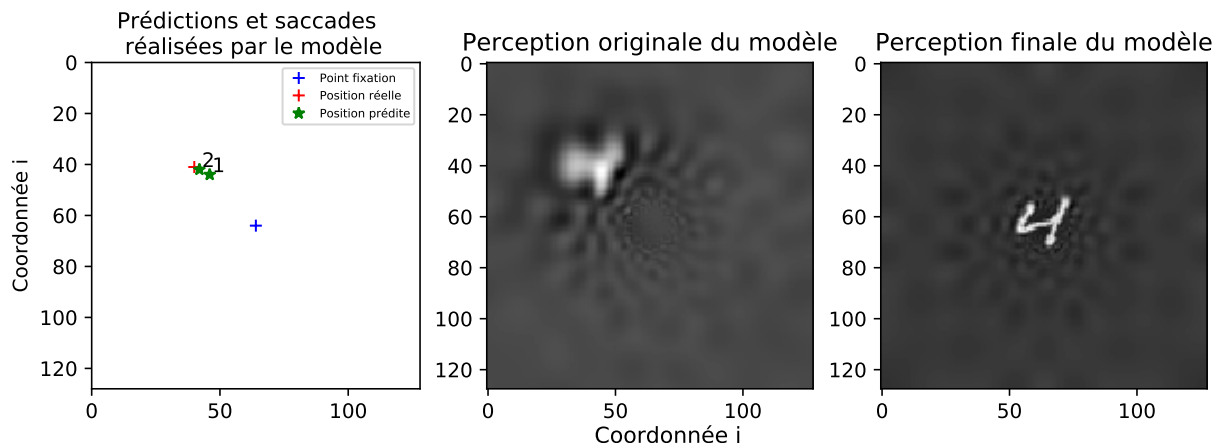


FIGURE A.15: Exemple de perception et comportement saccadique du modèle entraîné, dans le cadre d'un filtre *LogPolar*

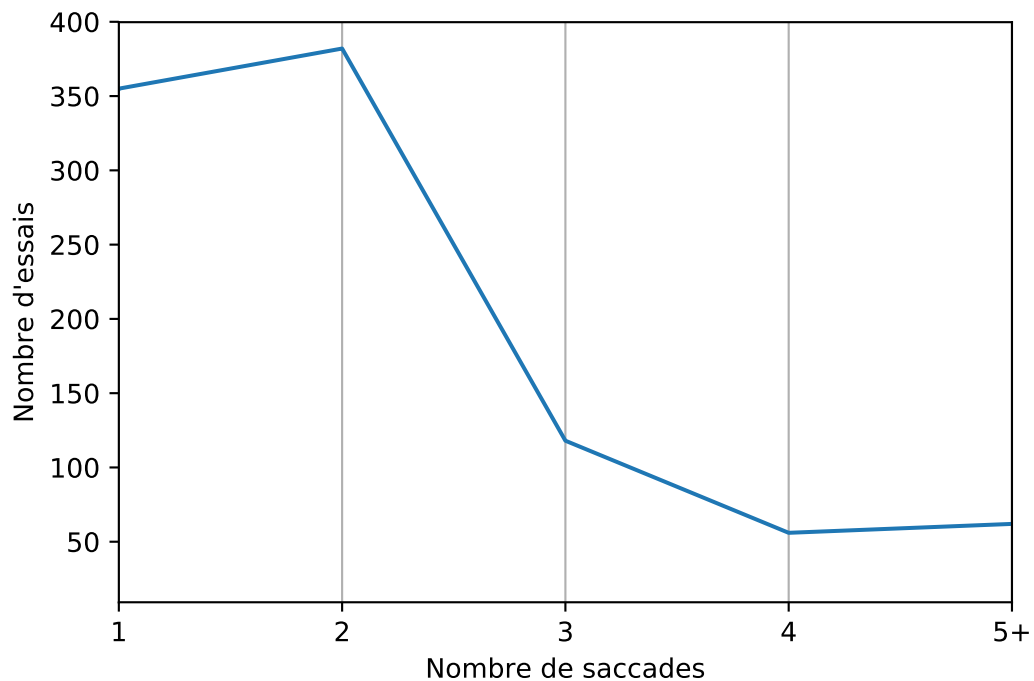


FIGURE A.16: Nombre de saccades nécessaires pour atteindre la position de la cible au cours de 1000 essais, dans le cadre d'un filtre *LogPolar* (taille de la base d'apprentissage : 1000, nombre d'itérations : 100,  $\alpha_{detect} = 0.0015$ ,  $\alpha_{classif} = 0.3$ )

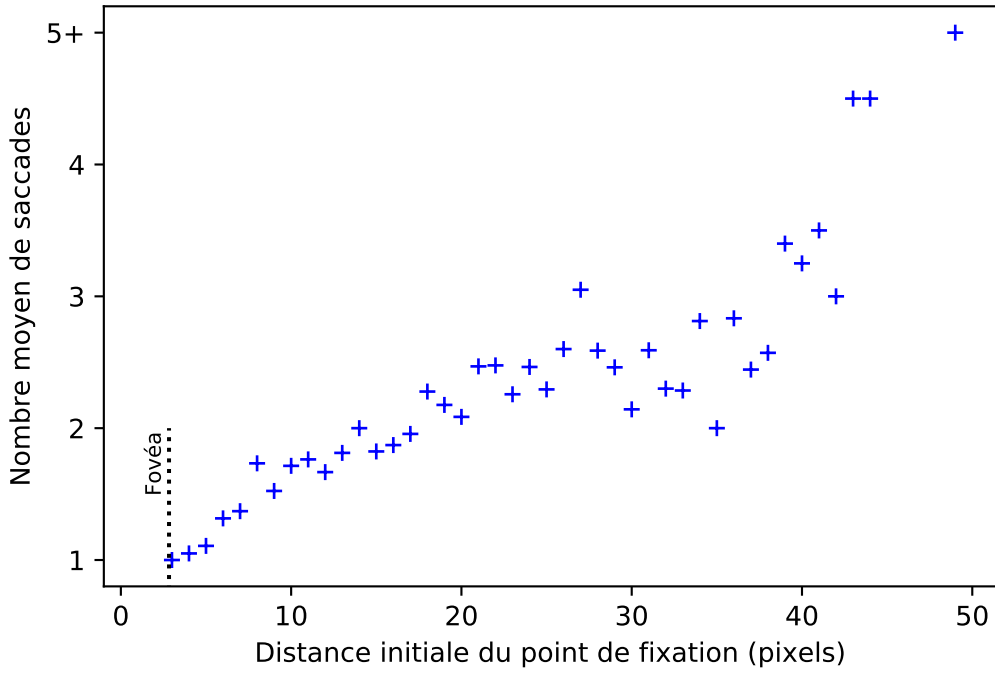


FIGURE A.17: Nombre moyen de saccades nécessaires pour atteindre la position de la cible en fonction de sa distance initiale du point de fixation au cours de 1000 essais, dans le cadre d'un filtre *LogPolar* (taille de la base d'apprentissage : 1000, nombre d'itérations : 100,  $\alpha_{detect} = 0.0015$ ,  $\alpha_{classif} = 0.3$ )

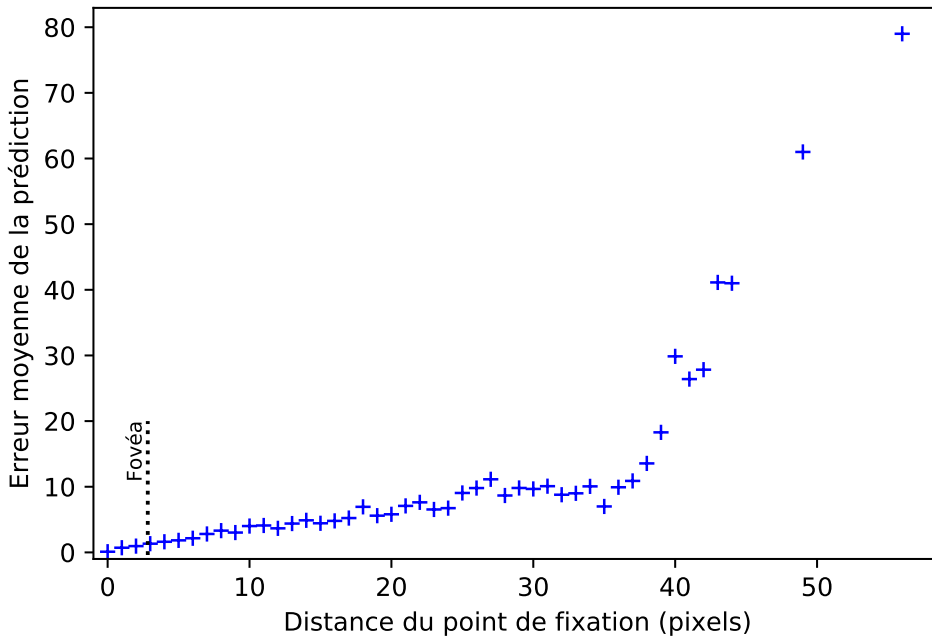


FIGURE A.18: Erreur moyenne lors de la prédiction de la position de la cible en fonction de sa distance du point de fixation au cours de 1000 essais, dans le cadre d'un filtre *LogPolar* (taille de la base d'apprentissage : 1000, nombre d'itérations : 100,  $\alpha_{detect} = 0.0015$ ,  $\alpha_{classif} = 0.3$ )



## B Code source

L'ensemble du code source du modèle, ainsi que de ce rapport et d'autres documents complémentaires (dont les notes personnelles) sont entièrement disponibles **en ligne**.