AIX-MARSEILLE UNIVERSITÉ

MÉMOIRE DE MASTER

Modélisation de la localisation de cible visuelle dans la voie magno-cellulaire

---

Auteur:

Pierre ALBIGÈS

Superviseur:

Emmanuel DAUCÉ

Un mémoire présenté 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

Modélisation de la localisation de cible visuelle dans la voie magno-cellulaire

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 termes de performance et d'adaptabilité à l'environnement. L'un de ces défauts est la puissance de calcul faramineuse nécessaire pour obtenir une vision artificielle haute performance.

Afin de proposer une solution à ce problème, nous nous sommes inspirés des agents markoviens pour développer un modèle 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 neuromimétique modifiant la perception de l'image par le modèle, avant que soit réalisé tout traitement d'information. Ce filtre, disponible selon 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

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. La vision est 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.[13]

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. Les cônes et les bâtonnets sont différentiables 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.[13]

Le champs visuel peut être divisé en deux parties. La vision centrale (environ 2° chez l'humain) repose anatomatiquement sur 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 bonne perception des couleurs. La vision périphérique, quant à elle, comprend majoritairement des bâtonnets. Elle 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). Plus précisemment, 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 par le nerf optique estimée à 10^2 bits/s, soit une réduction de plus de 99% de la quantité d'information). C'est ce qui a été théorisé par Barlow en 1961 sous le principe du codage efficace.[7,8,13,14]

Lors de l'exploration de son environnement visuel, un agent va pouvoir détecter des stimuli dans sa vision périphérique mais ne va pas y posséder 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.[8,13]

L'activité rétinienne est ainsi transmise le long des voies nerveuses visuelles jusqu'au cortex visuel, où sera réalisée la majorité du traitement des informations -notamment haut niveau- qu'elle code.[13]

Entre la rétine et le cortex visuel 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.[13]

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. 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.[2,13]

Cette multiplicité de voies visuelles est conservée au delà de V1, où l'on décrit deux voies sortantes: les voies **ventrale** (transportant et traitant majoritairement les informations provenant de la voie P) et **dorsale** (transportant et traitant majoritairement les informations provenant de la voie M).[4,6,13]

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.[4,6,13]

Parmi ce réseau dorsal, on trouve l'**aire intrapariétale latérale** (LIP), qui recoit 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 champs visuel périphérique. Des travaux ont d'ailleurs relié l'activité des neurones du LIP à la représentation spatiale des objets visuels et à la plannification des saccades oculaires.[13]

La planification et l'exécution des saccades oculaires implique un réseau cérébral complexe et dont on ne connait 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).[14]

1.2 Vision artificielle

La vision représentant notre modalité perceptive principale, les aires dévouées à traiter ses informations occupent une part significative de notre cortex cérébral (jusqu'à 50% chez certains primates). L'étudier, de ses aspects les plus moléculaires jusqu'aux fonctions les plus intégrées, permet de mieux comprendre le fonctionnement général de notre système nerveux. Parmi ces domaines d'étude, les neurosciences computationnelles se basent sur les données expérimentales (anatomiques, physiologiques et comportementales) pour proposer des modèles mathématiques sur le fonctionnement d'une partie ou de l'ensemble de la modalité visuelle. Idéalement, ces modèles doivent pouvoir expliquer l'activité visuelle dans l'ensemble des contextes observables, mais aussi pouvoir prédire son comportement dans de nouveaux contextes jamais rencontrés jusqu'alors. Ces modèles permettent rarement la compréhension exhaustive du système, mais ils peuvent néanmoins permettre d'identifier des défauts ou des zones d'ombre à notre compréhension et donc diriger les études expérimentales vers ces points.[14]

Au delà de la possible validation ou invalidation des modèles, les modèles mathématiques 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.[11]

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) et s'inspire fortement du fonctionnement de la voie visuelle magno-cellulaire. 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 précisémment 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 plus élevée.[14]

Pour cela, plusieurs 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 entièrements cachés.[1,7]

Les **modèles de contrôle**, tentent quant à eux de prédire quelles règles le système devra suivre pour réaliser au mieux une tâche. Il s'agit de définir une politique qui, dans un contexte donné, choisit l'action la plus à même de remplir l'objetif (ici, il s'agit de placer la cible visuelle au centre de la rétine). A chaque saccade, de nouvelles informations sur l'environnement sont collectées et permettent de changer l'opinion de l'agent sur le monde, et de définir dans ce nouveau contexte la meilleure action à produire.[1]

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 language 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).[1] 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.[1,11]

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.[1]

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).[1,11]

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

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).[1]

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

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

Ainsi, dans le cas qui nous intéresse, l'état de l'environnement est constitué de la position et de l'identité de la cible visuelle, et le but des actions (les saccades) est de permettre d'estimer au mieux (après t observations) l'état de l'environnement.

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 sur une image au fond blanc de 128x128 pixels afin de réduire sa taille relative 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 champ 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). Chaque méthode présente des avantages et des inconvénients, mais dans les deux cas la transformation mathématique imposée par le filtre est calculée à l'avance, permettant de l'appliquer à chaque nouvel exemple et après chaque saccade oculaire (donc à chaque nouvelle observation), tout en économisant au maximum la puissance de calcul disponible.[8]

Le **filtre *Wavelets*** consiste en un encodage pyramidal (c'est donc une approche purement mathématiques): on applique sur l'image une grille de résolution, définissant des anneaux concentriques contenants des superpixels dont la taille (et le nombre de pixels qu'ils contiennent) augmente avec

l'excentricité de l'anneau. Le niveau de gris d'un superpixel correspondant à la moyenne des pixels qu'il contient, la résolution de l'image à laquelle est appliquée ce filtre diminue à chaque nouvel anneau, et donc par palier (les effets de ce filtre sont visibles sur la figure A.6).[8]

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 précisemment, 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.[3,10]

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

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

2.4 Apprentissage supervisé

Notre modèle de reconnaissance visuelle repose sur un **régresseur linéaire multivarié** qui utilise comme entrée sa perception visuelle rétinienne et prédit la position dans l'espace visuel (x, y) de la cible. L'apprentissage des paramètres du modèle se fait par un apprentissage supervisé sous la forme d'une **descente de gradient**. Cette méthode de *machine learning* à 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_θ (équation 2.4) sur la répartition des stimuli en réalisant le produit scalaire du vecteur d'entrée $\overrightarrow{x}$ (l'image après transformation par le filtre rétinien) à une matrice de poids θ, puis en ajoutant éventuellement un biais b.

$$h_\theta(x) = \theta^T x + b \tag{2.4}$$

Ces poids sont ensuite optimisés par descente de gradient (équation 2.5), où ils sont comparés à la position réelle y de la cible (correspondant aux informations que l'on veut apprendre, ici ses coordonnées (i, j)) pour un nombre d'exemples (m au total) et d'itérations fixées. Le paramètre d'apprentissage α influence très fortement l'entrainement et sa valeur doit être adaptée pour éviter un sous- ou un sur-apprentissage (révélant respectivement une valeur trop faible ou trop importante).

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

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 décroitre au cours du temps, l'optimisation du modèle peut se faire en tentant de la réduire au maximum.

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

Nous construisons ainsi un modèle linéaire à une couche dont le comportement s'insipire de celui d'un agent POMDP (algorithme A.10). Pour cela, nous réalisons l'entraînement avec en entrée deux vecteurs: l'entrée x, produit d'une transformation mathématique de l'image après application de l'un des filtres rétiniens, ainsi que le label y comprenant les coordonnées (x, y) de la cible.

Cette couche, que l'on appellera (*détecteur*), est ainsi entraînée à prédire la position du stimulus dans l'image perçue à partir du champs rétinien, permettant de réaliser une saccade jusqu'aux coordonnées prédites et d'approcher la cible de sa fovéa. Chaque saccade vise donc à améliorer la détection du signal.[5]

Une seconde couche, que l'on appellera (*classifieur*), est actuellement en cours de développement pour ce modèle mais son implémentation n'étant pas achevée au moment de l'écriture de ce rapport, celle-ci n'apparaitra de que manière anecdotique dans les figures. L'objectif est d'entraîner cette nouvelle couche à prédire la catégorie du stimulus dans l'image perçue à partir du champs rétinien et d'amorcer l'arrêt de l'exploration de l'image lorsque sa prédiction présente une certitude assez élevée (elle deviendrai donc la condition d'arrêt en ligne 6 de l'algorithme A.10).

3 Résultats

Cette partie contient des résultats préliminaires, le modèle étant encore en cours de développement et d'optimisation lors de l'écriture de ce rapport. Les idées émises dans ce chapitre ainsi que le suivant restent pour le moment des hypothèses qui devront nécessairement être confirmées lors de travaux ultérieurs.

3.1 Apprentissage supervisé

L'étude d'étalonnage du paramètre d'apprentissage α (équation 2.5) permet de rendre compte de son importance sur l'efficacité de l'apprentissage et du modèle. On peut ainsi observer que certaines valeurs entraînent un sur-apprentissage très important (figures A.11), tandis que d'autres semblent représenter des valeurs utilisables, voire optimales, pour réaliser l'apprentissage (figure A.12).
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 α (respectivement α_{detect} et $\alpha_{classif}$) et l'on peut donc calculer leurs coûts indépendamments (figure A.13).
On peut observer qu'indépendamment du filtre utilisé et du nombre de couches entraînées, l'apprentissage par descente de gradient permet d'optimiser le modèle, en modifiant les poids θ itération après itération. Cette optimisation est révélée par une diminution graduelle du coût, représentant une différence entre la réalité et ce qui est prédit par le modèle.

3.2 Prédiction de la position

Après avoir été entraîné, le modèle semble capable de détecter la cible dans son environnement visuel et de prédire précisemment sa position dans l'espace (figure A.14 et A.15). L'agent est ensuite capable

d'utiliser ces connaissances pour réaliser une saccade jusqu'aux coordonnées prédites de la cible visuelle, ce qui modifie en conséquence sa perception de l'environnement et donc de la cible.

L'agent doit parfois réaliser plus d'une saccade pour atteindre la cible (figure A.16). On observe un nombre important d'essais où une ou deux saccades sont suffisantes pour atteindre sa position réelle, puis un nombre décroissant d'essais pour un nombre de saccades supérieure à 3.
Le nombre de saccades pour atteindre la cible semble dépendre de la distance à laquelle elle se trouve lorsqu'elle est présentée pour la première fois à l'agent (figure A.17). On observe que cette relation semble suivre une loi linéaire croissante jusqu'à un premier seuil (pour une distance initiale d'environ 25 pixels) après lequel on observe un plateau du nombre moyen de saccades. On peut ensuite observer un second seuil (pour une distance initiale d'environ 35 pixels) où le nombre moyen de saccades augmente à nouveau fortement. A noter que l'agent ne réalise pas de saccade lorsque la cible est directement présentée dans sa fovéa.
Cette relation entre le nombre de saccades nécessaires pour placer la cible dans la fovéa et la distance initiale à laquelle est présentée la-dite cible pourrait dépendre de la taille des erreurs que va commettre l'agent lorsqu'il va prédire la position de la cible (figure A.18). On observe en effet que cette erreur tend à croitre linéairement avec la distance à laquelle se trouve la cible lors de la prédiction. On retrouve ici l'un des seuils de la figure précédente (pour une distance d'environ 35 pixels) où l'erreur de prédiction augmente fortement et en rupture avec la croissance qui avait lieu jusque là. Cette augmentation de la taille de l'erreur peut se traduire par une diminution de la précision des prédictions de la position de la cible. Plus la cible est éloignée de sa fovéa, plus l'agent semble imprécis.

4 Discussion et perspectives

Notre modèle semble ainsi capable de s'inspirer du fonctionnement d'un modèle POMDP (figure A.4) pour réaliser 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 accuité possible, et donc avec le plus grand taux de réussite. Ce comportement ''atteindre-et-décrire'' instauré par le fonctionnement séquentiel semble cohérent avec ce que l'on peut observer dans les systèmes biologiques.[9,13]

En complément à ces observations sur le comportement général du modèle, nous avons pu observer que plus une cible est éloignée de la fovéa de l'agent lorsqu'elle est présentée, 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 qui a pu être observée lors d'études psychophysiques utilisant l'*eye-tracking* chez l'Humain. 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 visuel humain dans un contexte où une cible visuelle est présentée peu de temps (150ms). Lorsqu'une cible visuelle est présentée pendant un temps relativement long (1s), les performances du système visuel humain concernant la précision des prédictions et des saccades oculaires ne semblent pas être influencées par l'excentricité de la cible par rapport à la fovéa de l'agent. Des comparaisons quantitatives seront ici nécessaires pour confirmer ou infirmer ces similarités, mais aussi afin d'explorer certaines autres caractéristiques du comportement oculaire naturel, tels que la distance optimale de saccade ou l'asymétrie spatiale des mouvements oculaires.[9,12]

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 réalisation de ces étapes pourra concerner des travaux ultérieurs, pouvant prendre la forme d'un sujet de thèse.

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 amélioration permettrait de traiter la perception du modèle comme une carte de probabilité 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 une distribution de probabilités dont l'ecart-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 (puisqu'une nouvelle observation de l'environnement est alors réalisée), il devient possible de cumuler les probabilités au cours du temps pour aboutir à un modèle doté d'une mémoire.[1,9]

Une seconde étape sera de reconsidérer 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-stimulis 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.[9]

Finalement, lorsque le modèle aura été entraîné, évalué et optimisé dans un ensemble de contextes artificiels, il deviendra envisageable de soumettre son fonctionnement à des contextes écologiques en l'implémentant dans un système autonome. L'introduction effective du paradigme POMDP permettra à l'agent de choisir l'action la plus adaptée en fonction de sa perception de son environnement et de la tâche à réaliser, tandis la présence de l'apprentissage automatisé et de l'approche neuromimétique (vision rétinienne) permettra de conserver une performance élevée (notamment en adaptabilité et en vitesse de réaction) tout en diminuant au maximum la puissance de calcul et la consommation d'énergie nécessaires à son fonctionnement.[11]

En prenant pour example un agent robotisé mobile, tel qu'un drone, dont la tâche serait de reconnaitre un objet ou un visage dans son environnement visuel, ce modèle pourrait lui permettre de choisir de façon active et autonome quelle action choisir (réaliser une saccade et si oui en visant quelle position, réaliser un mouvement et si oui lequel parmi 6 directions et 6 rotations disponibles) suivant son contexte environnemental et les informations qui lui sont accessibles (position estimée de l'agent par rapport à la cible, taux de certitude de l'identification, environnement observé, etc) afin de réaliser sa tâche de façon optimale.[11]

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. arXiv: 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. 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.

[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. arXiv: 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.

[7] Laurent Itti and Christof Koch. "A saliency-based search mechanism for overt and covert hifts of visual attention". In: *Vision Research* 40.10-12 (2000), pp. 1489–1506. ISSN: 0042-6989. DOI: 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.

[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] Laurent Perrinet. *SLIP : a Simple Library for Image Processing*. URL: https://github.com/bicv/SLIP.

[11] Christian Potthast et al. "Active multi-view object recognition: A unifying view on online feature selection and view planning". In: *Robotics and Autonomous Systems* 84 (2016), pp. 31–47. ISSN: 09218890. DOI: 10.1016/j.robot.2016.06.013. URL: http://dx.doi.org/10.1016/j.robot.2016.06.013.

[12] 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.

[13] John S. Werner and Leo M. Chalupa, eds. *The new visual neurosciences*. MIT Press. 2014, p. 1675. ISBN: 9780262019163.

[14] Li Zhaoping. *Understanding vision : theory, models and data*. Oxford Uni. 2014, p. 383. ISBN: 9780199564668

A Figures

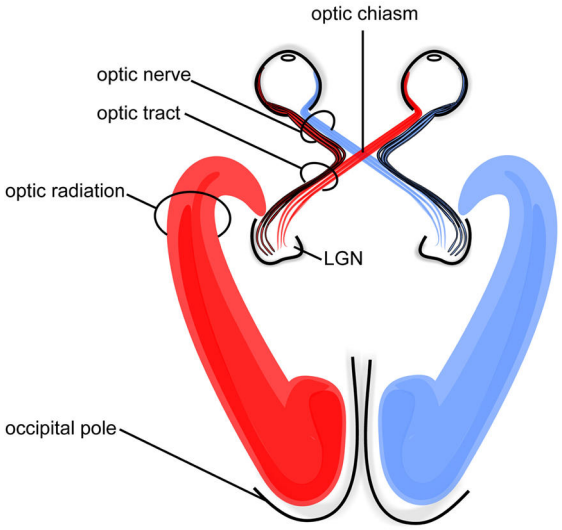

FIGURE A.1: Schéma des voies visuelles précorticales humaines (adapté de Hofer S. et al., 2010 via Wikimedia Commons [CC BY 3.0])

	Identifiant	Système d'explotation	Processeur	Mémoire vive	Carte graphique
Machine physique	ASUS ROG G75VW	Windows 7 64-bit SP1	Intel Core I7-3610QM 2,30GHz (8CPU)	8 GB (DDR3)	NVIDIA GeForce GTX670M
Machine virtuelle (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

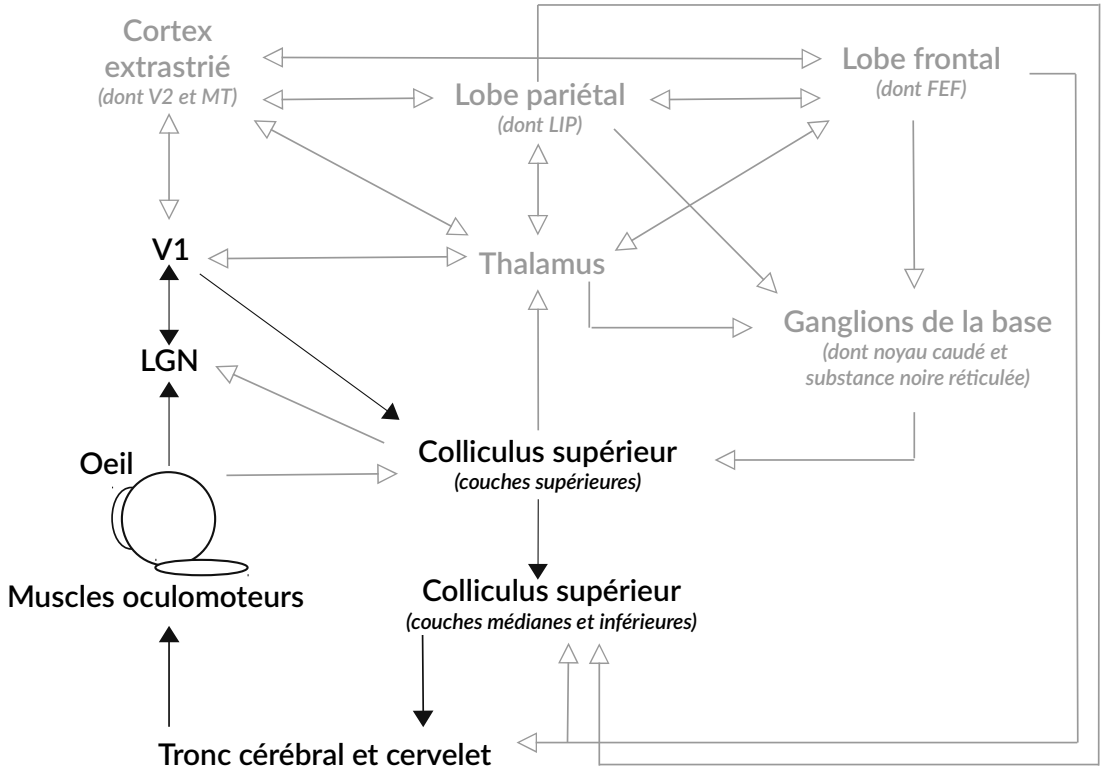

FIGURE A.2: Schéma simplifié du réseau nerveux impliqué dans la planification et l'exécution des saccades oculaires. Les parties noires correspondent aux composantes dont notre modèle tente de modéliser les fonctions, tandis que celles grises correspondent au reste du réseau (adapté de [14])

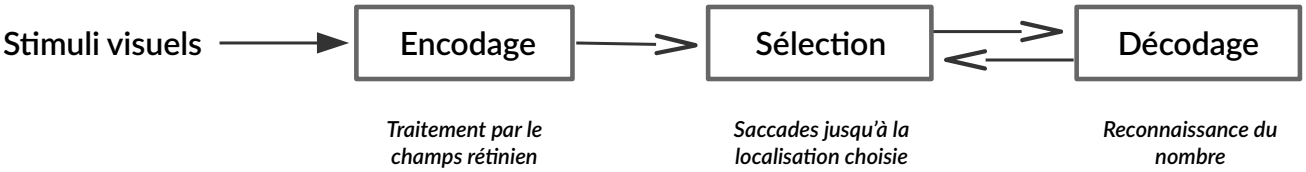

FIGURE A.3: Schéma simplifié du fonctionnement du système visuel avec son *équivalence dans le modèle* (adapté de [14])

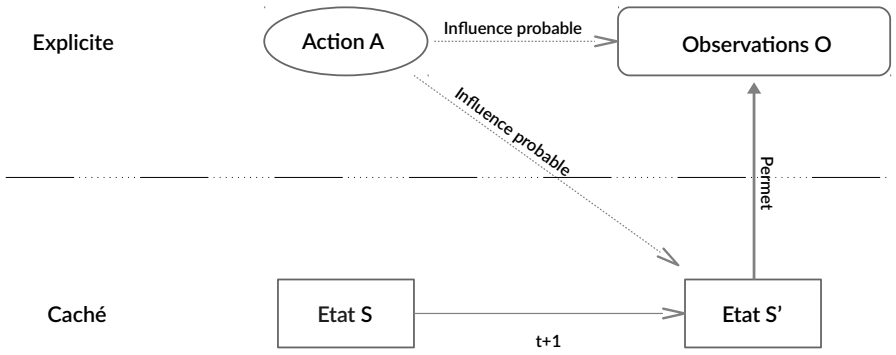

FIGURE A.4: Schéma des interations entre l'agent et son environnement au cours du temps dans un modèle POMDP

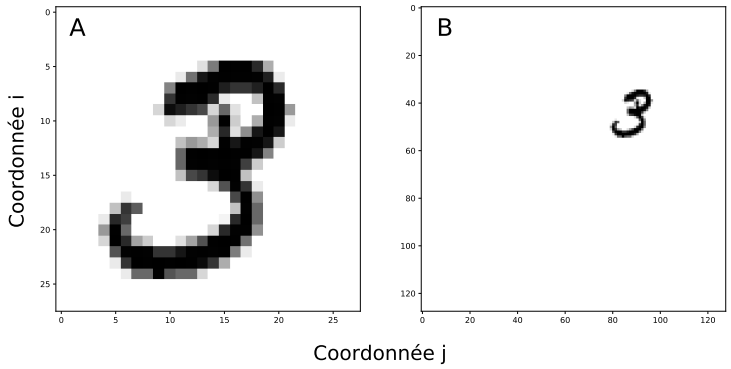

FIGURE A.5: **A.** Image originale tirée de la base MNIST ; **B.** Image après translation aux coordonnées $(i = -20, j = 25)$ sur un fond blanc 128×128

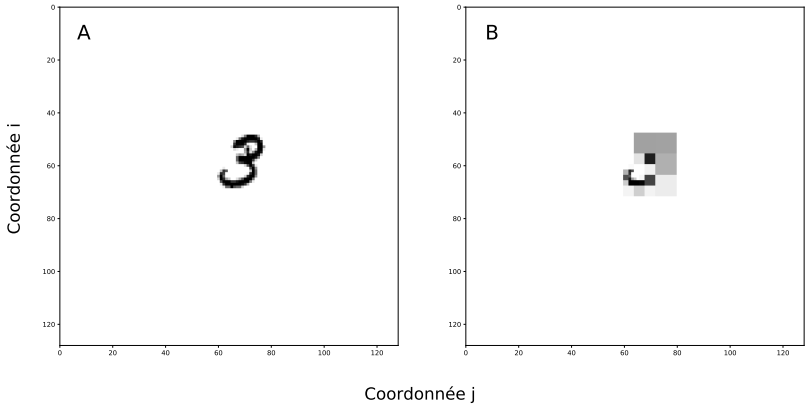

FIGURE A.6: **A.** Image avant encodage pyramidal par ondelettes avec cible positionnée aux coordonnées $(i = -6, j = 6)$; **B.** Image reconstruite d'après les valeurs des 70 coefficients d'ondelettes (taux de compression: $1 - \frac{70}{128 \times 128}$)

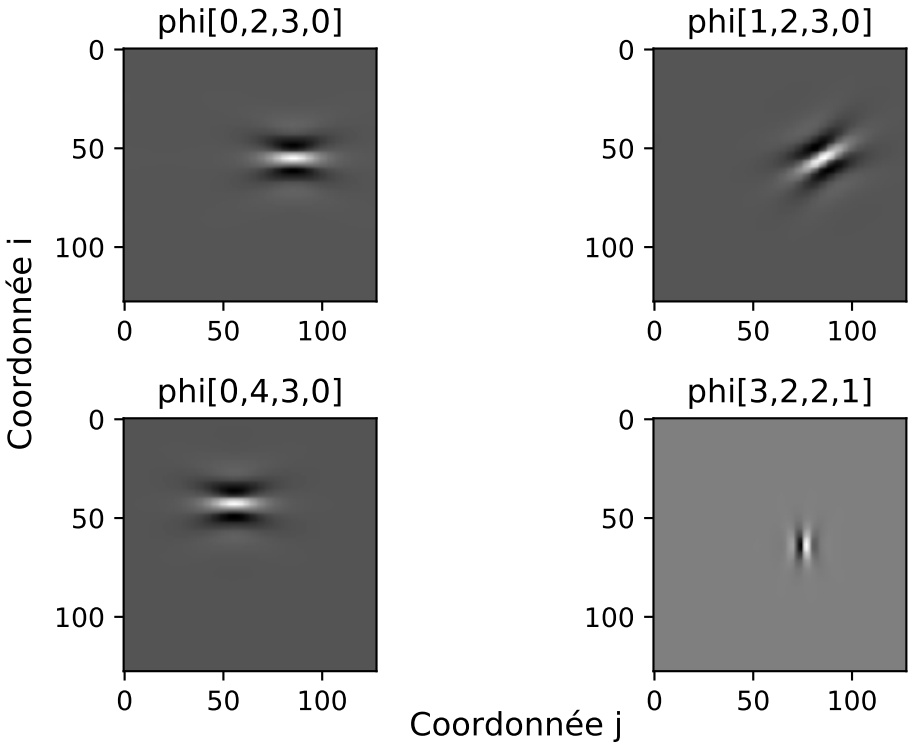

FIGURE A.7: Exemples de filtres LogGabor tels qu'appliqués sur l'image originale, avec leurs paramètres respectifs (nombre total de filtre constituant *LogPolar*: 400)

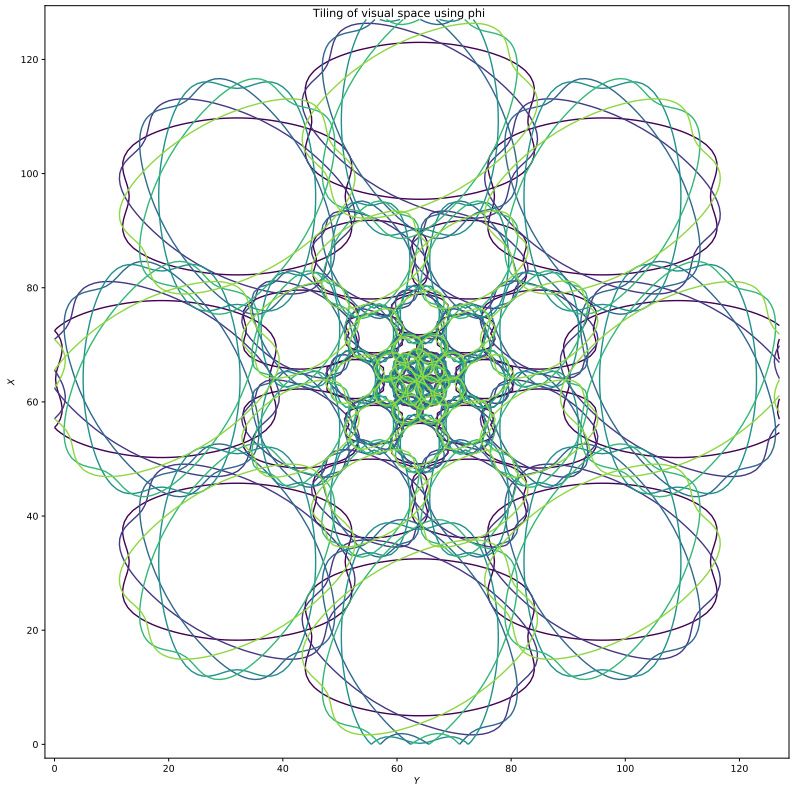

FIGURE A.8: Représentation graphique du filtre *LogPolar* ($N_{theta} = 6$,$N_{orient} = 8$,$N_{scale} = 5$,$N_{phase} = 2$)

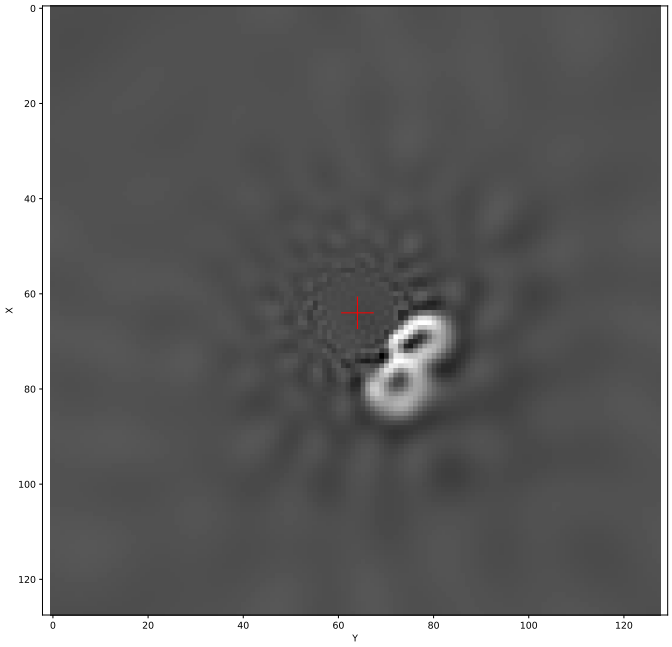

FIGURE A.9: Image reconstruite après transformation par le filtre *LogPolar* ($N_{theta} = 6$,$N_{orient} = 8$,$N_{scale} = 5$,$N_{phase} = 2$)

$Image28 \leftarrow new\_MNIST\_example$;
$(i, j) \leftarrow tirage\_aleatoire$;
$Image128 \leftarrow creer\_Image128(Image28, (i, j))$;
$x \leftarrow transformation(Image128(0, 0))$;
$\hat{i}, \hat{j} \leftarrow (0, 0)$;
while $\|(i - \hat{i}, j - \hat{j})\| > 2$ **do**
    $\triangle i, \triangle j \leftarrow detecteur(x)$;
    $\hat{i} \leftarrow \hat{i} + \triangle i$;
    $\hat{j} \leftarrow \hat{j} + \triangle j$;
    $x \leftarrow transformation(Image128, (\hat{i}, \hat{j}))$;
end

FIGURE A.10: Algorithme du modèle de reconnaissance visuelle

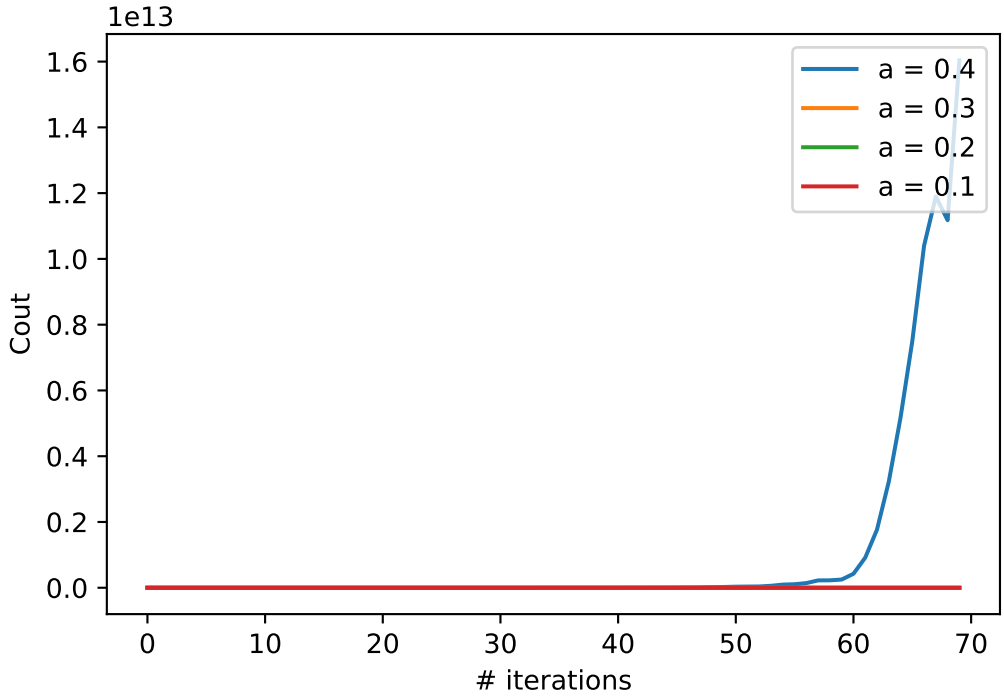

FIGURE A.11: Effet du paramètre alpha sur l'apprentissage dans le cadre d'un filtre *Wavelets*

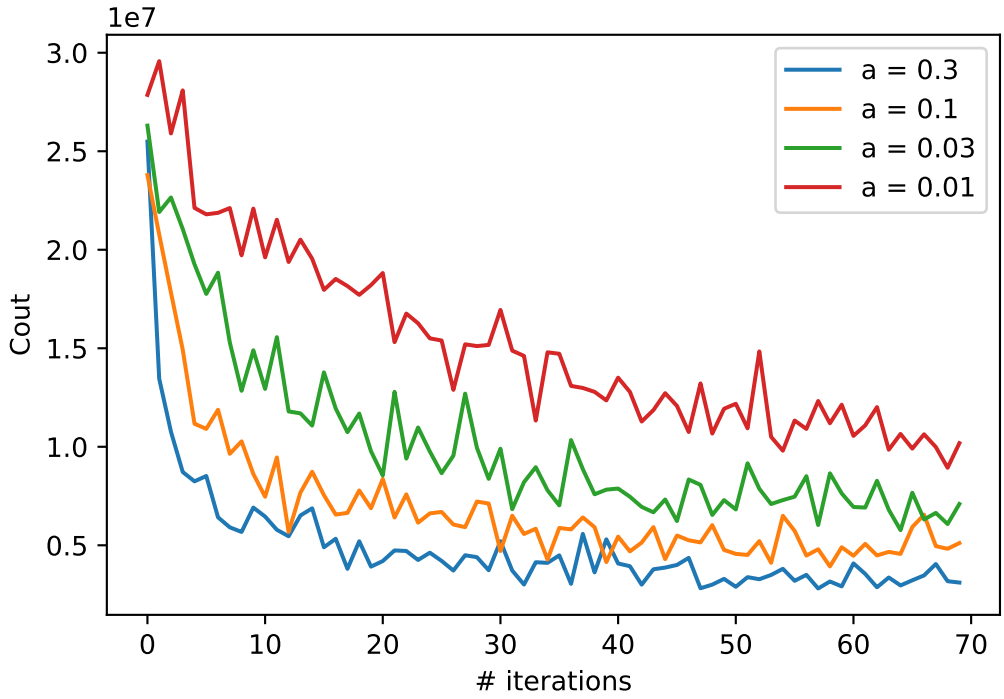

FIGURE A.12: Effet du paramètre alpha sur l'apprentissage dans le cadre d'un filtre *Wavelets*

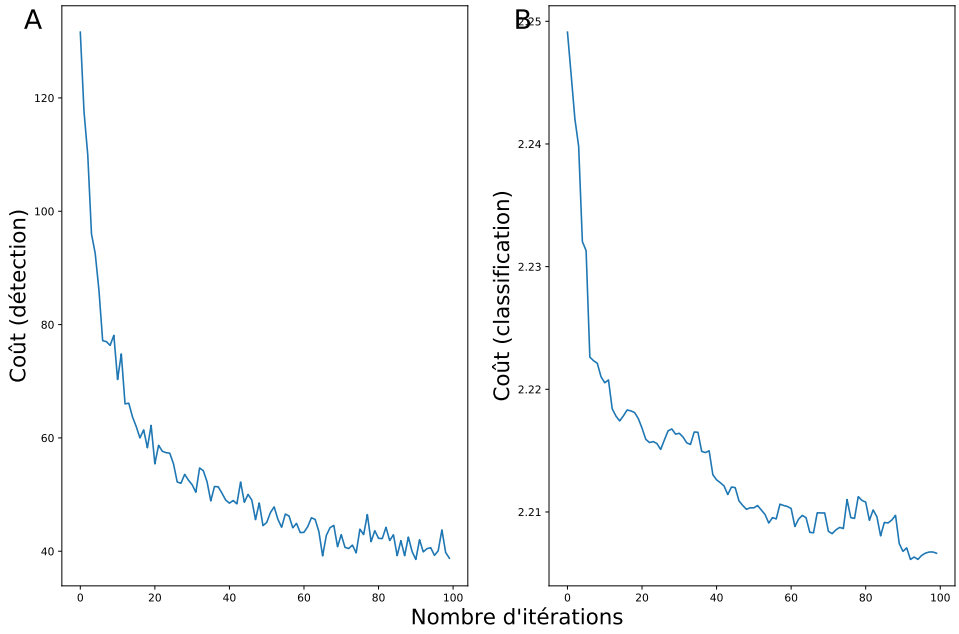

FIGURE A.13: Quantification de la fonction de coût des couches *détecteur* (gauche) et *classifieur* (droite) lors de l'apprentissage, dans le cadre d'un filtre *LogPolar* (taille de l'échantillon d'apprentissage : 10000, nombre d'itérations : 100, $\alpha_{detect} = 0.0015$, $\alpha_{classif} = 0.3$)

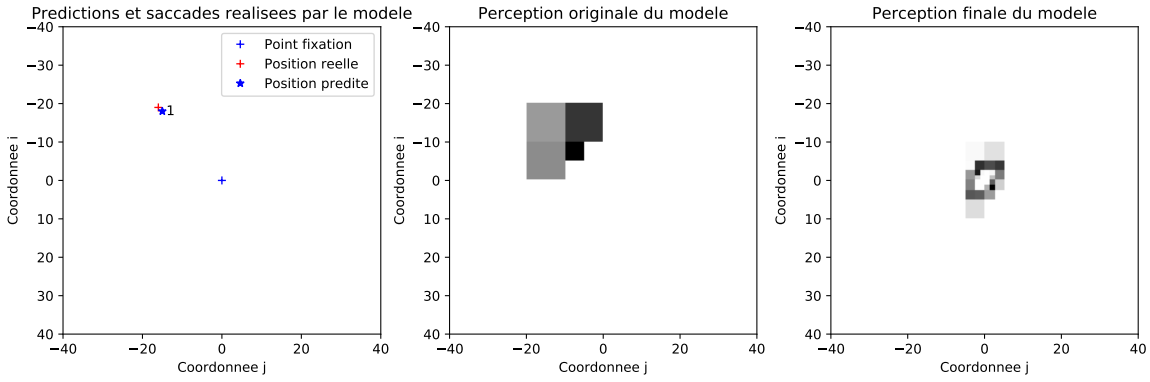

FIGURE A.14: Exemple de perception et comportement saccadique du modèle entraîné, dans le cadre d'un filtre *Wavelets*
A gauche: Position de la fovéa avant et après saccade jusqu'à la position de la cible
Au centre: Perception de l'agent avant saccade
A droite: Perception de l'agent après saccade

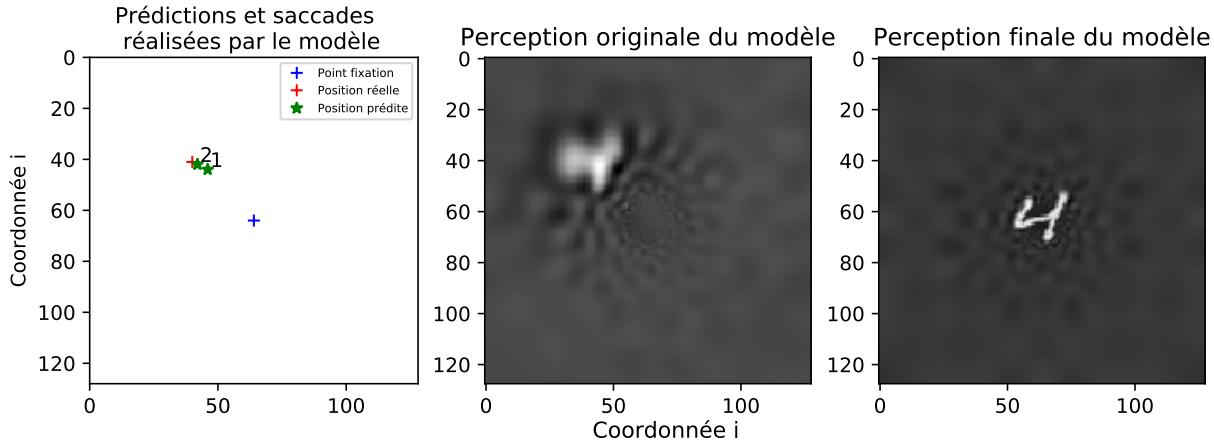

FIGURE A.15: Exemple de perception et comportement saccadique du modèle entraîné, dans le cadre d'un filtre *LogPolar*
A gauche: Position de la fovéa avant et après saccades jusqu'à la position de la cible
Au centre: Perception de l'agent avant saccades
A droite: Perception de l'agent après saccades

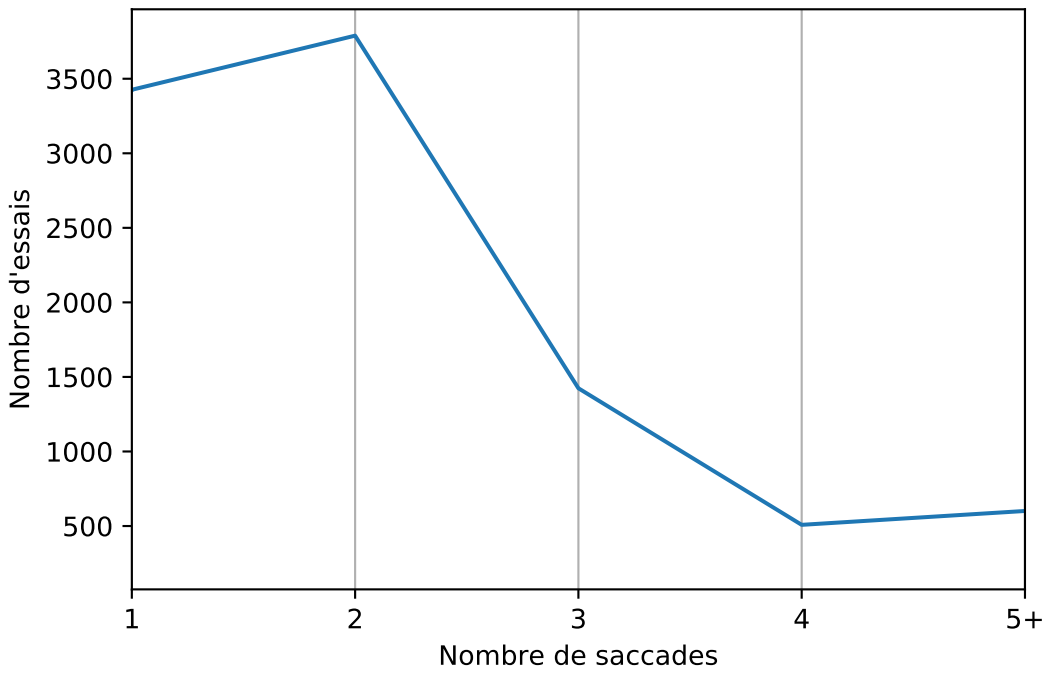

FIGURE A.16: Nombre de saccades nécessaires pour atteindre la position de la cible au cours de 10000 essais, dans le cadre d'un filtre *LogPolar* (taille de l'échantillon d'apprentissage : 10000, nombre d'itérations : 100, $\alpha_{detect} = 0.0015$, $\alpha_{classif} = 0.3$)

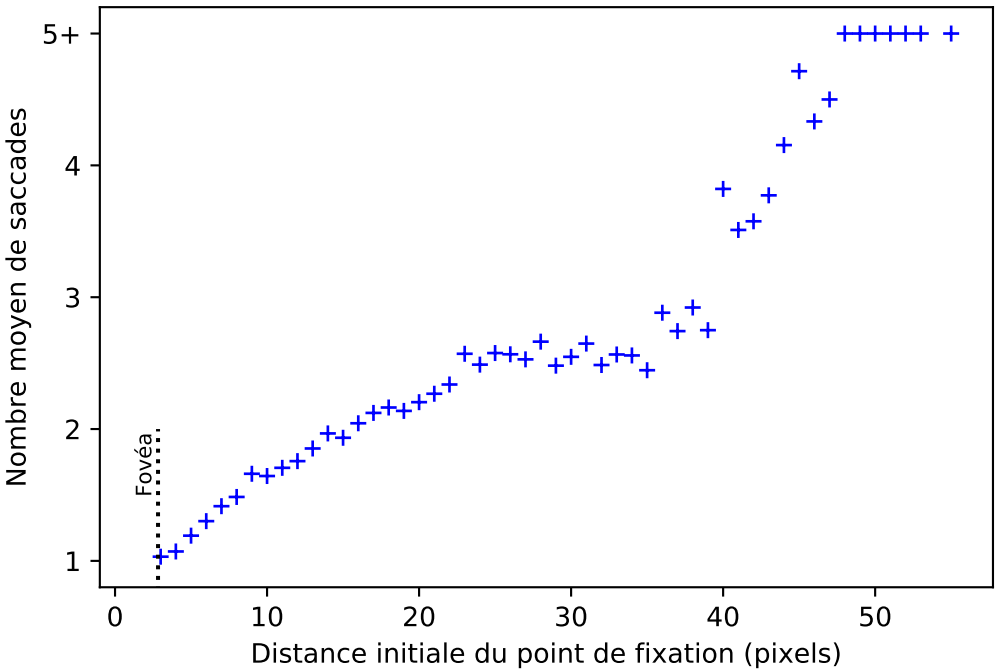

FIGURE A.17: Nombre moyen de saccades nécessaires pour atteindre la position de la cible en fonction de sa distance initiale du point de fixation au cours de 10000 essais, dans le cadre d'un filtre *LogPolar* (taille de l'échantillon d'apprentissage : 10000, nombre d'itérations : 100, $\alpha_{detect} = 0.0015$, $\alpha_{classif} = 0.3$)

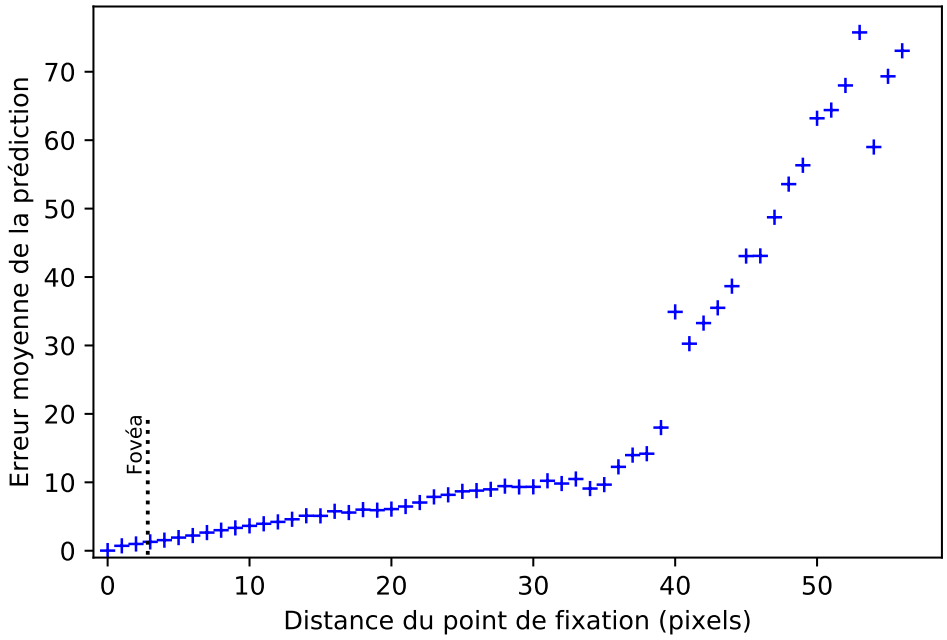

FIGURE A.18: Erreur moyenne lors de la prédiction de la position de la cible en fonction de sa distance du point de fixation au cours de 10000 essais, dans le cadre d'un filtre *LogPolar* (taille de l'échantillon d'apprentissage : 10000, nombre d'itérations : 100, $\alpha_{detect} = 0.0015$, $\alpha_{classif} = 0.3$)

B Code source et documents complémentaires

L'ensemble du code source du modèle sous forme de ipython notebooks, de ce rapport au format LaTeX ainsi que de l'ensemble des autres documents issus de ce travail (dont les notes personnelles) sont entièrement disponibles en ligne ou en contactant directement l'auteur.