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

"Wolves have no Kings"

Robin Hobb

AIX-MARSEILLE UNIVERSITÉ

Abstract

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

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too…

Acknowledgements

The acknowledgments and the people to thank go here, don't forget to include your project advisor…

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. 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.[8]

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.[8]
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 cette dernière de rester fonctionnelle et donc de nous fournir des informations pertinentes dans de nombreux contextes, y compris lorsque la luminance est très faible (le seuil absolu de la rétine humaine correspondant à 70 photons).[8]

Le champs visuel peut être divisé en deux parties. La vision centrale (environ 2° chez l'humain) est anatomatiquement 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 récépteurs les plus petits de la rétine) et une forte perception des couleurs.[8]
La vision périphérique quant à elle comprends majoritairement des bâtonnets, présente une faible acuité et une perception des couleurs très faible (voir nulle). Elle est par contre très sensible à des variations de luminance et de fréquence spatiale (donc aux mouvements).[8]
L'acuité visuelle diminue en fait 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%) .[7,9,8,6]
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.[7,8]

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.[8]
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.[8]

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.[2,8]
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 caractèristiques semblent notamment lier la voie M au traitement de la luminance et des mouvements.[2,8]

Cette multiplicité de voies visuelles est conservée au delà de V1, où l'on décrit deux courant 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).[8,3,5]
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.[8,3,5]

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 plannification des saccades oculaires et à la représentation spatiale des objets visuels.[8]

1.2 Vision artificielle

La vision représentant notre modalité perceptive principale et les aire dévouées 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.[9]

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.[9]

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.[9]

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.2).[9]

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 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 entièrements cachés.[1,6]

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

2 Matériel et méthodes

2.1 Support physique et numérique

L'ensemble des modélisations ont été réalisés sur un ordinateur portable hébergeant une machine virtuelle. Leurs caractéristiques sont rassemblées dans le tableau A.1.
Les modélisation 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.3), 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 et de catégorie du stimulus) en fonction des observations et des actions réalisées jusqu'ici.[1]

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.
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]

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

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]

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

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

Nous construisons ainsi un modèle en deux couches. La première, *détecteur*, est entraînée à prédire la position du stimulus dans l'image perçu 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 d'optimiser la détection du signal.[4]
La seconde, *classifieur*, est entraînée à prédire la catégorie du stimulus dans l'image, perçu au travers du champs rétinien et permet d'arrêter l'exploration de l'image lorsque sa prédiction présente une certitude assez élevée.

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 des niveaux de gris sur 28x28 pixels. Cette image est placée au hasard (mais de façon normocentrée) sur une image au fond blanc de 128x128 pixels (figure A.4).
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, tout en économisant au maximum la puissance de calcul disponible.[7]
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.[7] Les effets de ce filtre sont visibles sur la figure A.5.
La représentation graphique du filtre *LogPolar* utilisé dans ces travaux et ses effets sur l'un de nos stimuli visuels sont visibles sur les figures A.6 et A.7 .
L'une des forces du filtre *LogPolar* est sa capacité à modéliser la distribution des champs récepteurs rétinien, mais aussi (en modifiant seulement quelques paramètres) celle d'aires corticales 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.[3]

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**.
Pour cela, nous calculons une hypothèse h_θ (équation 2.4) sur la répartition des stimuli en multipliant chacune des valeurs x de l'entrée à un poids θ qui lui est spécifique.

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

Ces poids sont ensuite optimisés par descente de gradient (équation 2.5), où ils sont comparés aux labels y des entrées pour un nombre d'exemples et d'itérations fixées. Le paramètre d'apprentissage α influence très fortement l'apprentissage 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.

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

3 Résultats

3.1 Apprentissage supervisé

L'étape de benchmarking du paramètre d'apprentissage α (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*, 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.8 et A.9), tandis que $\alpha = 0.3$ semble représenter une valeur optimale pour l'apprentissage (figure A.10).
Dans le cadre du filtre *LogPolar*, deux jeux de poids indépendants doivent être optimisés pendant l'apprentissage, correspondant respectivement aux couches *détecteur* et *classifieur*. Chaque couche possède ainsi son propre paramètre α (α_{detect} et $\alpha_{classif}$) et l'on peut donc calculer leurs coûts indépendamments (figure A.11).

3.2 Prédiction de la position

Après entraînement, le modèle est capable de détecter la cible dans son environnement et de prédire précisemment sa position (figure A.12 et A.13). Il est ensuite 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.14) et le nombre de saccades nécessaires augmente avec la distance initiale de la cible du centre de fixation (figure A.15). 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.16)

3.3 Prédiction de la catégorie

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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

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

A Figures

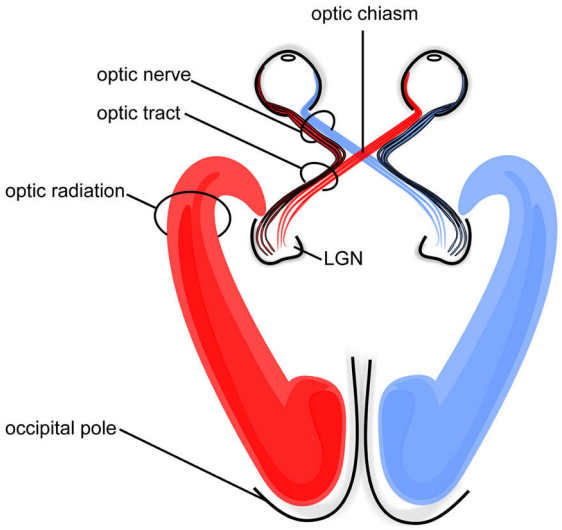

FIGURE A.1: Schéma des voies visuelles précorticales humaines (adapté de Hofer S. et al., 2010 via Wikimedia Commons [CC BY 3.0])

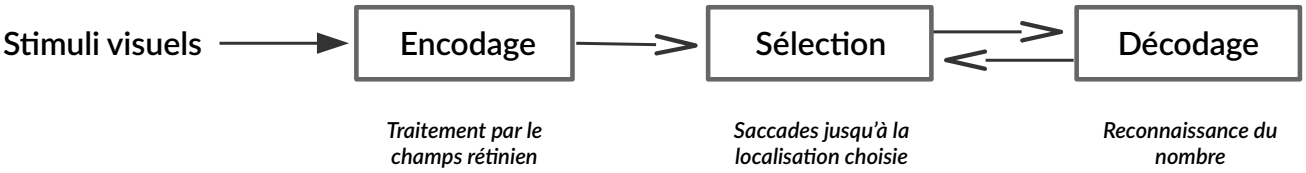

FIGURE A.2: Schéma simplifié du fonctionnement du système visuel avec son *équivalence dans le modèle* (adapté de [9])

	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

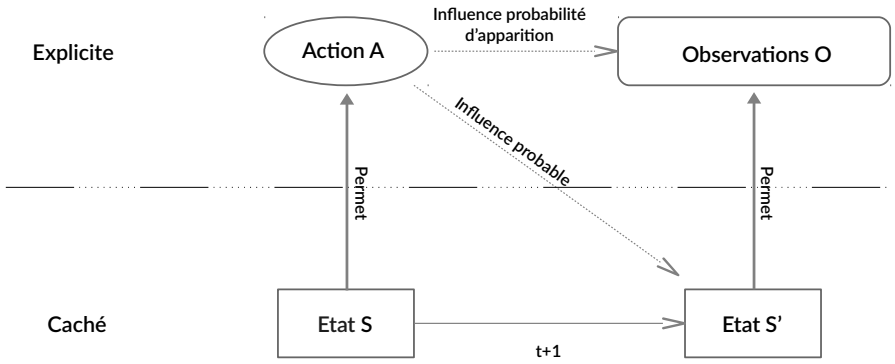

FIGURE A.3: Schéma des interations entre l'agent et son environnement au cours du temps dans un modèle POMDP

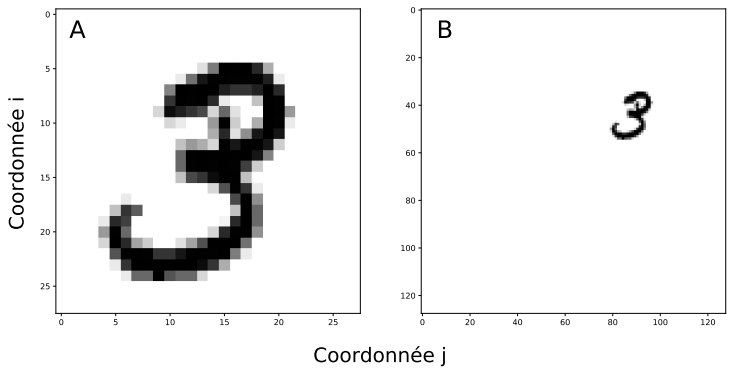

FIGURE A.4: **A.** Image originale tirée de MNIST ; **B.** Image après transformation géométrique et placé aux coordonnées $(i = -20, j = 25)$

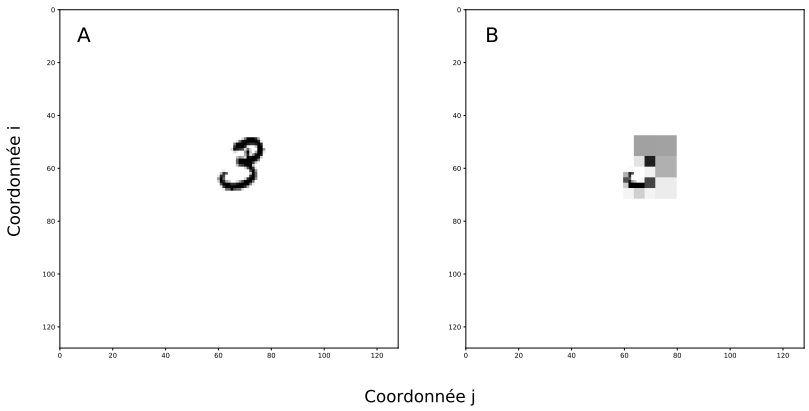

FIGURE A.5: **A.** Image avant application d'un filtre et placé aux coordonnées $(i = -6, j = 6)$; **B.** Image après transformation par le filtre *Wavelets*

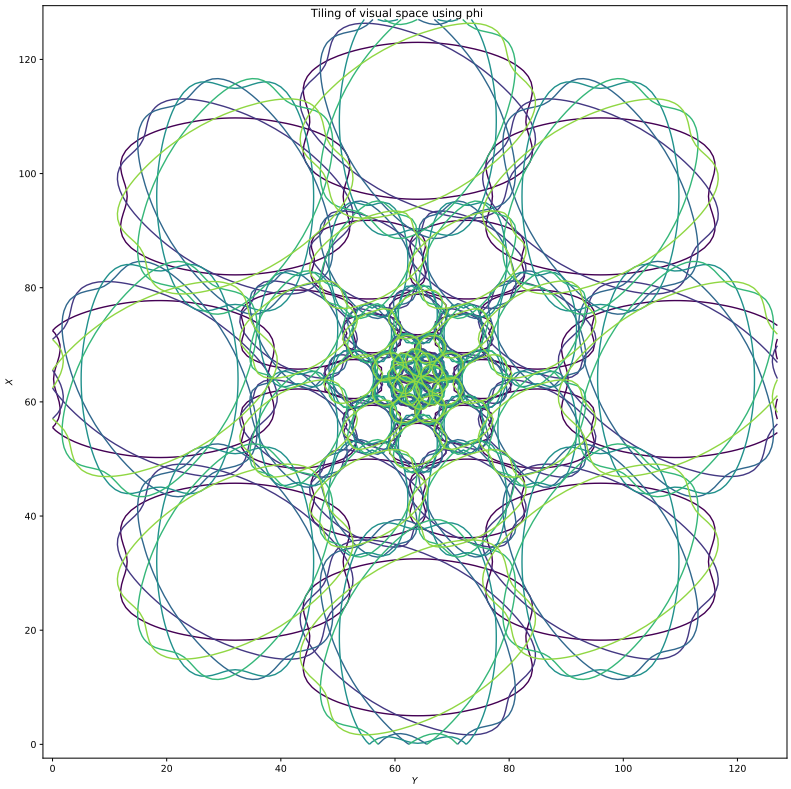

FIGURE A.6: Représentation graphique du filtre *LogPolar* ($N_{theta} = 6$,$N_{orient} = 8$,$N_{scale} = 5$,$N_{phase} = 2$)

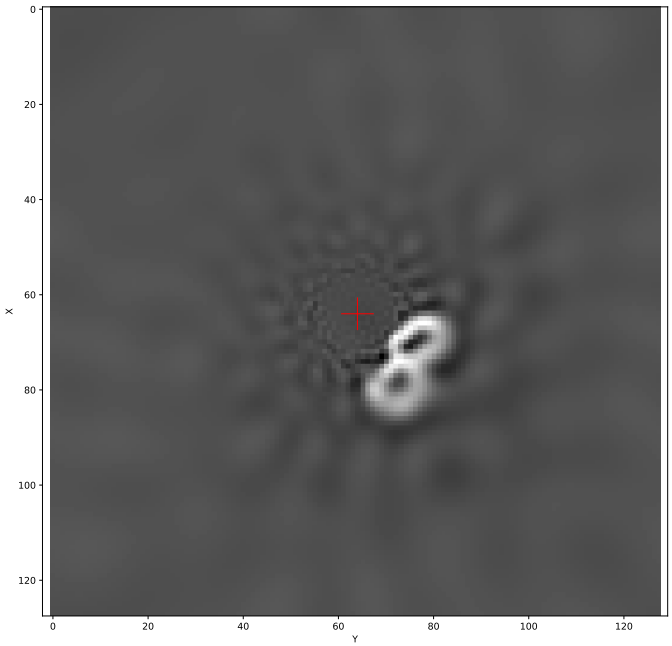

FIGURE A.7: Image après transformation par le filtre *LogPolar* ($N_{theta} = 6$,$N_{orient} = 8$,$N_{scale} = 5$,$N_{phase} = 2$)

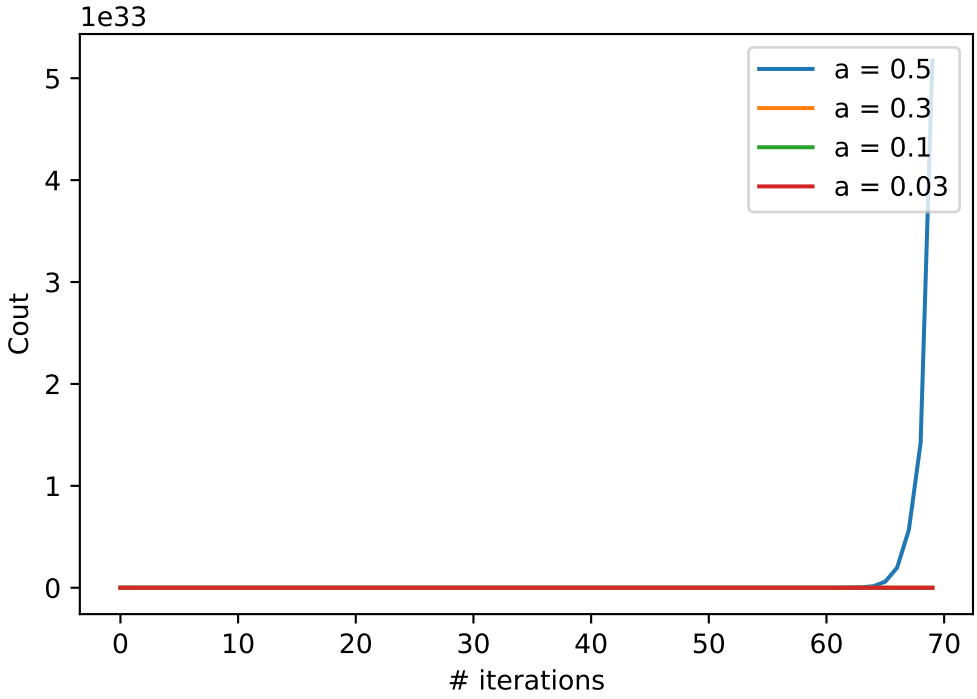

FIGURE A.8: Effet du paramètre alpha sur l'apprentissage dans le cadre d'un filtre *Wavelets*

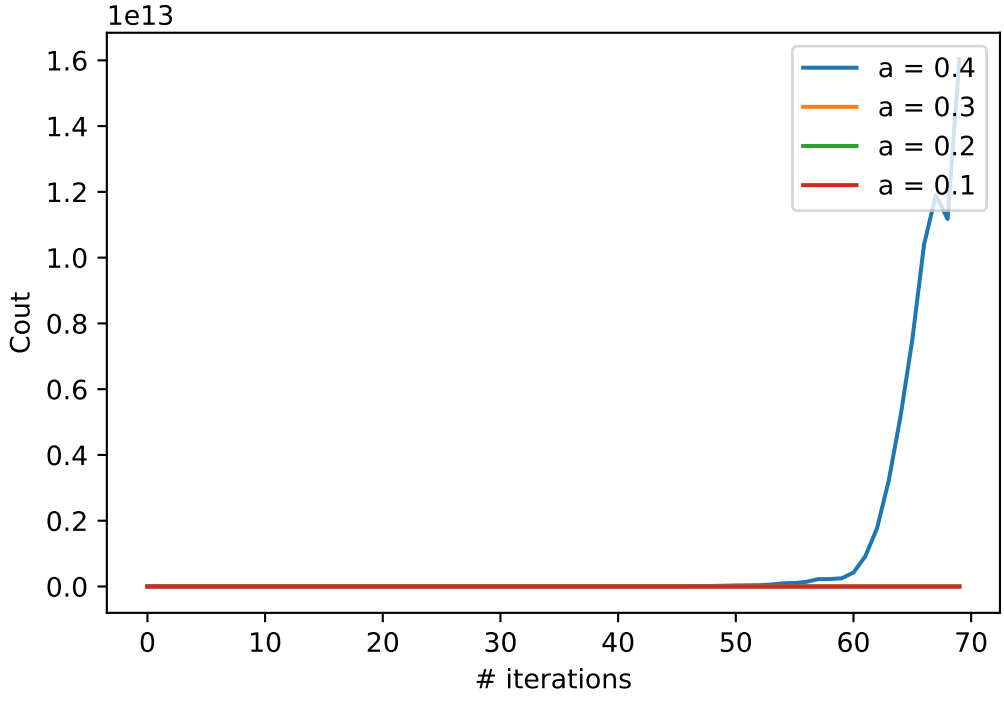

FIGURE A.9: Effet du paramètre alpha sur l'apprentissage dans le cadre d'un filtre *Wavelets*

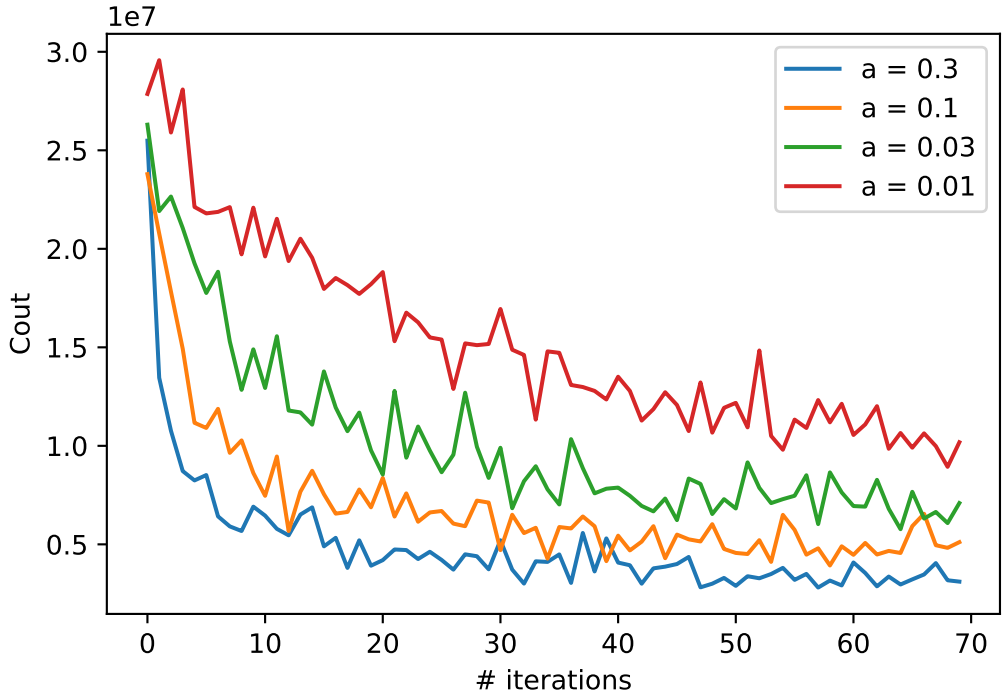

FIGURE A.10: Effet du paramètre alpha sur l'apprentissage dans le cadre d'un filtre *Wavelets*

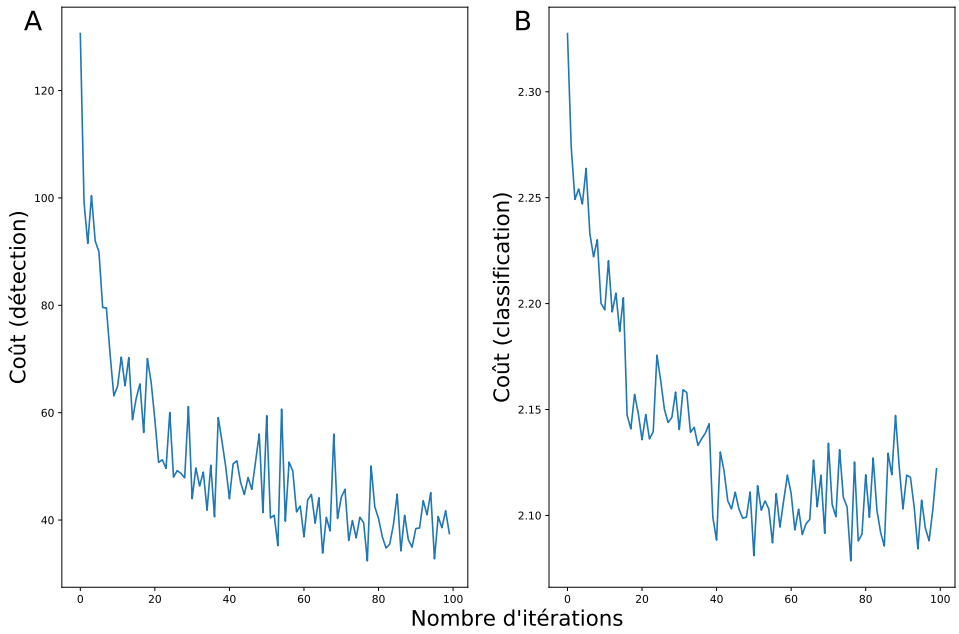

FIGURE A.11: Réduction du coût des couches *détecteur* et *classifieur* lors de l'apprentissage, dans le cadre d'un filtre *LogPolar* (taille de la base d'apprentissage : 1000, nombre d'itérations : 100, $\alpha_{detect} = 0.0015$, $\alpha_{classif} = 0.3$)

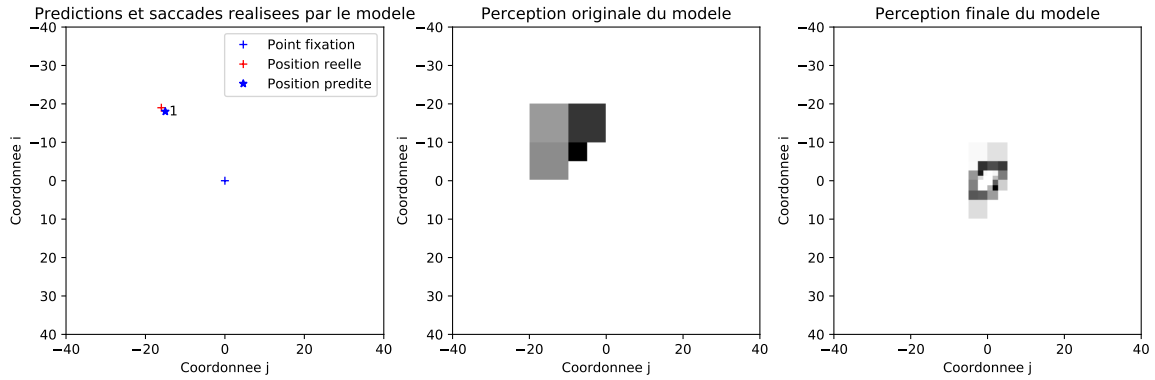

FIGURE A.12: Exemple de perception et comportement saccadique du modèle entraîné, dans le cadre d'un filtre *Wavelets*

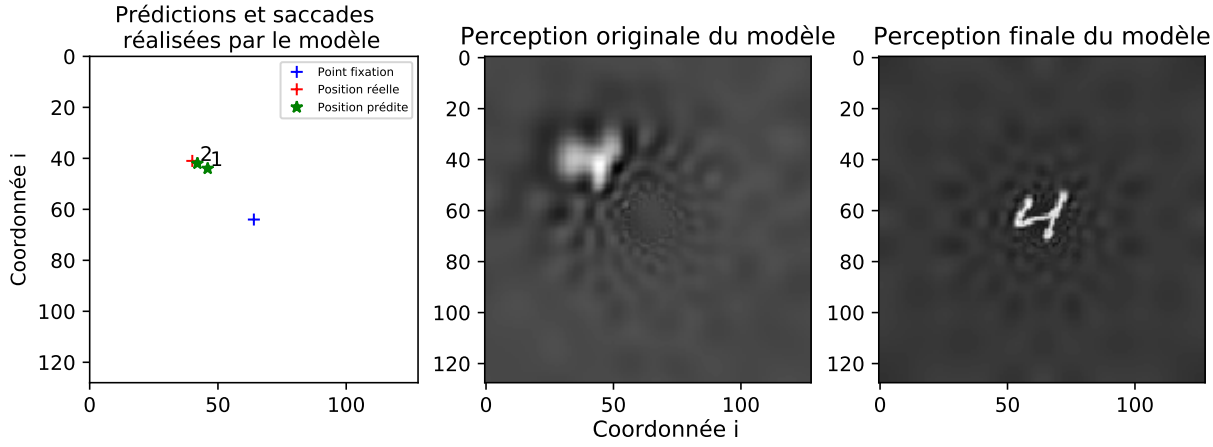

FIGURE A.13: Exemple de perception et comportement saccadique du modèle entraîné, dans le cadre d'un filtre *LogPolar*

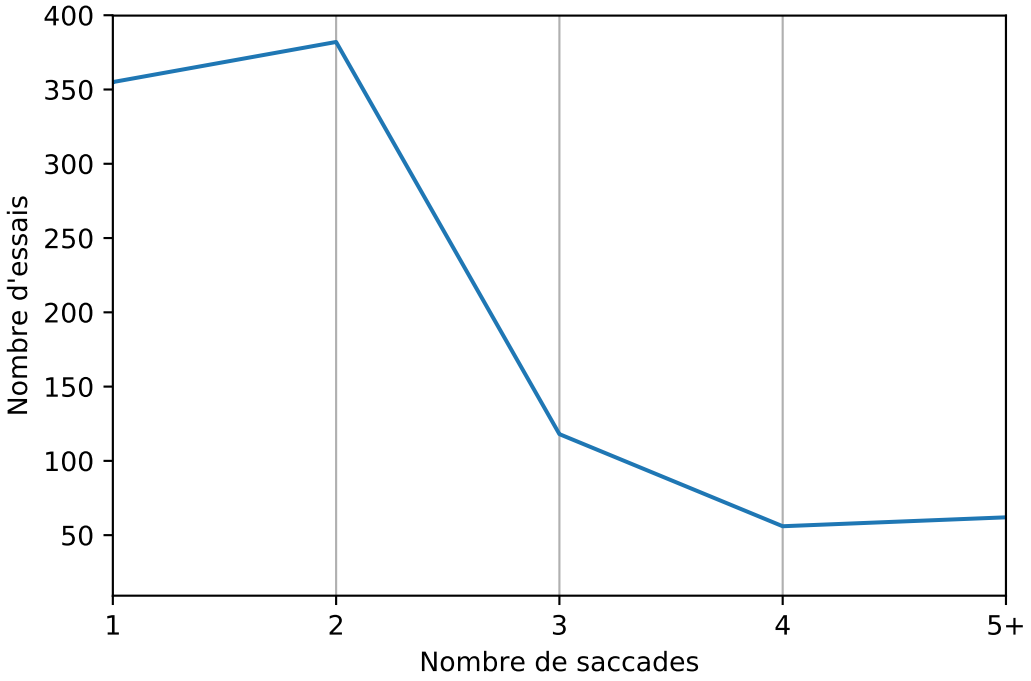

FIGURE A.14: Nombre de saccades nécessaires pour atteindre la position de la cible au cours de 1000 essais, dans le cadre d'un filtre *LogPolar* (taille de la base d'apprentissage : 1000, nombre d'itérations : 100, $\alpha_{detect} = 0.0015$, $\alpha_{classif} = 0.3$)

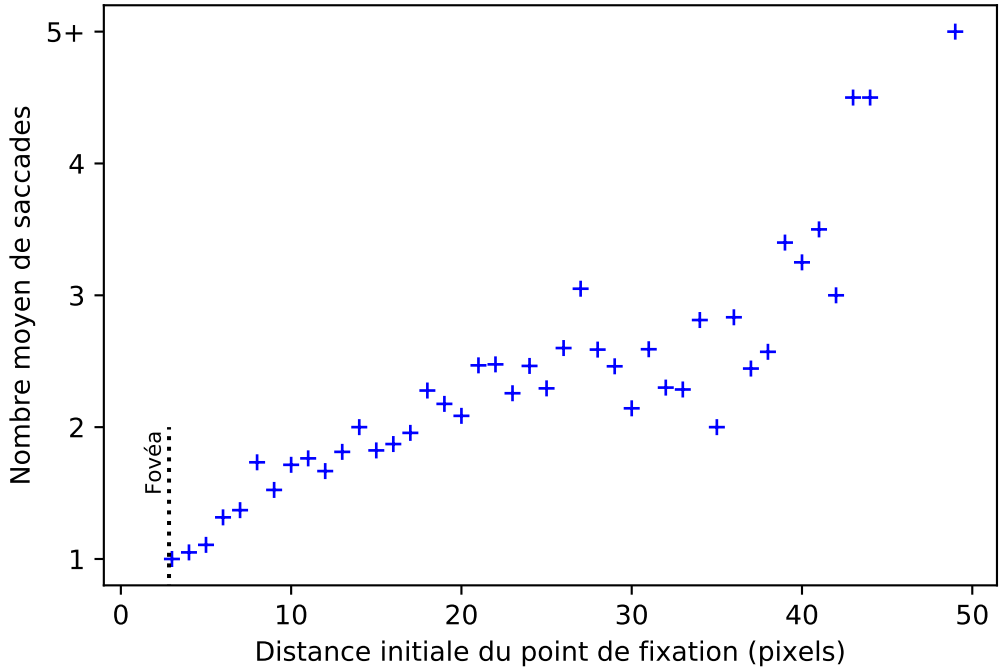

FIGURE A.15: Nombre moyen de saccades nécessaires pour atteindre la position de la cible en fonction de sa distance initiale du point de fixation au cours de 1000 essais, dans le cadre d'un filtre *LogPolar* (taille de la base d'apprentissage : 1000, nombre d'itérations : 100, $\alpha_{detect} = 0.0015$, $\alpha_{classif} = 0.3$)

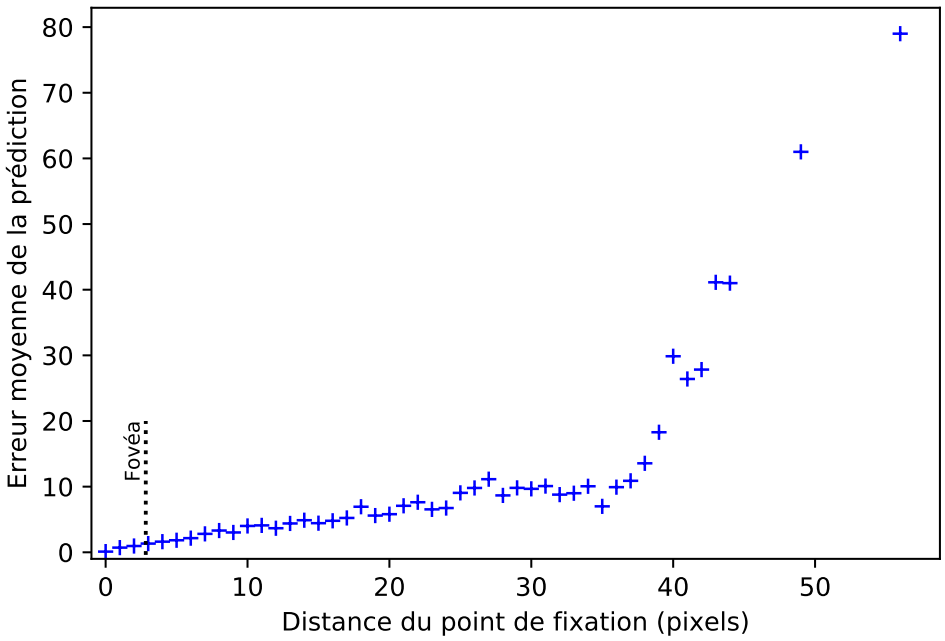

FIGURE A.16: Erreur moyenne lors de la prédiction de la position de la cible en fonction de sa distance du point de fixation au cours de 1000 essais, dans le cadre d'un filtre *LogPolar* (taille de la base d'apprentissage : 1000, nombre d'itérations : 100, $\alpha_{detect} = 0.0015$, $\alpha_{classif} = 0.3$)

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.