Modélisation de processus attentionnels dans la perception multimodale d'un robot virtuel

Sylvain Chevallier, Helene Paugam-Moisy

Modélisation de processus attentionnels dans la perception multimodale d'un robot virtuel

Sylvain CHEVALLIER [a,1] Hélène PAUGAM-MOISY [b,2]

[a] *LIMSI, UPR CNRS 3251, Orsay, Paris-Sud*

[b] *Institut des Sciences Cognitives, UMR CNRS 5015, Lyon*

Résumé

Nous avons modélisé deux types de processus attentionnels sur un robot virtuel, dont le "cerveau" est constitué d'un modèle connexionniste, distribué, de mémoire associative multimodale. Celui-ci évolue dans un environnement virtuel dynamique, avec proies et prédateurs, qui nous permet d'évaluer les apports de la modélisation des mécanismes visuo-attentionnels. Le second type de processus modélisés est issu des récents travaux, en neurosciences, qui démontrent l'existence d'interactions précoces entre les modalités perceptives, les interactions cross-modales.

Mots clefs: processus attentionnels, interactions cross-modales, robotique, vision

1 Introduction

La perception visuelle que nous expérimentons quotidiennement est riche et détaillée, elle démontre la grande virtuosité de nos appareils perceptifs. La modélisation de la vision soulève un double problème décrit par [Marr, 1982] : quels traitements de l'information réalise ce système perceptif et quelles représentations émergent de ces traitements ? La première de ces questions peut réfèrer à un traitement de l'information visuelle qui permet de réduire les quantités d'informations perçues [Tsotsos, 1990] en se focalisant sur certains points plus pertinents : l'attention [Treisman and Souther, 1985].

Nous présentons ici les modélisations de deux types de mécanismes attentionnels sur un robot virtuel [Reynaud and Puzenat, 2001]. Celui-ci utilise un modèle de mémoire associative multimodale, simulant une intégration multisensorielle,

[1] email : sylvain.chevallier@limsi.fr

[2] email : hpaugam@isc.cnrs.fr

[Crépet et al., 2000,Reynaud, 2002] issu d'un modèle d'architecture fonctionnelle de la psychologie cognitive [Kosslyn and Koenig, 1995]. Tout d'abord, nous décrivons ce modèle (partie 2) avant de présenter l'ajout de mécanismes de pré-attention et d'attention visuelle [Wolfe, 2000] [Itti and Koch, 2001] [Fontaine, 2003] (partie 3) et l'amélioration du comportement que cela apporte (partie 4). Nous expliquons ensuite la modélisation des interactions cross-modales (partie 5), qui simulent une influence de l'audition sur la vision [Giard and Perronet, 1999] [Fort et al., 2002] [Falchier et al., 2002].

2 Modèle de mémoire associative multimodale

Les travaux de [Kosslyn and Koenig, 1995] en psychologie cognitive décrivent l'architecture fonctionnelle de la mémoire, vue comme un ensemble de sous-systèmes indépendants et traitant les informations en cascades. Un *buffer visuel* engrange les percepts visuels, la *fenêtre attentionnelle* s'y déplace et délimite la partie du *buffer visuel* à transmettre aux sous-systèmes suivants. La phase de *reconnaissance* est suivie d'un traitement à plus haut niveau : la phase d'*identification*. L'hypothèse d'architectures fonctionnelles similaires pour toutes les modalités perceptives conduit à dupliquer les sous-systèmes bas-niveaux (un par modalité, *reconnaissances* simultanées, indépendantes) et à considérer une unique mémoire, associative et multimodale, qui réalise la fusion des données perceptives et produit une *identification* globale. Cette architecture modulaire, au sens fodorien du terme, est

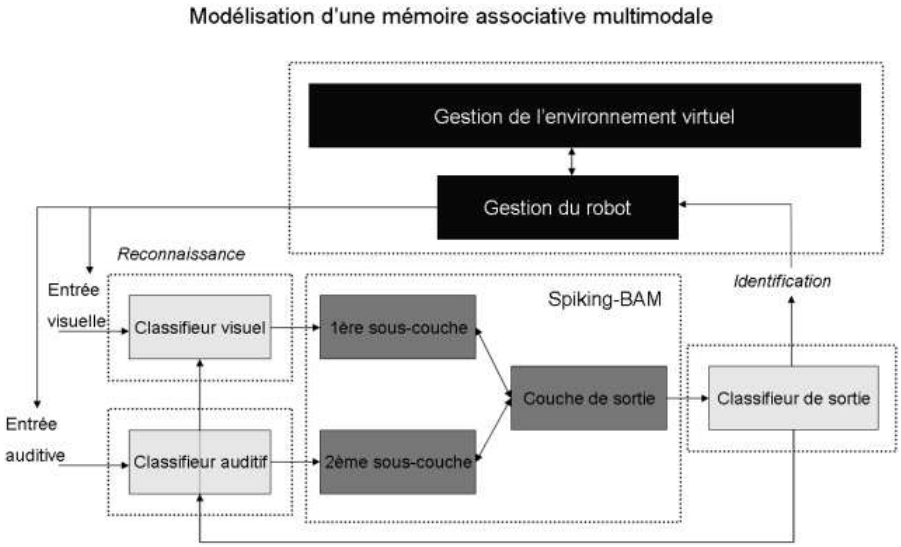

FIG. 1. Architecture du modèle de mémoire associative multimodale. Les cadres en pointillés entourent des processus indépendants qui communiquent par échanges de messages.

exploitée par une modélisation connexionniste distribuée (figure 1), où chaque module ou sous-système est implanté sur un processeur dédié [Bouchut et al., 2003]. Dans cette implémentation, des classifieurs d'entrées, visuel et auditif, procèdent à la *reconnaissance* et une *Bidirectionnal Associative Memory* (BAM), opère la fusion des prétraitements effectués dans chaque modalité perceptive. Un classifieur de sortie permet l'*identification* de la sortie de la BAM [Reynaud, 2002].

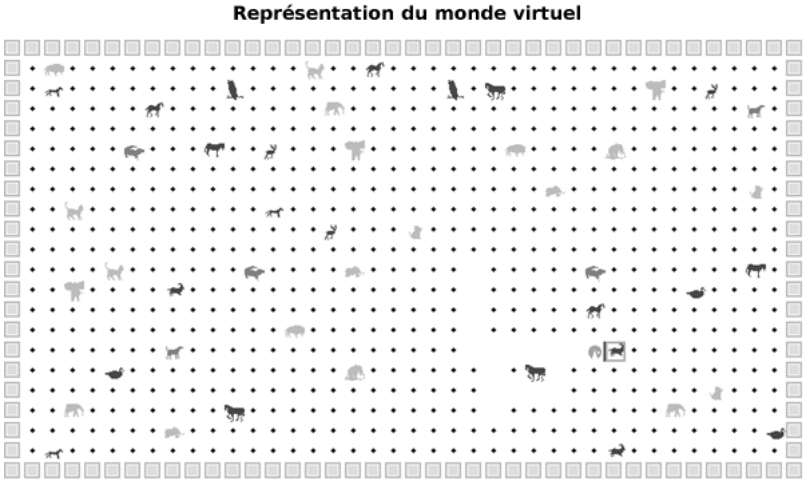

FIG. 2. Illustration de l'environnement virtuel utilisé pour évaluer le comportement du modèle (cf. [Reynaud and Puzenat, 2001]).

Pour être évalué, ce modèle est "incarné" comme constituant le "cerveau" d'un robot virtuel [Reynaud and Puzenat, 2001], doté de champs perceptifs auditifs et visuels distincts. Son champ visuel est un arc de cercle de 120°, dirigé par son regard, et son champ auditif est omnidirectionnel, mais de plus courte portée. Ce robot évolue dans un environnement dynamique simulé (figure 2) où il interagit avec des animaux, des proies qu'il doit capturer et des prédateurs à éviter. Les différences entre ses champs perceptifs, visuel et auditif, engendrent des situations dans lesquelles il entend un animal et en voit un autre. D'autre part, l'éloignement d'un stimulus est simulé par la dégradation de ses perceptions. Ces deux propriétés induisent des cas où les classifieurs échouent à *reconnaître* ou à *identifier* un animal proche du robot.

3 Attention visuelle

La vision pré-attentive regroupe les mécanismes qui peuvent guider l'attention visuelle [Treisman and Souther, 1985] : le "pop-out" mis en évidence, par exemple, quand il faut trouver une cible verte parmi des distracteurs rouges. Les traits caractéristiques sur lesquels reposent ces traitements pré-attentifs chez l'humain sont : la couleur, le mouvement, la luminance, l'orientation, *etc* [Wolfe, 2000]. Pour la vision artificielle, il faut nécessairement recourir à l'utilisation de processus attention-

nels. Sans ce type de stratégie d'exploration, la puissance computationnelle requise explose dès que les images sont grandes ou détaillées [Tsotsos, 1990]. Ces processus attentionnels requièrent les trois composants suivants [Tsotsos et al., 1995] : la sélection d'une région d'intérêt, la sélection de dimensions caractéristiques et de valeurs d'intérêt (comme celles citées ci-avant) et le déplacement dans le temps d'une région sélectionnée vers la suivante. Nous nous baserons sur un modèle d'attention visuelle [Itti and Koch, 2000,Itti and Koch, 2001] qui utilise les composants décrits par Tsotsos et qui permet de sélectionner rapidement, dans une image, des points saillants qui déterminent la sélection d'une région d'intérêt.

Pour modéliser ces mécanismes pré-attentifs, nous avons considéré les points suivants : (i) il est possible de discerner, dès le stade pré-attentif, la nature de l'animal (proie ou prédateur), (ii) ce discernement peut être réalisé en faveur des prédateurs, *i.e.* rendre les prédateurs plus saillants que les proies, (iii) cette distinction ne peut être effectuée que pour des percepts situés suffisamment près du robot. Le nombre important de dimensions caractéristiques existant au stade pré-attentif, détaillées dans [Wolfe, 2000], offre un choix permettant de discerner la nature de l'animal. Il est possible d'identifier quelles dimensions caractéristiques sont les plus pertinentes pour discriminer au mieux la nature des animaux. Une fois ces dimensions choisies, en déterminant quelle importance on accorde à chacune d'elles, il est possible de rendre les prédateurs plus *saillants* que les proies (voir figure 3). Cette possibilité est démontrée par des modélisations telles que celles de [Itti and Koch, 2001].

4 Tests et résultats

La modélisation des mécanismes visuo-attentionnels retenus permet de désambiguïser certaines situations, comme le montre la figure 3. De manière intuitive, on peut noter que le robot ne s'approchera pas d'une proie si celle-ci est trop proche d'un prédateur et qu'il tiendra moins compte des animaux neutres quand ceux-ci seront entourés de proies ou de prédateurs. Pour évaluer les changements de comportement du robot suite à l'ajout de processus attentionnels, nous avons réalisé plusieurs tests. Nous mesurons le nombre de "vies" perdues par le robot (quand il est touché par un prédateur), le nombre de proies "mangées", son nombre de déplacements, total et vers l'avant, et l'étendue d'exploration de son environnement.

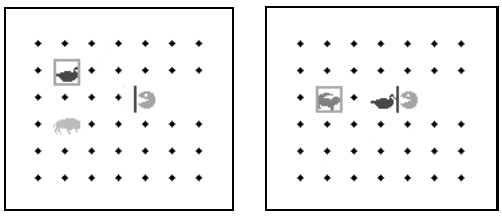

FIG. 3. Situations illustrant les cas où l'ajout de mécanismes visuo-attentionnels change la réponse comportementale. La direction du regard est symbolisée par la barre placée à côté du robot.

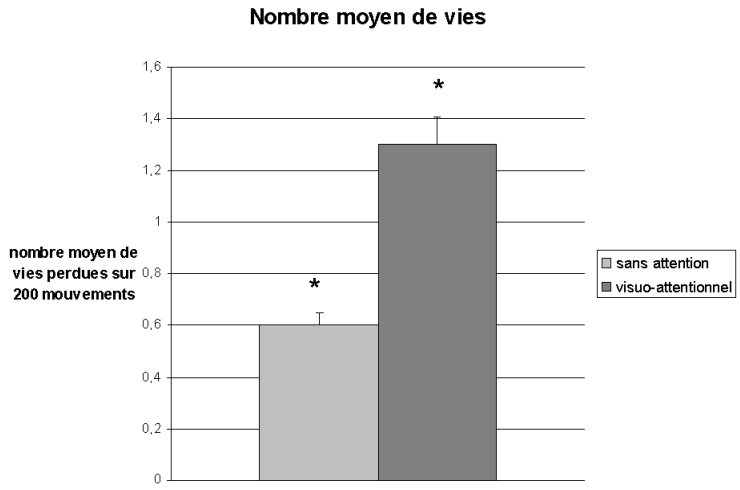

FIG. 4. Le diagramme donne le nombre de vies moyen pour 10 exécutions où le nombre de mouvements total est limité à 200.

Une première évaluation a été réalisée en lançant dix fois le programme avec, comme critère d'arrêt, une limite de 200 mouvements, selon deux conditions, avec et sans processus visuo-attentionnels. Les résultats sont présentés dans la figure 4 et montrent un gain significatif (Test de Student, $p < 0,05$) en terme de nombre de vies et de zone explorée pour le robot disposant de mécanismes visuo-attentionnels. Le second test, où le nombre de vies est fixé à trois, démontre que le nombre de mouvements (total et vers l'avant) ainsi que les zones explorées sont significativement plus importants ($p < 0,05$) pour le robot doté de processus attentionnels (figure 5). Ces deux diagrammes traduisent l'apport significatif, au modèle de mémoire associative multimodale, de la modélisation de processus attentionnels. Le robot virtuel "attentif" est plus apte à éviter les prédateurs et il explore mieux son environnement.

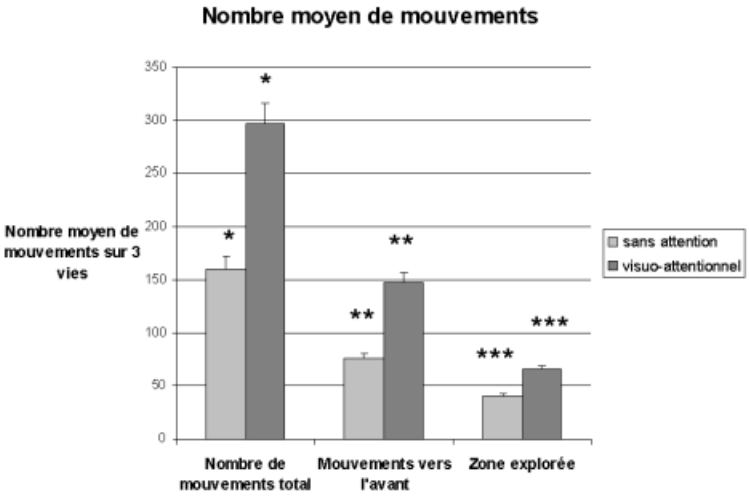

FIG. 5. Diagramme exposant plusieurs mesures moyennes du mouvement, calculées sur 10 exécutions, avec un nombre de vies limité à 3.

5 Interactions cross-modales

Les travaux récents de neurophysiologie et de neuroanatomie ont mis en évidence les interactions cross-modales qui se produisent entre les systèmes perceptifs, aussi bien chez les humains [Fort et al., 2002] que chez les animaux [Falchier et al., 2002] [Giard and Perronet, 1999]. Dès les premiers traitements perceptifs (40 ms après la perception), les modalités interagissent, à bas niveau, et facilitent ainsi la *reconnaissance*. Le cerveau est capable de rediriger son attention visuelle vers une source sonore périphérique. Le modèle de mémoire associative multimodale, dans sa version la plus récente [Meunier and Paugam-Moisy, 2004], possède une *BAM* implémentée avec des neurones impulsionnels (*spiking neurons*). Ce formalisme de neurone artificiel, biologiquement plus plausible, nous permet de prendre en compte ce type d'interactions. Cette *Spiking-BAM* est capable de réaliser la fusion des différentes modalités de façon dynamique, en intégrant les percepts *reconnus* au fur et à mesure de leur arrivée.

Nous proposons une modélisation de l'interaction cross-modale de l'audition sur la vision, qui repose sur les deux hypothèses suivantes : (i) la perception d'un stimulus auditif saillant à la périphérie du champ visuel entraîne une nouvelle focalisation visuelle, (ii) un stimulus auditif est considéré comme saillant lorsque l'animal qui l'émet est suffisamment proche. Ainsi, lorsque le robot voit un animal V et entend un animal A très proche (figure 6), il déclenche les procédures de *reconnaissance* dans les deux modalités. Directement après la détection de ce stimulus auditif proche, le robot redirige son regard vers l'animal A et déclenche une nouvelle procédure de *reconnaissance* visuelle. Celle-ci sera intégrée par la *BAM* au cours du traitement d'identification.

FIG. 6. Illustration d'une situation où le mécanisme d'interaction cross-modale intervient. Un animal proche (crocodile) est hors du champ visuel. Le robot l'entend et redéclenche une procédure de vision.

La figure 7 présente une visualisation des patterns d'activation de cette intégration en cours de traitement dans la *Spiking-BAM*. Trois couches de neurones sont dédiées, respectivement, à l'intégration des stimuli visuels, des stimuli auditifs et à la fusion des modalités. Au temps d'intégration R1, le robot voit et entend l'animal V, à R2 il voit l'animal V et entend un animal proche A. Il redirige son regard et, dès l'intégration suivante, il voit (à l'instant R3) et entend (à R4) l'animal A. On

observe que la couche interne de la *Spiking-BAM*, qui fusionne les données, est modifiée entre R2 et R4, ce qui signifie que le robot intègre bien ce nouveau percept visuel.

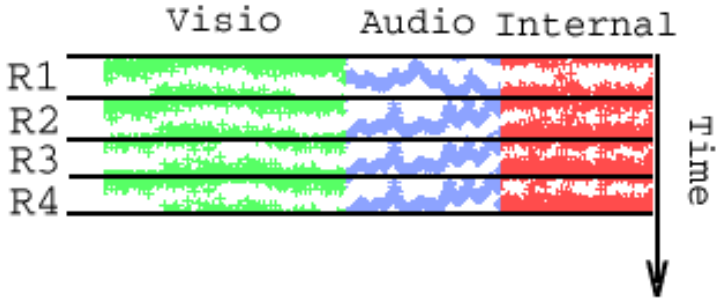

FIG. 7. Diagramme, en fonction du temps, des patterns d'activations des 3 couches neuronales de la *Spiking-BAM* du robot virtuel.

6 Conclusion

En nous appuyant sur un modèle connexionniste de mémoire associative multimodale [Reynaud, 2002] [Bouchut et al., 2003] [Meunier and Paugam-Moisy, 2004], nous avons proposé une modélisation de processus attentionnels, à partir des données de la psychologie cognitive [Treisman and Souther, 1985,Wolfe, 2000] et des travaux en vision artificielle [Tsotsos et al., 1995] [Itti and Koch, 2000] ou encore [Itti and Koch, 2001]. Cette modélisation permet une amélioration comportementale du robot virtuel qui adapte mieux son comportement à l'environnement dans lequel il évolue. D'autre part, nous avons simulé, dans ce modèle, des interactions cross-modales qui ont été mises en évidence par de récents travaux en neurosciences [Giard and Perronet, 1999] [Falchier et al., 2002] ou bien [Fort et al., 2002]. Ces améliorations s'inscrivent dans la perspective d'utiliser ce type de modèle informatique distribué en robotique autonome. Les processus attentionnels jouent un rôle important sur le gain en temps de traitements pour tendre vers des modèles fonctionnant en temps réel.

Références

[Bouchut et al., 2003] Bouchut, Y., Paugam-Moisy, H., and Puzenat, D. (2003). Asynchrony in a distributed modular neural network for multimodal integration. In *Int. Conf. on Parallel and Distributed Computing and Systems*, pages 588–593, Marina del Rey, US. IASTED, ACTA Press.

[Crépet et al., 2000] Crépet, A., Paugam-Moisy, H., Reynaud, E., and Puzenat, D. (2000). A modular neural network for binding several modalities. In Ed., H. A., editor, *Int. Conf. of Artificial Intelligence*, pages 921–928, Las Vegas. IC-AI, CSREA Press.

[Falchier et al., 2002] Falchier, A., Clavagnier, S., Barone, P., and Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *The Journal of Neuroscience*, 22(13) :5749–5759.

[Fontaine, 2003] Fontaine, M. (2003). *Mise en œuvre d'un système de vision attentionnelle fondé sur l'identification de points d'intérêt*. Rapport de DEA - Université Paris-Sud/LIMSI.

[Fort et al., 2002] Fort, A., Delpuech, C., Pernier, J., and Giard, M.-H. (2002). Early auditory-visual interactions in human cortex during nonredundant target identification. *Cognitive Brain Research*, 14 :20–30.

[Giard and Perronet, 1999] Giard, M.-H. and Perronet, F. (1999). Auditory-visual integration during multimodal object recognition in humans : a behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, 11(5) :473–490.

[Itti and Koch, 2000] Itti, L. and Koch, C. (2000). A saliency based search mecanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12) :1489–1506.

[Itti and Koch, 2001] Itti, L. and Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3) :194–203.

[Kosslyn and Koenig, 1995] Kosslyn, S. and Koenig, O. (1995). *Wet Mind : The New Cognitive Neuroscience*. New York : Free Press, 2ème edition.

[Marr, 1982] Marr, D. (1982). *Vision : a computational investigation into the human representation and processing of visual information*. W.H. Freeman and Compagny, New-York.

[Meunier and Paugam-Moisy, 2004] Meunier, D. and Paugam-Moisy, H. (2004). A "spiking" bidirectional associative memory for modeling intermodal priming. In *Int. Conf. on Neural Networks and Computational Intelligence*, pages 25–30, Grindelwald, SUISSE. IASTED, ACTA Press.

[Reynaud, 2002] Reynaud, E. (2002). *Modélisation connexionniste d'une mémoire associative multimodale*. PhD thesis, Institut National Polytechnique de Grenoble.

[Reynaud and Puzenat, 2001] Reynaud, E. and Puzenat, D. (2001). A multisensory identification system for robotics. In IJCNN, editor, *Int. Joint Conf. on Neural Networks*, pages 2924–2929, Washington DC. IJCNN'2001, INNS.

[Treisman and Souther, 1985] Treisman, A. and Souther, J. (1985). Search asymmetry : a diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology - General*, 114 :285–310.

[Tsotsos, 1990] Tsotsos, J. K. (1990). Analysing vision at the complexity level. *Behavioral and brain sciences*, (13) :423–469.

[Tsotsos et al., 1995] Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., and Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial intelligence*, 78 :507–545.

[Wolfe, 2000] Wolfe, J. (2000). *Seeing*, chapter Visual attention, pages 335–386. Academic Press, 2ème edition.