Scrutation de scènes naturelles par un agent conversationnel animé

Antoine Picot, Gérard Bailly, Frédéric Elisei & Stephan Raidt

Institut de la Communication Parlée, 46 av. Félix Viallet, 38031 Grenoble - France
Correspondance: gerard.bailly@icp.inpg.fr

RÉSUMÉ. *Cet article présente un système de gestion du regard d'un agent conversationnel animé doté de capacités à percevoir l'environnement physique dans lequel il intervient. Ce système est inspiré des composantes connues de l'attention visuelle et intègre ses limitations en termes d'acuité, de sensibilité au mouvement, de mémoire à court-terme et de suivi de mouvements. Le but de ce couplage entre animation et analyse de scènes est de fournir à l'usager des gages de présence et d'attention aux changements de l'environnement immédiat d'interaction. Après une courte introduction à ce projet de recherches et un bref état de l'art, cet article détaille les composantes de notre système et présente une première confrontation à des données oculométriques enregistrées sur des sujets observant les mêmes scènes.*

ABSTRACT. *We present here a system for controlling the eye gaze of a virtual embodied conversational agent able to perceive the physical environment in which it interacts. This system is inspired by known components of human visual attention system and reproduces its limitations in terms of visual acuity, sensitivity to movement, limitations of short-memory and object pursuit. The aim of this coupling between animation and visual scene analysis is to provide sense of presence and mutual attention to human interlocutors. After a brief introduction to this research project and a focused state of the art, we detail the components of our system and confront simulation results to eye gaze data collected from viewers observing the same natural scenes.*

MOTS-CLÉS : *agents conversationnels animés, interaction face-à-face, regard, tête parlante, analyse de scène visuelle.*

KEYWORDS: *embodied conversational agents, face-to-face interaction, eye gaze, talking face, visual scene analysis.*

1. Introduction

Notre regard se porte sur les éléments de notre environnement suivant des trajectoires complexes qui dépendent de multiples facteurs : la proéminence perceptive de ces éléments par rapport à leur environnement (couleur, forme, mouvement, etc.), la pertinence de ceux-ci par rapport à l'objectif de notre scrutation (chercher un visage, un objet, etc.), l'attention que nous leur portons (poursuite d'éléments en particulier, etc.) ainsi que la connaissance *a priori* que nous avons sur chaque élément de cette scène (familiarité de l'objet ou du visage, etc.). L'objectif de ce travail est de déterminer automatiquement la succession des divers centres d'intérêt probables d'une scène animée en gérant de manière optimale l'ensemble de ces contraintes et d'y asservir les saccades oculaires d'un agent conversationnel animé (ACA).

Ce travail s'inscrit dans un projet plus ambitieux visant à doter les ACA de capacités à percevoir l'environnement dans lequel ils sont plongés afin de les rendre sensibles aux changements de cet environnement, induits ou non par leurs propres actions. L'environnement inclut évidemment le ou les interlocuteurs : l'objectif de cet ancrage sur le monde réel, de cette attention portée à l'environnement de l'interaction, à l'interlocuteur, à ce qu'il regarde, désigne ou dit, est de donner aux partenaires humains des gages de présence et de mesurer l'impact de ces gages de présence sur le dialogue (compréhension de l'information linguistique et paralinguistique échangée, charge cognitive, croyance en l'information, etc.).

Après un rapide état de l'art où nous allons particulièrement détailler deux travaux majeurs qui ont inspiré ce travail, nous décrivons et illustrons par des exemples concrets les principales innovations de notre proposition. A cette description technique succède une première confrontation de notre système à des données oculométriques.

(a) **(b)**

Figure 1 : Gestion du regard par ROBITA signalant aux interlocuteurs (a) qu'il suit effectivement les tours de parole ; (b) qu'il décode effectivement leurs gestes déictiques.

2. Etat de l'art

Afin de pouvoir planifier leurs déplacements dans un environnement changeant, les robots mobiles sont dotés de capacité à analyser l'espace qui les entoure. La plupart des robots anthropoïdes ou robots de compagnie sont de même dotés de systèmes sophistiqués d'analyse de scène pour analyser le comportement de leurs interlocuteurs, de façon à planifier leurs actions de manière adéquate. Ainsi les robots sociaux du MIT (Brooks, Breazeal et al. 1999) intègrent des boucles de perception/action où la gestion de l'attention mutuelle est essentielle pour acquérir et maintenir un espace de représentation partagée. Ces boucles exploitent les résultats du suivi de la direction du regard de l'usager et la capacité à mouvoir tête et globes oculaires pour signaler cette capacité de perception et d'analyse. De même ROBITA, développé à l'Université WASEDA au Japon (cf. Figure 1), donne des gages d'attention visuelle et auditive. Il est capable de suivre et d'intervenir dans une conversation multi-intervenants (Matsusaka, Tojo et al. 2003). Il est également capable de comprendre certains gestes tels que la désignation multimodale d'objets.

En l'absence de capacité à ancrer le contrôle de l'action sur la perception de l'environnement, de nombreux ACA exploitent des modèles de mouvement du regard basés sur certaines régularités statistiques (telles que la fréquence des clignements ou l'amplitude des saccades oculaires). Lee et al. (2002) ont ainsi développé un modèle basé sur des modèles empiriques de saccades oculaires (distinguant notamment les états production/perception de parole) et des modèles statistiques de données oculaires. Si les saccades calculées par ce système sont nettement préférées à un regard fixe ou aléatoire, les auteurs reconnaissent qu'un séquencement plus fin des tâches cognitives supposées de l'ACA est souhaitable (phases préparatoires à la prise de parole, etc.) ainsi qu'un ancrage sur l'environnement. Une solution à ce dernier point a été proposée par Courty (2002) mais seulement pour les scènes virtuelles dans lesquelles leur ACA était plongé.

Le robot Rackham, développé par le LAAS, intègre l'un de nos ACA. Un premier couplage de regard avec l'analyse de l'environnement qu'il effectue pour se déplacer et guider l'usager a été effectué (Clodic, Fleury et al. 2006). L'objectif de ce travail est de le doter d'un système de contrôle du regard plus complet, capable de reproduire les caractéristiques essentielles de l'attention visuelle humaine.

2.1. *Données comportementales*

Le processus d'échantillonnage par lequel notre œil explore notre champ de vision peut être vu comme une série de saccades, de fixations et de poursuites. Les saccades sont des mouvements rapides de l'œil (approx. 25-40ms, 200°/s, 150000 par jour) qui amènent la région d'intérêt de notre champ de vision dans le champ réceptif central de l'œil (fovea) à des fins d'analyse haute-résolution. Les fixations permettant alors cette analyse (approx. 300ms) sont caractérisées par des

microsaccades qui permettent notamment de compenser l'adaptation rétinienne. Ces deux composantes correspondent grossièrement à deux voies visuelles complémentaires (Grossberg 2003) : une voie ventrale vers le lobe pariétal « quoi » allocentrique qui est responsable de l'identification des objets et la voie dorsale vers le lobe temporal « où » égocentrique qui participe à la localisation des évènements multi-sensoriels de la scène. Un mécanisme additionnel – la poursuite lente – permet d'accrocher la fovea sur un objet d'intérêt se déplaçant lentement.

La scrutation d'une scène (une image fixe ou une vidéo) ne se résume pas à effectuer des saccades de l'objet le plus saillant de la scène au suivant. La saillance perceptive n'est pas le seul déterminant de l'intérêt : la demande cognitive a un impact fort sur la stratégie de scrutation et sur l'interprétation de la scène. Yarbus (1967) a ainsi montré que la trajectoire du regard est influencée par la tâche donnée au sujet lors de l'examen d'une image. De même, Vatikiotis-Bateson et al (1998) ont montré que la trajectoire du regard lors de perception de parole audiovisuelle est influencée non seulement par les conditions environnementales (rapport signal/bruit) mais aussi par la tâche de compréhension (segmentale *vs* suprasegmentale). La mobilisation de l'attention sur certains éléments de la scène demandée par la tâche peut alors conduire à une cécité au changement d'autres éléments tout aussi saillants (Simons and Chabris 1999).

2.2. *Modèles de scrutation de scènes visuelles*

De nombreux modèles d'attention visuelle inspirés par la vision humaine ont été proposés pour rendre compte de ces données comportementales. Il existe ainsi une large littérature sur les cartes de saillance, utilisées notamment pour coder efficacement des scènes visuelles. Deux modèles proposés dans la littérature ont retenu notre attention car ils possèdent tous deux des voies exogènes et endogènes tout en offrant des approches très complémentaires (cf. Figure 2).

Itti et al (2003) proposent un modèle neurobiologique d'attention visuelle et de mouvement des yeux et de la tête ainsi que son application à l'animation d'une tête humaine virtuelle réaliste. Le modèle d'attention visuelle exploite trois cartes visuelles : (a) une carte de saillance qui associe à chaque pixel de l'image une valeur de saillance : cette carte de saillance est obtenue par la combinaison de différentes cartes élémentaires (mouvement, orientation, intensité...), calculées à diverses échelles (ce que nous appellerons décomposition pyramidale). (b) une carte de pertinence module l'amplitude de la réponse de ces cartes en regard de la tâche de scrutation à réaliser (Navalpakkam and Itti 2005) ; (c) une carte d'attention gère le séquencement de la détermination des points d'intérêt en inhibant les zones de l'image déjà scrutées, c'est ce qu'on appelle l'inhibition de retour ou IOR (pour Inhibition Of Return).

Sun (2003) propose un modèle d'attention visuelle basé sur une segmentation préalable de la scène en objets. Dans cette approche on ne considère plus chaque

pixel de l'image indépendamment, la carte d'attention se chargeant d'inhiber la zone autour du pixel gagnant. Ainsi, il est possible de gérer une scrutation « syntaxique » de l'image, respectant un axe syntagmatique allant des objets les plus gros aux objets les plus fins et un axe paradigmatique allant de l'objet à ses composantes. Sun introduit en outre une inhibition de retour temporaire ou tIOR (pour temporary Inhibition Of Return) qui permet notamment à une zone de l'image d'être à nouveau sélectionnée si son contenu a changé.

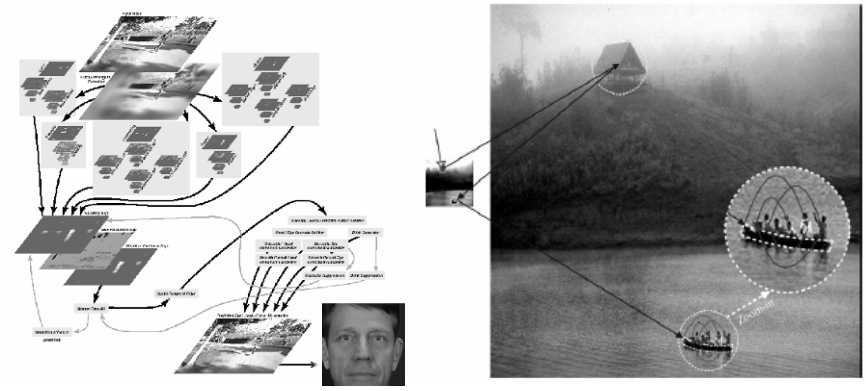

Figure 2: Modèles de scrutation de scènes naturelles. A gauche : les saccades oculaires prédites par Itti et al (2003) sont déclenchées par la détection de points d'intérêt sur la vidéo d'origine. A droite : Sun (2003) utilise une segmentation multi-échelles pour organiser de manière plus structurée la scrutation d'une image fixe.

3. Le modèle de scrutation proposé

Comme chez Itti et al., le modèle que nous proposons (voir Figure 3) s'appuie sur le calcul d'une carte de saillance sans segmentation préalable. La segmentation s'effectue après détermination d'un point d'intérêt en segmentant localement la région d'intérêt « accrochée » au point. Les caractéristiques colorimétriques de cette région sont alors caractérisées (notamment contrastées par rapport à la zone entourant la région). Elles sont mémorisées dans une pile d'attention et utilisées notamment pour suivre les mouvements de la zone.

Nous avons ainsi développé un mécanisme d'attention visuelle basé sur la gestion d'une pile d'attention mémorisant temporairement le contenu de la carte de saillance de régions préalablement visitées. Cette pile permet d'interrompre la scrutation d'une région du champ visuel pour traiter un stimulus exogène particulièrement saillant afin d'y revenir lorsque ce dernier a été traité. La pile permet aussi une inhibition temporaire de retour, la zone redevenant active si la zone d'intérêt a significativement changé par rapport au contenu de la pile. Nous avons également ajouté un mécanisme de poursuite lente basé sur le suivi des

caractéristiques colorimétriques de la zone d'intérêt par un filtre de Kalman. Un mécanisme additionnel de reconnaissance et de scrutation d'objets spécifiques a été ajouté afin de pouvoir faire émerger de la carte de saillance des zones à forte valeur informationnelle comme les visages.

Notons finalement que ce système réalise un couplage effectif entre génération des saccades oculaires et analyse rétinienne : un filtrage rétinien centré sur la position courante du cône fovéal est appliqué sur l'image et rend ainsi la carte de saillance sensible au regard porté sur l'image.

3.1. *Carte de saillance*

La carte de saillance combine les réponses de deux modules de traitement d'images : (a) le module « où » combine des cartes d'orientation (0 et 90°) et de mouvement calculées à diverses échelles sur l'image brute, remises à l'échelle la plus petite puis recombinées par simple adition (b) le module « quoi » combine des cartes de couleur et d'intensité calculées sur la sortie d'un filtre rétinien appliqué à l'image sur la résolution la plus haute. Ce filtre convolue chaque pixel avec un filtre gaussien d'ouverture proportionnelle à la distance du pixel avec le centre d'attention : l'image ainsi obtenue est de plus en plus floue au fur et à mesure que l'on s'éloigne du centre d'attention. La carte de saillance finale est obtenue en normalisant les cartes par leurs variances (obtenues expérimentalement) et les sommant. Le pixel qui remportera l'attention sera alors le pixel ayant la saillance la plus importante (Winner Take All ou WTA).

3.2. *Gestion de l'attention visuelle par une pile d'attention*

Contrairement à Itti et al, où l'inhibition de retour est réalisée par l'inhibition définitive de la zone alentour du point de saillance, la solution que nous avons retenue pour gérer le tIOR est une structure de pile, où l'on stocke la fenêtre locale de saillance correspondant à l'objet d'intérêt. La baisse d'attention de la zone est alors réalisée par différence entre cette fenêtre locale de saillance et la carte de saillance globale. Cela permet d'inhiber la zone même si sa saillance est élevée et de la réactiver si sa saillance change. La pile ne comprend que 3 éléments et fonctionne selon le principe FIFO (First In First Out) : s'il y a moins de 3 éléments, l'empilement d'un nouvel élément ajoute une zone inhibée supplémentaire. Un élément est enlevé si la saillance de zone empilée devient sensiblement différente de celle stockée ou si l'empilement d'un nouvel objet vient chasser l'élément le plus ancien. La pile peut aussi fonctionner en LIFO (Last In First Out), lorsque la fixation de l'élément courant (moyenne de 160ms) est interrompue par l'apparition d'un élément beaucoup plus saillant dans la scène – traité et éventuellement suivi – puis reprise là où on l'avait laissée.

3.3. *Suivi d'objet*

Dans les propositions d'Itti et de Sun, seule la carte de mouvement permet de rendre compte de la poursuite lente d'objets : on espère que l'élément le plus saillant de la carte de saillance restera ancré sur l'objet en mouvement. Ce n'est que rarement le cas, notamment lors du croisement de deux objets en mouvement. Nous avons ainsi ajouté un mécanisme spécifique de suivi d'objet exploitant un filtre de Kalman dont l'observation est actualisée par un simple modèle statistique colorimétrique de la région d'intérêt (analyse discriminante linéaire entre pixels appartenant à l'objet et au fond). Le suivi continue jusqu'à ce que la vitesse estimée de l'objet soit inférieure à un seuil. En mode suivi, le regard est ancré sur la trajectoire de l'objet (*œil de la mouche*) et la carte de saillance n'est pas utilisée de manière aussi prioritaire qu'en mode scrutation. En pratique, la carte de saillance n'est plus actualisée que toutes les 5 images : ce calcul permet à la fois de vérifier que l'objet que l'on suit est toujours saillant dans la scène et qu'un objet beaucoup plus saillant n'est pas apparu. Dans ce dernier cas, le nouvel objet devient le focus de l'attention et l'objet précédent est empilé.

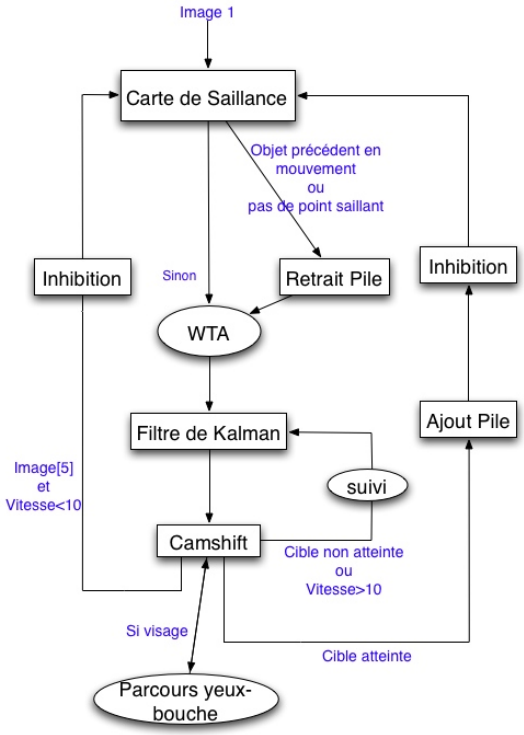

Figure 3 : Synopsis du système d'attention visuelle développé.

3.4. *Objet spécifique : détection de visages*

Le système de vision humain est informé : l'activité neuronale dans une zone spécifique du lobe temporal augmente ainsi de manière significative lorsqu'on observe des visages même tronqués *vs* des images comportant des visages déstructurés ou sans visage (Perrett, Rolls et al. 1982). Ces détecteurs câblés permettent de focaliser rapidement notre attention sur les objets de la scène les plus potentiellement porteurs d'informations. La carte de saillance de notre ACA a été ainsi augmentée d'une carte de détection de visage (utilisation du détecteur fourni dans OpenCV basé sur Lienhart and Maydt 2002). Un algorithme de détection de la position des yeux et de la bouche (Jiang, Binkert et al. 1998) a été adjoint permettant la scrutation informée des principaux éléments constitutifs du visage.

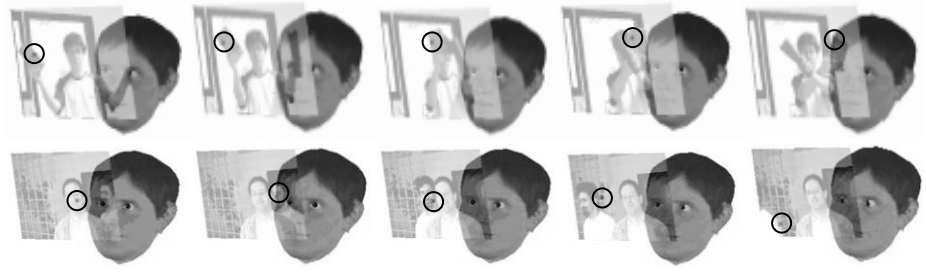

Figure 4: Exemples de scrutation de scènes naturelles par notre ACA. Pour les besoins de l'illustration, l'image ambiante acquise par la caméra de champ est incrustée dans un écran semi-transparent. Un cercle noir matérialise le point d'intérêt déterminé par notre système de gestion du regard. Les résultats sont donnés pour quelques images-clés de deux vidéos : en haut, le sujet agite un livre bleu en face de l'ACA et le module responsable de la poursuite d'objets contrôle le regard ; en bas, une personne va passer derrière l'interlocuteur, provoquant une saccade pour suivre ce nouvel objet d'intérêt.

3.5. *Résultats et évaluation*

Nous avons réalisé des tests sur 2 séquences vidéo réelles. Nous présentons Figure 4 les résultats obtenus sur des images-clés de ces séquences. Sur ces images, on peut voir le clone parlant en arrière plan avec la vidéo projetée en transparence devant ses yeux, le cercle noir représentant la position de son centre d'intérêt courant. Les résultats fournis par notre algorithme ont été comparés à des données oculométriques collectées sur 5 sujets auxquels on avait demandé de simplement décrire la scène observée. La Figure 5 montre la superposition des coordonnées écran du point d'attention calculé et des données collectées.

Dans la scène 1, un sujet agite successivement devant lui un objet bleu, un objet rouge, puis les 2 à la fois : cette scène permet de valider le suivi d'objet et la gestion de la pile d'attention (car le regard ne suit qu'un objet à la fois). La Figure 4 montre la partie de la vidéo où le sujet croise les deux objets : l'objet bleu pris par la main

gauche est suivi du regard alors que l'objet rouge est ignoré et ne sera jamais empilé. Cette cécité attentionnelle se retrouve dans les données oculométriques. Les différences majeures entre simulation et données sont au niveau du temps de réaction (notre système détecte le mouvement plus rapidement que les sujets) et du suivi de gestes tronqués (notre système reste sensible au mouvement du bras qui porte les objets bien que l'objet sorte du champ de vision alors que les sujets exploitent les liens de causalité entre les mouvements de ces segments).

Dans la scène 2, un sujet regarde la caméra alors que plusieurs autres sujets passent derrière lui, certains s'arrêtant pour regarder la caméra : cette scène permet de valider la détection de visage et le fonctionnement de la pile d'attention en mode LIFO (la scrutation du visage du sujet est souvent interrompue par le passage des autres sujets). La Figure 4 montre une partie de la vidéo où la scrutation du visage du sujet principal est interrompue au profit d'un suivi du visage en arrière plan. Les différences majeures entre simulation et données (voir Figure 5) sont au niveau des ordonnées : le système reste trop sensible aux couleurs saturées des vêtements portés par les sujets.

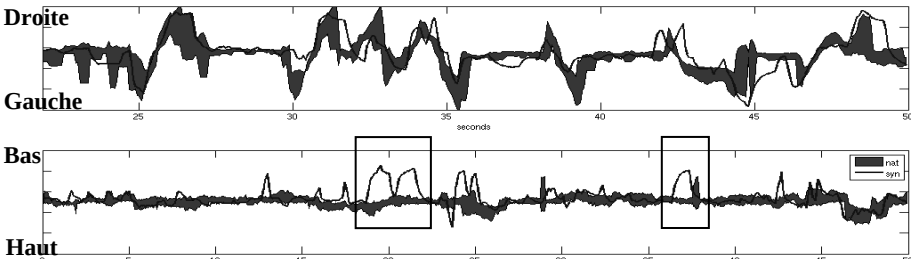

Figure 5: Comparaison des mouvements oculaires (haut: déplacement horizontal; bas : déplacement vertical) prédits par notre système avec celles enregistrées sur des sujets observant la même scène audiovisuelle où des personnes passent en arrière-plan de l'interlocuteur (le gabarit figuré en gras est à ±3 fois la variance des données sur 5 sujets). Les principales différences sont observées dans le déplacement vertical : en l'absence de toute stratégie de haut-niveau, le regard de l'ACA est parfois plus attiré par la couleur saturée des habits des sujets que par leurs visages.

3.6. *Performances*

Cet algorithme a été implémenté en langage C sous Linux Red Hat 9. Il utilise la bibliothèque OpenCV 0.9.9 de Intel. Les tests ont été réalisés sur un Pentium 4 (cadencé à 3.2 GHz). On obtient ainsi une vitesse d'exécution de 0.08s/image soit environ 12 images par seconde.

4. Conclusions et perspectives

Un système de scrutation de scènes naturelles a été développé et couplé au système de contrôle du regard d'un ACA. Il comporte un certain nombre de composants originaux tels que la gestion d'une pile d'attention, la détection de visages et le suivi d'objets en mouvement. La confrontation des simulations à des données oculométriques a montré la pertinence de ces choix et suggère quelques pistes d'amélioration, notamment l'ajout de composantes d'analyse descendante qui permettraient d'exploiter les connaissances *a priori* que nous avons sur les objets, leur mouvement et leurs éléments constitutifs les plus informatifs. Il serait à cet égard intéressant de gérer en miroir soit une pile d'intention comme suggéré par Chopra-Khullar et Badler (1999) soit une carte de pertinence (Navalpakkam and Itti 2005) afin que l'agent puisse se concentrer sur les tâches et être attentif aux évènements qui s'y rapportent.

Ce système combiné à un contrôle et un rendu plus précis du regard de l'ACA (voir Casari et al, ce volume) devrait nous permettre à moyen terme un test *in vivo* des capacités de notre agent à assurer une interaction face-à-face située efficace et crédible.

Remerciements

Ce travail a été financé par le projet « scrutation de scènes multimodales » d'ELESA, le projet « Deixis Multimodale » du GIS PEGASUS, et le projet « Présence » du cluster Infolog de la région Rhones-Alpes. Une première implémentation du modèle d'Itti et al a été réalisée par Romain Rossi.

Bibliographie

Brooks, R. A., C. Breazeal, et al. (1999). The Cog Project: Building a Humanoid Robot" in Computation for Metaphors, Analogy, and Agents. Lecture Notes in Artificial Intelligence. C. Nehaniv. New York, Springer**:** 52–87.

Chopra-Khullar, S. and N. I. Badler (1999). Where to look? Automating attending behaviors of virtual human characters. Annual Conference on Autonomous Agents, New York: 16-23.

Clodic, A., S. Fleury, et al. (2006). Rackham: an interactive robot-guide. IEEE International Workshop on Robots and Human Interactive Communications, Hatfield, UK.

Courty, N. (2002). Animation référencée vision : de la tâche au comportement. PhD Thesis. IRISA. Rennes, INSA**:** 198 pages.

Grossberg, S. (2003). "How does the cerebral cortex work? development, learning, attention, and 3d vision by laminar circuits of visual cortex." Behavioral and Cognitive Neuroscience Reviews **2**: 47-76.

Itti, L., N. Dhavale, et al. (2003). Realistic avatar eye and head animation using a neurobiological model of visual attention. SPIE 48th Annual International Symposium on Optical Science and Technology, San Diego, CA: 64-78.

Jiang, X., M. Binkert, et al. (1998). Detection of glasses in facial images. Asian Conference on Computer Vision, Hong Kong - China: 726-733.

Lee, S. P., J. B. Badler, et al. (2002). "Eyes alive." ACM Transaction on Graphics **21**(3): 637-644.

Lienhart, R. and J. Maydt (2002). An extended set of haar-like features for rapid object detection. IEEE International Conference on Image Processing, Rochester - NY: 900-903.

Matsusaka, Y., T. Tojo, et al. (2003). "Conversation Robot Participating in Group Conversation." IEICE Transaction of Information and System **E86-D**(1): 26-36.

Navalpakkam, V. and L. Itti (2005). "Modeling the influence of task on attention." Vision Research **45**(2): 205-231.

Perrett, D., E. Rolls, et al. (1982). "Visual neurones responsive to faces in the monkey temporal cortex." Exp Brain Research **47**: 329-342.

Simons, D. J. and C. F. Chabris (1999). "Gorillas in our midst: sustained inattentional blindness for dynamic events." Perception **28**(9): 1059-1074.

Sun, Y. (2003). Hierarchical object-based visual attention for machine vision. PhD Thesis. Institute of Perception, Action and Behaviour. School of Informatics. Edinburgh, University of Edinburgh**:** 169 pages.

Vatikiotis-Bateson, E., I.-M. Eigsti, et al. (1998). "Eye movement of perceivers during audiovisual speech perception." Perception & Psychophysics **60**: 926-940.

Yarbus, A. L. (1967). Eye movements during perception of complex objects. Eye Movements and Vision'. L. A. Riggs. New York, Plenum Press. **VII:** 171-196.