01234567890123456789012345678901234567890123456789012345678901234567890123

Université de Cergy-Pontoise

THÈSE

présentée pour obtenir le titre de DOCTEUR en Informatique et Traitement de l'Information

De la reconnaissance des expressions faciales à une perception visuelle partagée: une architecture sensori-motrice pour amorcer un référencement social d'objets, de lieux ou de comportements.

Sofiane Boucenna

E-mail : sofiane.boucenna@ensea.fr

Soutenue le 5 mai 2011 devant le jury composé de :

P. Gaussier,	directeur de thèse
F. Kaplan,	Rapporteur
M. Paindavoine,	Rapporteur
L. Hafemeister,	Co-directrice de thèse
J. Nadel,	Examinateur
D. Bazalgette,	Examinateur
R. Alami,	Examinateur

Cette thèse se concentre sur les interactions émotionnelles en robotique autonome. Le robot doit pouvoir agir et réagir dans un environnement naturel et faire face à des perturbations imprédictibles. Il est donc nécessaire que le robot puisse acquérir une autonomie comportementale, à savoir la capacité d'apprentissage et d'adaptation en ligne. En particulier, nous nous proposons d'étudier quels mécanismes introduire pour que le robot ait la capacité de se constituer une perception des objets de son environnement qui puisse être partagée par celle d'un partenaire humain. Le problème sera de faire apprendre à notre robot à préférer certains objets et à éviter d'autres objets. La solution peut être trouvée en psychologie dans ce que l'on appelle "référencement social" ("social referencing") qui consiste à attribuer une valeur à un objet grâce à l'interaction avec un partenaire humain. Dans ce contexte, notre problème est de trouver comment un robot peut apprendre de manière autonome à reconnaître les expressions faciales d'un partenaire humain pour ensuite les utiliser pour donner une valence aux objets et permettre leur discrimination.

Nous nous intéresserons à comprendre comment des interactions émotionnelles avec un partenaire peuvent amorcer des comportements de complexité croissante tel que le référencement social. Notre idée est que le référencement social aussi bien que la reconnaissance d'expressions faciales peut émerger d'une architecture sensori-motrice. Sans connaissance de ce que l'autre est, le robot devrait réussir à apprendre des tâches "sociales" de plus en plus complexes. Nous soutenons l'idée que le référencement social peut être amorcé par une simple cascade d'architectures sensori-motrices qui à la base ne sont pas dédiées aux interactions sociales.

Cette thèse traite de plusieurs sujets qui ont comme dénominateur commun l'interaction sociale. Nous proposons tout d'abord une architecture capable d'apprendre à reconnaître de manière autonome des expressions faciales primaires grâce à un jeu d'imitation entre une tête expressive et un expérimentateur. Les interactions avec le dispositif robotique commençent par l'apprentissage de 5 expressions faciales prototypiques. Nous proposons ensuite une architecture capable de reproduire des mimiques faciales ainsi que leurs différents niveaux d'intensité. La tête expressive pourra reproduire des expressions secondaires par exemple une joie mêlée de colère. Nous verrons également que la discrimination de visages peut émerger de cette interaction émotionnelle à l'aide d'une rythmicité implicite qui se crée entre l'homme et le robot. Enfin, nous proposerons un modèle sensori-moteur ayant la capacité de réaliser un référencement social. Trois situations ont pu être testées: 1) un bras robotique capable d'attraper et de fuir des objets selon les interactions émotionnelles venant du partenaire humain. 2) un robot mobile capable de rejoindre ou d'éviter certaines zones de son environnement. 3) une tête expressive capable d'orienter son regard dans la même direction que l'humain tout en attribuant des valeurs émotionnelles aux objets via l'interaction expressive de l'expérimentateur.

Nous montrons ainsi qu'une séquence développementale peut émerger d'une interaction émotionnelle de très bas niveau et que le référencement social peut s'expliquer d'abord à un niveau sensori-moteur sans nécessiter de faire appel à un modèle de théorie de l'esprit.

Je tiens à remercier Messieurs Frédéric Kaplan et Michel Paindavoine pour avoir accepté de rapporter ce travail de thèse et ceci malgré leurs emplois du temps très chargés. Je remercie Madame Jacqueline Nadel et Messieurs Rachid Alami et Didier Bazalgette d'avoir accepté d'être examinateurs de ma thèse.

Je remercie Laurence Hafemeister et Philippe Gaussier pour m'avoir encadré pendant trois ans durant lesquels j'ai pu bénéficier de leurs expériences, de leurs conseils et de leur disponibilité. Merci pour votre soutien et pour votre vision des sciences cognitives. Merci également pour vos discussions scientiques où l'on repart très souvent avec la tête pleine d'idées mais qui finissent trop souvent à la trappe...

Je tiens également à remercier Inbar Fijalkow, directrice du laboratoire ETIS, ainsi que les tous les membres constituants ce chaleureux laboratoire où il fait bon vivre ensemble.

Je tiens également à remercier tout particulièrement les membres de l'équipe Neurocyber pour leur gentillesse, leur disponiblité et leur discussion. Merci à Arnaud, Pierre, Julien, Philippe, Cyril, Antoine, David, Mathias, Christophe, Benoît, Fred (j'espère avoir oublié personne!)

Mes dernières pensées vont à ma famille sans qui rien n'aurait pu être possible... Merci mamam, merci papa. Merci à ma femme pour son soutien et de supporter mon caractère...

Table des matières

Première partie

Introduction et Outils

Chapitre 1

Introduction

1.1 Les robots dans la société

Les robots ont pris une place importante dans notre vie quotidienne allant du simple robot jouet à des robots beaucoup plus sophistiqués dotés de capacités bien spécifiques. Les robots ont depuis bien longtemps suscité énormément d'attention pour l'homme, débouchant parfois sur des films captivant l'intérêt du spectateur. Ces oeuvres qui mêlent fiction et réalité, incarnent la vie future où le robot a sa place à part entière dans la vie quotidienne. Les robots sont à l'image de l'humain tant sur le plan physique que sur le plan des compétences (raisonnement, mémoire, conscience ...) possédant même des émotions. Une harmonie se crée entre l'homme et le robot laissant suggérer une parfaite cohabitation. L'imaginaire du téléspectateur voyage dans le futur laissant supposer la possibilité d'une coopération entre ces deux entités l'une biologique et l'autre mécanique.
Le retour à la réalité est souvent brusque, des dizaines de laboratoires de recherche travaillent

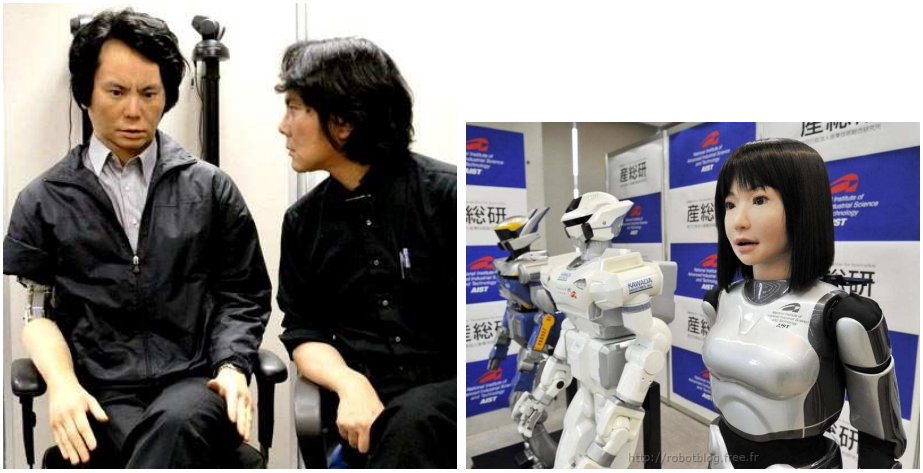
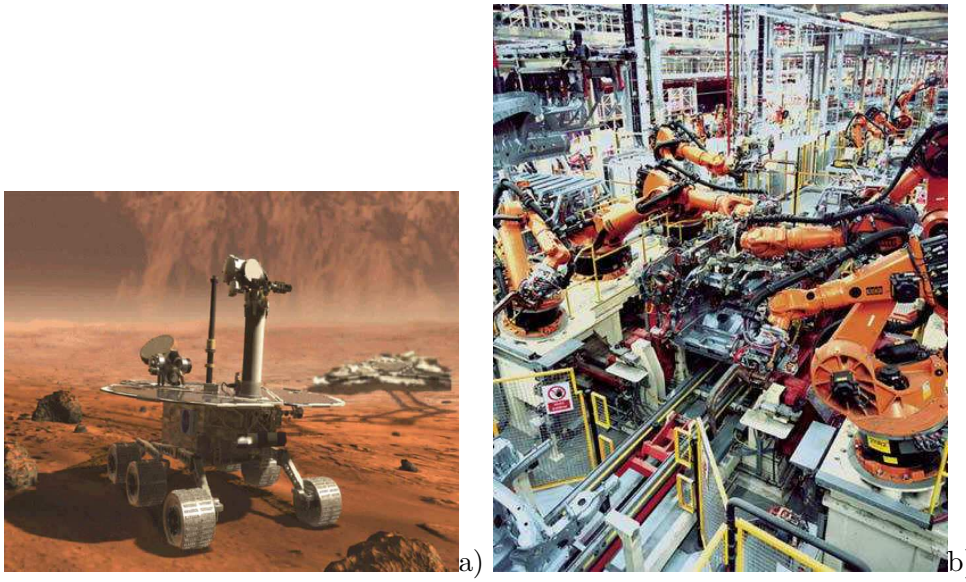


Fig. 1-1 – Robot humanoide ayant des caractéristiques physiques très proches de celles de l'humain.

d'arrache pied au développement de capacités "cognitives" tandis que d'autres essaient de créer des robots ayant une ressemblance physique avec l'homme (Figure 1-1). Cependant, le robot "intelligent" n'appartient pas à la majorité des utilisations. De nombreux robots existent dont les utilisations et les finalités sont aussi diverses que variées.

La première distinction importante est la capacité du robot à interagir avec un expérimentateur humain. Dans une perspective développementale, certains robots agissent sans l'aide de l'humain (Oudeyer et al., 2007; Hasson and Gaussier, 2010) tandis que d'autres sont contrôlés soit en partie soit totalement par ce dernier (Calinon et al., 2007). En général, lorsque les tâches sont peu complexes et répétitives, le robot est programmé pour agir seul, comme par exemple les robots travaillant à la chaine dans les usines. Néanmoins, les robots qui sont utilisés dans l'armée (drones) ou ceux utilisés dans les centrales nucléaires sont télé-opérés par un expert humain (le droit à l'erreur n'est pas permis). Ce premier aspect insiste sur la capacité du robot à être autonome vis à vis de l'humain.

La deuxième distinction met l'accent sur l'adaptation comportementale du robot. Il s'agit de comportements non stéréotypés qui sont généralement fonctions des changements environnementaux. Le robot s'adapte en permanence aux flux d'entrée. En revanche, les robots travaillant à la chaîne ne s'adaptent pas aux perturbations de l'environnement, contrairement aux robots explorateurs qui ont la capacité à éviter plusieurs types d'obstacles (Brooks, 1986) (Figure 1-2). L'adaptation est une capacité primordiale pour le robot afin qu'il se mouvoit dans son environnement.



a) b)

Fig. 1-2 – a) robot explorateur Spirit b) robot travaillant dans une usine

Enfin, la troisième distinction et non la moindre, porte sur l'acquisition de nouveaux comportements. L'apprentissage est un moyen d'acquérir de nouvelles capacités par l'intermédiaire de l'expérience. Certains mécanismes de "curiosité" implantés dans les robots, les poussent à s'intéresser aux choses nouvelles (Oudeyer et al., 2007). Les interactions avec le monde physique et social permettent au robot d'apprendre des compétences de plus en plus évoluées, un peu comme le fait l'enfant.

Ces distinctions mettent en évidence plusieurs aspects importants dans une recherche en robotique. L'autonomie, l'adaptation et l'apprentissage sont indispensables pour concevoir un robot ayant des comportements ressemblant à l'humain. Ainsi, le robot doit pouvoir interagir avec des expérimentateurs humains afin de développer des compétences sociales.

1.2 Pourquoi utiliser des robots?

Certains robots sont utilisés pour remplacer l'homme dans des tâches contraignantes (les usines automobiles), dangereuses (navigation dans des zones radioactives ou manipulation de produits toxique) ou les zones où l'homme n'accède pas (les planètes du système solaire). D'autres robots commencent à apparaître pour le grand public, ils sont utilisés dans les maisons réalisant des tâches simples comme nettoyer le sol ou tondre la pelouse comme par exemple le Roomba (société IRobot de Rodney Brooks). Les avancées technologiques ont permis leur introduction dans la vie quotidienne, ils sont dotés de capacités relativement simples (évitement d'obstacles, se déplacer dans des environnements nouveaux et inconnus). Une dernière catégorie de robots est celles des robots ludiques dont certains annoncent qu'ils constituent une nouvelle révolution comparable à celle de l'automobile. L'AIBO (Figure 1-3) et plus récemment le NAO représentent ce type de robot nouvelle génération, utilisés dans les laboratoires d'intelligence artificielle et possédant des capacités d'interactions sociales pour attirer l'attention de l'homme lorsqu'il est en manque "d'affection".

Fig. 1-3 – Le célèbre robot AIBO de sony utilisé dans les expériences de Kaplan et Oudeyer.

Enfin, les robots sont aussi considérés comme un outil dans les laboratoires de recherche. Notamment, en psychologie du développement où l'utilisation du robot peut faciliter l'interaction (médiateur). Le robot est considéré comme un outil pour tester certaines compétences cognitives. Cependant, de plus en plus de laboratoires considèrent le robot comme une vie artificielle à part entière ayant un cerveau et un corps. Les recherches qui en découlent, essayent de comprendre comment une vie artificielle peut-elle se développer? Les différents comportements qui sont acquis par le robot, sont le produit d'une interaction complexe entre controlleur, corps et environnement. Les modèles développés essaient de prendre en compte des interactions complexes entre le robot et l'environnement tant physique que social. Ces robots sont le substitut du jeune enfant permettant de tester certains modèles psychologiques et en retour d'aider les psychologues en leur faisant part des résultats robotiques par exemple en réalisant des lésions dans le modèle afin de voir les effets sur le comportement du robot. Le comportement du robot peut être comparé à celui du jeune bébé pour une fonctionnalité ou un ensemble de mécanismes donnés.

1.3 Le robot vu comme un enfant

Les bébés peuvent être comparés à des ordinateurs extrêmement puissants où les neurones sont des éléments indispensables afin de développer des capacités cognitives. L'enfant est façonné au travers l'évolution contrairement à l'ordinateur qui est programmé par un ingénieur expert

dans un domaine spécifique. Les bébés ont à leur disposition des mécanismes d'apprentissage extrêmement puissants leurs permettant de modifier, d'adapter et de restructurer leurs connaissances contrairement à l'ordinateur qui est dans l'incapacité de développer ce type de faculté extraordinaire faisant du bébé un enfant, de l'enfant un adolescent et de l'adolescent un adulte.

Le bébé est capable de s'imprégner du flux de données complexes venant du monde extérieur et de lui donner sens en agissant sur l'environnement. Le travail des psychologues du développement se situe à ce niveau là, ils travaillent à plusieurs niveaux d'abstraction qui sont les suivants: trouver quels sont les mécanismes utilisés par le bébé, déterminer comment ces mécanismes sont codés dans le cerveau et comment ils se développent. Le travail de cette communauté tente de répondre à la question philosophique de la connaissance. Pour nous roboticiens, les questions sont similaires, à savoir: comment un robot est-il capable de se développer de manière autonome avec un programme (modèle) minimal. L'adaptation prend une place importante dans tous les modèles que nous développons aussi bien sur des aspects de navigation, de reconnaissance d'objets ou encore de reconnaissance d'expressions émotionnelles. Les modèles développés offrent la possibilité au psychologue du développement d'analyser la plausibilité de nos modèles ou bien même encore de réfuter leur propre modèle.

Cependant, le bébé dispose du meilleur enseignant: l'adulte et plus particulièrement les parents. L'adulte lui même est conçu pour aider le bébé à apprendre. Ce soutien joue un rôle tellement crucial dans le développement du bébé qu'on ne peut pas dissocier les deux entités. De notre point de vue, le bébé et l'adulte ne forment qu'un seul système interagissant dans une harmonie perpétuelle. Les recherches tendent à montrer que l'immense finesse de l'homme concernant sa capacité d'adaptation, sa lutte pour la survie, est son extraordinaire facilité à apprendre lorsqu'il est bébé et à enseigner quand il est adulte. Jerome Bruner (Bruner, 1983; Wood et al., 1976) propose que la période d'immaturité du bébé pendant laquelle l'adulte le protége, permettrait à l'enfant de découvrir son environnement physique et social.

Cette capacité du "vivre ensemble" semble primordiale pour la survie et le développement de l'espèce. Il nous paraît important de prendre cette dimension en compte pour la construction de robot autonome. De la même manière que pour les bébés, il paraît évident que le développement d'un robot doit forcément être vu d'une manière développementale et implique par conséquent l'aide d'un expérimentateur où l'interaction entre les deux entités ne forme qu'un seul système indissociable.

1.4 L'importance des émotions pour l'homme et par conséquent pour le robot

L'exemple de Phineas Gage montre l'importance des émotions dans les comportements sociaux (Damasio, 1994). En 1848, Phineas Gage est blessé à la tête alors qu'il travaillait à la construction d'un chemin de fer dans le Vermont aux Etats-Unis. Suite à une explosion, une tige de fer pointue transperça le crâne de Phineas Gage entrant par la joue et ressortant dans le haut du crâne (Figure 1-4). Contre toute attente, il survécut à ce terrible accident, réussissant même à parler quelques minutes plus tard. Le docteur Harlow soigna Phineas Gage pendant de long mois. Il semblait n'avoir plus aucun déficit neurologique apparent. Gage utilisait tous ses membres et pouvait parler. Malgré son rétablissement physique, son comportement social avait changé. Phineas n'était plus le même, il était d'humeur changeante, grossier, inconstant et très capricieux. Son statut social n'allait qu'en s'aggravant, il ne dirigeait plus les travaux de ses chantiers et finit comme attraction dans un célébre cirque de New York.

Les dommages causés à son cortex frontal ont induit une perte des inhibitions sociales avec un comportement social parfois très déplacé. L'histoire de cet homme est sans aucun doute un tournant important en neurosciences des émotions. L'étude de ce pauvre homme montre que les lésions corticales (cortex préfrontal, amygdale) n'induisent pas forcément des déficits au niveau du langage, de la mémoire, ou même du raisonnement logique mais peuvent induire des troubles du comportement émotionnel et social.

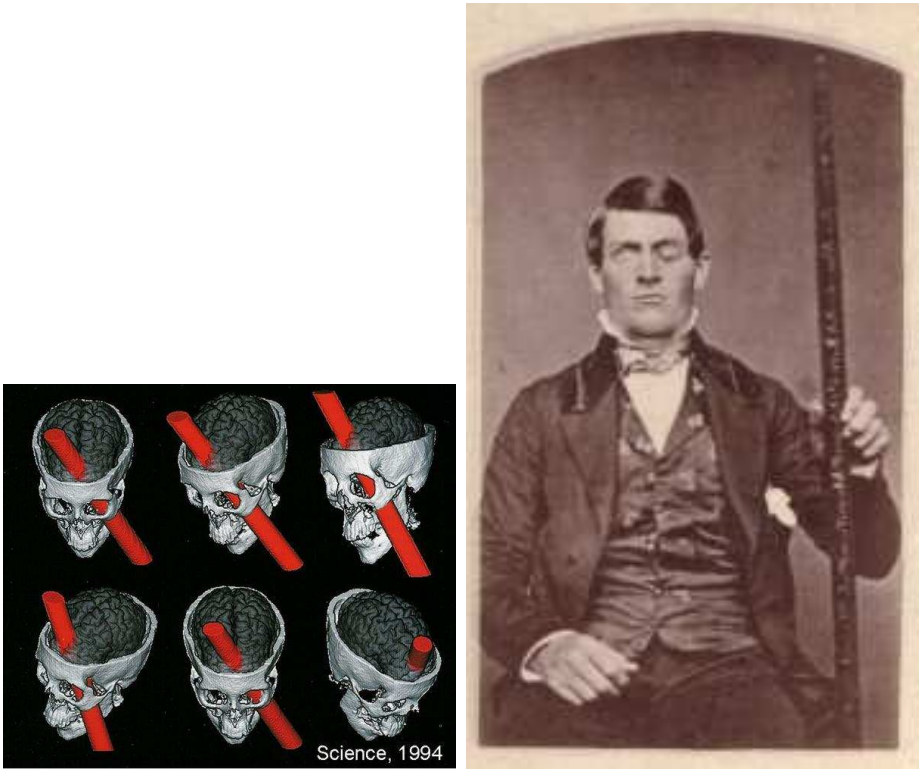


Fig. 1-4 – Phineas Gage.

Dans l'erreur de Descartes (Damasio, 1994), Damasio montre que l'émotion permet l'élaboration de plans d'action et l'anticipation des comportements. Elle permet d'avoir des raisonnements sociaux appropriés mais également la prise de décisions rationnelles. L'émotion influence donc le processus de raisonnement. Sans les mécanismes émotionnels, les prises de décisions et les raisonnements en groupe peuvent être altérés, nous faisons le choix d'investir dans cette direction de recherche pour la création de robots autonomes.

1.5 Contributions

Ma thèse essaie de montrer dans quelle mesure les émotions permettent le partage d'une expérience. Le terme de "perception partagée" utilisé dans le titre de ce manuscrit suscite certainement des interrogations. La question qui en découle, est comment une expérience peut être partagée entre deux entités (homme-homme, homme-robot, robot-robot). L'idée principale défendue dans cette thèse, est que l'émotion est un excellent moyen de communication dans un

contexte non-verbal et peut être le facilitateur de ce partage. En d'autres termes, nous montrerons comment l'émotion, et plus particulièrement les expressions faciales émotionnelles donnent un sens à l'environnement. Ainsi des objets, des lieux et parfois même des comportements acquièrent des valeurs émotionnelles laissant suggérer que la transmission de messages émotionnels entre un homme et un robot a pour but de partager une expérience commune.

Nous nous proposons d'étudier quels mécanismes introduire pour que le robot ait la capacité de se constituer une perception des objets de son environnement qui puisse être partagée par celle d'un partenaire humain. L'enjeu sera de faire apprendre à notre robot à préférer certains objets et à en éviter d'autres. Une réponse plausible peut être trouvée en psychologie dans ce que l'on appelle "référencement social" ou "social referencing" (Feinman, 1982) qui consiste à attribuer une valeur à un objet grâce à l'interaction avec un partenaire humain. Dans ce contexte, nous étudierons comment un robot peut apprendre de manière autonome à reconnaître les expressions faciales d'un partenaire humain pour ensuite les utiliser pour donner une valence aux objets et permettre leur discrimination.

Nous nous intéresserons à comprendre comment des interactions émotionnelles avec un partenaire peuvent amorcer des comportements de complexité croissante allant de la reconnaissance des expressions faciales au référencement social. Sans connaissance de ce que l'autre est, le robot devrait réussir à apprendre des tâches "sociales" de plus en plus complexes. Nous soutenons l'idée que le référencement social peut être amorcé par une simple cascade d'architectures sensori-motrices qui à la base ne sont pas dédiées aux interactions sociales.

L'apprentissage des différentes capacités est principalement basé sur l'ambiguité de la perception, à savoir sur l'incapacité à faire la différence entre son propre corps et celui de l'autre s'il y a corrélation entre ses propres actions et celles d'autrui. On montrera que l'ambiguité de la perception couplée avec un système homéostatique est suffisant pour apprendre à reconnaître les expressions faciales de l'expérimentateur, à attribuer des valeurs émotionnelles à des objets arbitrairement positionnés dans l'environement du robot et à apprendre à un bras robotique à rejoindre ou à attraper des objets.

Après avoir présenté l'état de l'art sur les émotions et leurs fonctions régulatrices du comportement dans le chapitre 2 ainsi que les outils utilisés pour le développement de nos modèles (chapitre 3), nous commencerons par décrire les expériences portant sur les interactions diadiques. La première expérience réalisée que nous décrirons au chapitre 4 porte sur la capacité d'un robot à apprendre à reconnaître de manière autonome des expressions faciales primaires grâce à un jeu d'imitation entre une tête expressive et un expérimentateur. Dans le chapitre 5, nous verrons également que la discrimination de visages peut émerger de cette interaction émotionnelle à l'aide d'une rythmicité implicite qui se crée entre l'homme et le robot. Dans le chapitre 6, nous proposons une architecture capable de reproduire des mimiques faciales à différents niveaux d'intensité. La tête expressive reproduira des expressions secondaires par exemple une joie mêlée de colère. Les contributions majeures de cette première partie sont:

- le développement d'un modèle ayant la capacité d'apprendre à reconnaître en ligne et de manière autonome les expressions faciales de l'humain.
- l'interaction émotionnelle amorce la détection de visage. Le robot est capable d'apprendre à discriminer un visage d'un non visage en ligne et de façon autonome au travers l'interaction avec l'expérimentateur.
- la capacité à reproduire des intensités d'expressions faciales primaires et secondaires.

Enfin la dernière partie de cette thèse se focalisera essentiellement sur des interactions triadiques entre le partenaire social, le robot et le reste de l'environnement (figure 1-5). Nous proposerons un modèle sensori-moteur ayant la capacité de réaliser un référencement social. Trois situations sont

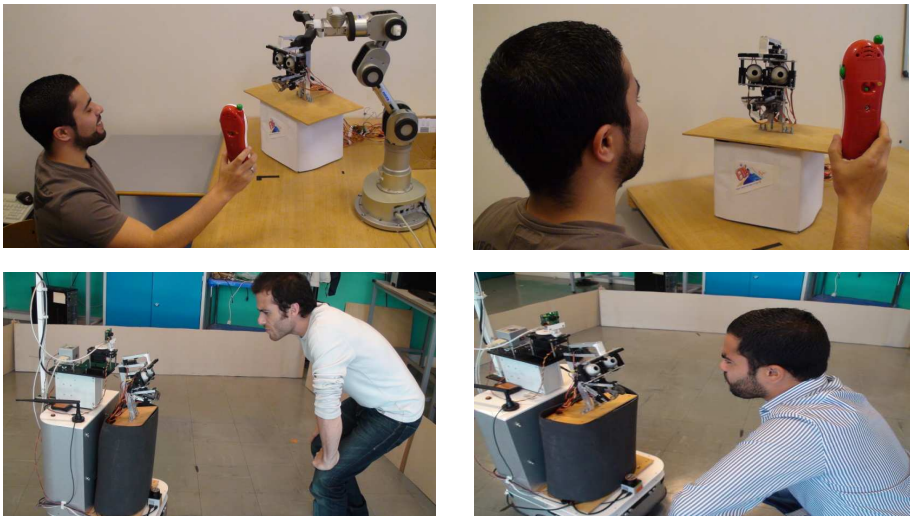


Fig. 1-5 – Les différents set-up expérimentaux montrant des interactions triadiques entre l'expérimentateur, l'objet (ou le lieu) et le robot.

décrites dans les derniers chapitres. Dans le chapitre 7 nous montrerons qu'un robot humanoïde doté d'un bras et d'une tête expressive peut apprendre à attraper ou à fuir des objets selon ses interactions émotionnelles avec le partenaire humain. Dans le chapitre 8, nous mettrons en évidence qu'un robot mobile est capable de rejoindre ou d'éviter certaines zones de son environnement grâce à l'expressivité faciale de l'expérimentateur. Le chapitre 9 décrit une tête expressive capable d'orienter son regard dans la même direction que l'humain tout en attribuant des valeurs émotionnelles aux objets via l'interaction expressive de l'expérimentateur. Nos études ont montré:

- une séquence développementale émergeant d'une interaction émotionnelle de très bas niveau.
- le référencement social peut s'expliquer d'abord à un niveau sensori-moteur sans nécessité de faire appel à un modèle de théorie de l'esprit.
- Des cascades d'architectures sensori-motrices peuvent résoudre des interactions triadiques (homme-objet-robot) dans des cadres expérimentaux différents: manipulation d'objets, naviguation, attention conjointe.

L'ensemble de ces contributions nous amènerons à une conclusion au chapitre 10.

Chapitre 2

Les émotions et leurs fonctions régulatrices du comportement

2.1 Introduction

Qu'est-ce que l'émotion? Selon le Larousse:

- Trouble subit, agitation passagère causés par un sentiment vif de peur, de surprise, de joie, etc. : Parler avec émotion de quelqu'un.
- Réaction affective transitoire d'assez grande intensité, habituellement provoquée par une stimulation venue de l'environnement.
- Sous l'Ancien Régime, révolte populaire non organisée et généralement de courte durée.

Le Petit Robert définit l'émotion (du latin "ex-movere" qui signifie mouvement vers l'extérieur) comme un état affectif intense, caractérisé par une brusque perturbation physique et mentale où sont abolies, en présence de certaines excitations ou représentations très vives, les réactions appropriées d'adaptation à l'événement (au sens affaibli, on parle d'un état affectif, plaisir ou douleur, nettement prononcé).

Sur wikipédia, l'émotion (action de mouvoir, mouvement) est une manifestation physique liée à la perception d'un événement dans l'environnement (externe), ou dans "l'espace mental" (interne). De vastes réseaux de structures cérébrales sont impliquées dans la perception, le traitement et la régulation des émotions, qui influencent d'autres phénomènes psychologiques tels que l'attention, la mémoire ou le langage (verbal et non-verbal).

Quand à lui, Damasio donne une définition un peu plus riche: "On peut définir les émotions tout simplement comme une modification transitoire de l'état de l'organisme dont les causes sont spécifiques. De même, on peut définir simplement le fait d'éprouver une émotion comme la représentation de cette modification transitoire de l'état de l'organisme sous forme de configurations neuronales et des images induites par ces dernières. Lorsque quelques instants plus tard, ces images s'accompagnent du sentiment de soi que l'on a dans l'acte de connaître et qu'elles parviennent sur le devant de la scène, elle deviennent conscientes. On a alors véritablement affaire à un sentir de sentiment". Damasio propose de réserver "le terme sentiment à l'expérience mentale et privée d'une émotion, et d'utiliser au contraire le terme émotion pour désigner l'ensemble de réponses qui, pour bon nombre d'entre elles, sont publiquement observables".

Ces définitions montrent que les émotions sont constitués d'un ensemble de réactions complexes où l'expressivité émotionnelle peut être visible de l'extérieur (expression faciale, geste, réponse d'évitement ...) mais peut également être vécue intimement (non visible) par exemple

les réponses physiologiques (accélération du rythme cardiaque). L'expérience émotionnelle est une caractéristique importante des émotions, elle décrit les sensations subjectives, c'est à dire les choses que l'on ressent (la peur, la joie, la tristesse ...). L'expressivité émotionnelle et l'expérience émotionnelle interagissent afin de décrire l'émotion.

Les travaux sur les émotions ont des intêrets différents. Dans ce chapitre, nous aborderons différents thèmes: les différents modèles théoriques des émotions, le développement émotionnel du jeune enfant, la résonance émotionnelle et enfin les interactions triadiques avec les aspects de référencement social.

2.2 Les différents modèles théoriques

2.2.1 James-Lange vs Cannon-Bard

James et Lange ont proposé l'une des premières théories sur les émotions en 1884 (James, 1884) et 1885 (Lange, 1885) (défendant chacun de leur côté une théorie amenant beaucoup de controverse concernant le ressenti émotionnel). Ils abordent le problème théorique d'une manière tout à fait originale où pour eux ce qui était considéré comme la conséquence d'une émotion est maintenant considérée comme une cause. Le déclenchement d'une émotion spécifique serait déterminé par la perception d'un motif spécifique. Autrement dit, "j'ai peur car je constate que je tremble" (Figure 2-1). Leur théorie amène une touche de renouveau pour l'époque et prend toutes les conceptions classiques de l'émotion à contre pied comme celle de Spinoza (Spinoza, 1677) qui explique que trois éléments permettent de décrire toutes les expériences émotionnelles: la joie, le désir et la tristesse ou encore celle de Descartes (Descartes, 1649) reposant sur 6 primitives émotionnelles: l'émerveillement, l'amour, le désir, la joie, la haine et la tristesse. L'idée sous-jacente est que chaque émotion posséderait son propre motif de changements physiologiques (Rime and Giovannini, 1986).

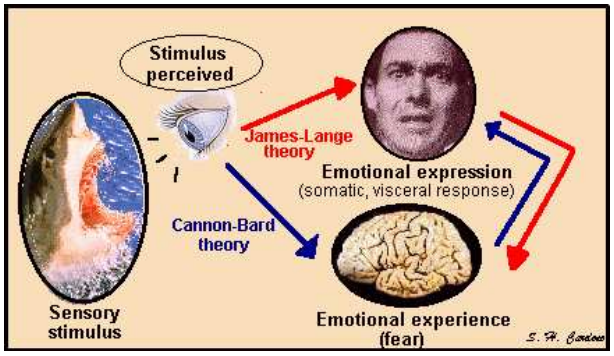


Fig. 2-1 – Illustration de la théorie de James-lange versus Cannon-Bard

Au contraire, Cannon en 1927 (Cannon, 1927) et Bard en 1928 (Bard, 1928) mettent en avant l'importance du système nerveux central (notamment du thalamus) dans le déclenchement d'une émotion. Par conséquent, les changements physiologiques ne sont pas vus comme des causes mais comme des conséquences d'une émotion (Figure 2-1). Cannon a réalisé de nombreuses études empiriques mettant en défaut la théorie de James-Lange. Par exemple, il constate que des réactions viscérales semblables semblent se produire dans de nombreuses émotions et également pour des états non émotionnels, il constate également que la suppression des afférences viscérales n'affecte pas les émotions. Cependant, ces résultats sont apprendre avec des précautions car

certains auteurs comme (Fraisse, 1963) montre que les critiques de Cannon présenterait des faiblesses.

Ces deux théories diamétralement opposées dans leur conception nourrissent encore de nombreux débats. Damasio (Damasio, 1994) reprend l'idée de James-Lange où l'accent est mis sur le rôle causal des changements corporels dans les émotions. L'introduction des marqueurs somatiques par Damasio soutiennent la théorie de James-Lange. Les marqueurs somatiques sont vus comme des réactions physiologiques associées à des événements passés. Ils seraient activés pendant le traitement d'un nouvel événement biaisant les prises de décisions en lien avec les conséquences potentielles de cet événement. D'autre part, les théories de la rétroaction faciale qui montrent que les mouvement faciaux modulent le ressenti émotionnel ont pour base la théorie de James-Lange, notamment James (James, 1892) l'avait déjà mentionné.

2.2.2 Théorie bifactorielle des émotions

Schachter (Schachter and Singer, 1962) introduit la théorie bifactorielle où l'émotion est déterminée par une interaction entre:

- une activation physiologique
- une cognition concernant la situation déclenchante de cette activation physiologique

L'excitation physiologique détermine l'intensité mais non la qualité de l'émotion. Elle est totalement indépendante de l'émotion (non spécifique à une émotion). L'interprétation de la situation permet d'identifier la nature de l'émotion ressentie. Pour Schachter, la cognition détermine si l'état d'activation physiologique sera labellisé comme "joie", "colère", "tristesse", ...

Cette théorie est en accord avec Cannon-Bard car les changements physiologiques ne sont pas spécifiques à une émotion mais elle est également en accord avec James-Lange car une activation physiologique est nécessaire pour la production d'une émotion. Schachter est l'un des pionniers dans le sens où cet auteur a une approche cognitive des émotions. L'émotion est déterminée comme l'existence d'une cognition modulée par l'activation physiologique. La théorie bifactorielle utilise la dimension sociale (informations disponibles dans l'environnement).

2.2.3 Théorie des émotions de base

Certains chercheurs ont introduit l'existence d'un nombre limité d'émotions universelles qui posséderaient chacune une fonction évolutionnaire appelées généralement émotions primaires (Ekman, 1982; Izard, 1971). Les différents auteurs considèrent la colère, la peur, la joie, la tristesse, le dégoût comme des émotions primaires même si des débats existent toujours notamment pour la surprise. Cette théorie a comme point de départ les travaux de Darwin en 1872 traduit dans (Darwin, 1965) qui a une approche évolutionniste vis à vis des expressions émotionnelles, une certaine continuité existe entre les différentes espèces animales. Darwin s'intéresse à plusieurs aspects notamment la question de la sélection des expressions émotionnelles au cours de l'évolution ainsi que leurs fonctions adaptatives à l'environnement. Il s'intéresse particulièrement aux expressions faciales où il souligne leur fonction communicative (Figure 2-2). L'évolution a joué un rôle majeur pour les émotions, plus particulièrement la notion d'adaptation. Les émotions se seraientt développées pour arbitrer certains mécanismes importants pour l'organisme comme l'attention, l'apprentissage ou la mémoire (Cosmides and Tooby, 2000). Les émotions sont définies comme des réactions transitoires, bio-psycho-sociales conçues pour aider les individus à s'adapter et à faire face à des événements qui ont des implications pour leur survie et leur bien-être (Matsumoto and Ekman, 2009).

Fig. 2-2 – Expressions faciales émotionnelles. Ces expressions faciales représentent les émotions universelles de P. Ekman aussi appelées émotions primaires

Ekman (Ekman, 1992) défend l'idée que les émotions de base partagent des propriétés communes. Elles seraient présentes chez plusieurs espèces, se déclencheraient rapidement et automatiquement et apparaitraient spontanément. Les émotions de base possèdent des motifs expressifs spécifiques car l'étude des émotions est souvent étroitement liée aux expressions faciales. Pour Ekman, l'expression faciale est vue comme le pivot de communication entre les hommes. (Ekman and Friesen, 1971) montre l'universalité des expressions faciales. Les expressions faciales n'ont pas de frontières quelque soit l'endroit du globe l'homme possède la capacité à reconnaître les expressions faciales de peuples et de cultures différentes. Un résultat relativement intéressant nous montre que certaines expressions faciales comme la peur permettent d'augmenter l'acquisition sensorielle par exemple l'ouverture des yeux, du nez et de la bouche (Susskind et al., 2008). Elles seraient peut être la conséquence évolutionniste de mécanismes qui au départ n'avaient pas toute la complexité des émotions.

Enfin, les émotions de base seraient liées à des régions cérébrales spécifiques. Ledoux (LeDoux, 1996) postule pour l'existence de systèmes neuronaux séparés pour les différentes émotions. De nombreux travaux suggèrent que l'amygdale traiterait la peur, l'insula traiterait le dégoût ou encore que l'accumbens s'occuperait de plaisir. Cependant, l'amygdale n'a pas une fonctionnalité unique: elle permet le conditionnement de manière générale (Bechara et al., 1995; Cardinal et al., 2002) ainsi que d'attribuer des valeurs émotionnelles aussi bien positives que négatives aux stimulations entrantes (Paton et al., 2006). Néanmoins, nous ne détaillerons pas ces différentes structures.

2.2.4 Les théories dimensionnelles

L'approche dimensionnelle est fondée sur l'existence de dimensions élémentaires indépendantes qui selon Russel (Russell and Feldman-Barrett, 1999) sont des propriétés basiques. Le modèle de Russell (Russell, 1980) décrit les émotions autour d'un cercle dont deux axes sont nécessaires: la valence correspondant au plaisir/déplaisir et l'activation faible/forte. Ce modèle circulaire est appelé circumplex (Feldman-Barrett and Russell, 2009) (Figure 2-3). D'autres théories di-

mensionnelles comme (Wundt, 1897) disposent de 3 dimensions de base: plaisant/déplaisant, tension/relaxation éprouvée et le caractère excitant/déprimant. Cette approche est soumise à

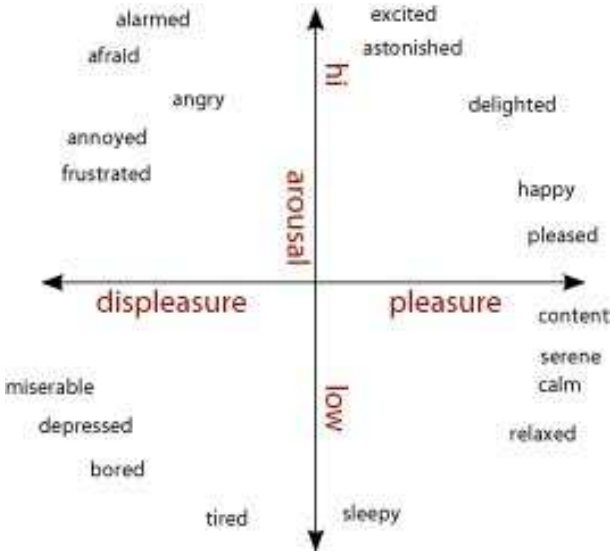


Fig. 2-3 – Le modèle circumplex (Feldman-Barrett and Russell, 2009)

de nombreuses critiques:

- La capacité à différencier les émotions. La peur et la colère se retrouvent généralement à la même place sur le cercle.
- Les dimensions élémentaires ne sont pas les mêmes pour tous les auteurs (Fontaine et al., 2007) montrant un manque de consensus.
- Le manque de rigueur sur une définition des concepts notamment sur les aspects d'intensité émotionnelle.

2.2.5 Les théories de l'évaluation cognitive de l'émotion

Cette approche est fondée sur un traitement cognitif de type évaluatif qui est à l'origine des émotions. Elle postule que l'évaluation d'un stimulus, d'un événement ou d'une situation déclenche une émotion comme dans les travaux de Sander (Sander et al., 2005). Ces modèles mettent en évidence que les organismes explorent leur environnement et réagissent aux stimuli pertinents. L'évaluation cognitive serait réalisée grâce à un ensemble de critères. Les critères d'évaluation sont généralement les mêmes: nouveauté, caractère plaisant, prédictibilité ... Les travaux d'Arnold (Arnold, 1960) ont été précurseurs pour cette approche théorique des émotions, des résultats expérimentaux ont renforcé l'idée que les différentes émotions peuvent être décrites en termes de pattern d'évaluation cognitives différentes (Frijda, 1987; Scherer, 1993).

Pour Frijda (Frijda, 1986), le processus émotionnel comporte plusieurs phases de traitement aboutissant sur une action, permettant une adaptation à la situation. L'approche de cet auteur considère que les émotions consistent à évaluer les événements induisant des actions.

Scherer (Scherer, 1989) quand à lui insiste plus particulièrement sur le processus d'évaluation qui est une succession rapide d'étapes de traitement des stimuli, élaborant ainsi le modèle des processus composants. Ces processus seraient le point de départ de la conception des émotions.

Le modèle défendu par Scherer nous questionne sur la séquence émotionnelle et les interactions dynamiques entre les différentes composantes constituant l'émotion. Cette approche met en évidence le versant dynamique des émotions. Scherer avance l'idée que les émotions possèdent plusieurs dimensions:

- une dimension évaluation cognitive permettant le déclenchement et la différenciation des émotions
- une dimension physiologique correspondant aux changements corporels
- une dimension expression motrice: expression faciale, vocale, posturale et gestuelle
- une dimension motivationnelle: tendances à l'action (approche, évitement)
- une dimension sentiment subjectif correspondant à l'expérience émotionnelle vécue

2.3 L'enfant et le développement des émotions

Les émotions chez l'enfant évoluent avec l'âge dans le but d'optimiser des stratégies adaptatives et régulatrices intra-personnelles et inter-personnelles. Les états émotionnels sont constitués d'activités psycho-biologiques, dont l'étude développementale permet d'une part d'avoir une chronologie de la construction du processus émotionnel et d'autre part de comprendre les raisons de certains comportements pouvant paraître minimaux. Certains auteurs comme (Brun and Nadel, 1997) postulent que les émotions permettent une adaptation aux contextes sociaux alors qu'elles étaient considérées dans le passé comme des états subjectifs pouvant brouiller le fonctionnement mental.

Cette manière originale d'appréhender les émotions induit de nombreuses études empiriques à différents niveaux aussi bien sur le versant moteur (expressivité émotionnelle) que sur le versant de la compréhension des émotions. Pour nous roboticien, les enjeux sont multiples à savoir quelles sont les compétences que les enfants sont capables de développer dès leur plus jeune âge? Qu'est ce qu'ils comprennent des autres? et comment font-ils le lien entre la compréhension émotionnelle et l'expressivité émotionnelle?

Cependant, les méthodes expérimentales sont généralement adaptées à l'âge des sujets que l'on teste. Pour les expressions faciales émotionnelles, les protocoles utilisés sont les méthodes de préférence visuelle (Fantz, 1961), méthodes d'habituation-deshabituation (Sokolov, 1963) et d'intégration de différentes modalités (Walkers, 1982).

2.3.1 Les premiers jours de la vie

De nombreux auteurs comme (Muir and Nadel, 1998) montrent que le nourrisson de 4 à 8 semaines est capable de sourire et que cet acte est synonyme d'expression de joie lorsqu'il s'adresse à un objet social. Autrement dit, très tôt la capacité à sourire devant un événement social est présente chez le nourrisson. Qui peut résister devant le sourire d'un bébé, ces sourires conduisent forcément à des contacts aussi bien physiques qu'émotionnels avec ce dernier. Cette empathie que, nous avons pour le jeune enfant est très certainement déclencheur d'apprentissage de compétences diverses et variées, tant pour le développement émotionnel que moteur. Dans les premières semaines de vie, le bébé ne possède pas la capacité de se mouvoir mais son répertoire moteur est suffisamment riche pour communiquer et susciter l'intérêt des personnes qui l'entourent.

Très tôt, le nourrisson est capable de produire des expressions prototypiques de joie, de tristesse, de dégoût, de colère ... et d'en différencier certaines en fonction de l'émotion mater-

nelle (Izard et al., 1995). Les bébés âgés de 2 mois sont capables de répondre aux expressions émotionnelles de manière appropriée. Ce résultat a été montré lors de la réciprocité de l'échange entre bébé et expérimentateur (Tronick et al., 1978; Trevarthen, 1993). Entre 2 mois et 2 mois et demi, le nourrisson possède cette capacité qui peut être vue comme un moyen de communication entre lui et la mère. Cependant, d'autres auteurs comme (Young-Browne et al., 1977) affirment que la capacité à discriminer des visages tristes ou joyeux et des visages tristes ou surpris n'a pas été trouvée chez des enfants de 3 mois tandis qu'ils sont capables de discriminer les visages joyeux, des visages surpris. Selon Serrano (Serrano et al., 1992), les enfants de 4 à 6 mois sont quand à eux capables de discriminer les expressions faciales de colère, de peur et de surprise. Ce sont sans doute les premières étapes d'une communication émotionnelles entre la mère et le bébé. Ce développement de l'expressivité faciale coïncide avec l'acquisition de capacités de discrimination perceptive plus fines des stimuli émotionnels. Entre 4 et 7 mois, le bébé acquiert la faculté de catégorisation perceptive (De Haan and Nelson, 1998). Selon (Ludemann and Nelson, 1988), la capacité à catégoriser les émotions n'est disponibles qu'à l'âge de 7 mois. (LaFreniere, 2000) montre que le bébé commence à avoir des réactions empathiques élémentaires, les stimuli émotionnels sont contagieux aux autres nourissons.

Certaines expériences se sont focalisées davantage sur la production des expressions faciales des bébés dans des situations plus réelles que l'analyse des photos où la situation contextuelle n'a pas d'influence. Ainsi, dans l'expérience de Muir avec la double vidéo (Muir et al., 2005), les auteurs montrent que les bébés perçoivent les changements émotionnels plus tôt. (Field et al., 1982) montrent que des nourrissons de 36 heures discriminent la joie, la tristesse et la surprise dans des interactions en face à face. (Nadel et al., 1999) montrent que les bébés de 2 mois réagissent de façon négatives devant une mère non contingente émotionnellement.

Ces compétences acquises durant les premiers mois de la vie nous montrent que le nourrisson se développe au fur et à mesure des interactions avec son environnement physique et social. Il commence par sourire comme s'il désirait envoyer des messages, demandant de lui porter de l'intérêt. Par la suite, des compétences plus fines comme la discrimnation et la catégorisation d'expressions émotionnelles. Ces facultés n'impliquent pas forcément la compréhension des émotions mais néanmoins elles sont une porte d'entrée pour y aboutir. Enfin, des réactions empathiques élementaires vis à vis de ses congénères émergent durant les tous premiers mois de la vie impliquant la capacité à ressentir les émotions.

2.3.2 L'acquisition de compétences durant la première année

Les études en psychologie du développement montrent que la réactivité émotionnelle est une capacité que le bébé possède très tôt par conséquent la production, la discrimination et la catégorisation des stimuli émotionnels sont des éléments qui sont indispensables à la construction du jeune bébé. Les études de (Walker-Andrews, 1986; Walker-Andrews, 1997) mettent en évidence que des bébés de 5 à 7 mois, possèdent la capacité à reconnaître une expression faciale correspondant à son expression sonore. Le protocole d'intégration intermodal est utilisé entre la vision et l'audition. On présente à l'enfant deux séquences vidéos correspondant chacune à des expressions faciales émotionnelles différentes (joie et colère) sur deux écrans placés l'un à coté de l'autre. Dans le même temps, un enregistrement sonore correspondant à l'une des deux vidéos est diffusé. Par la suite, le temps de fixation visuel permet de mettre en évidence, la capacité du bébé à réaliser une association entre le stimulus visuel et auditif. Ces travaux soulignent la faculté du bébé à extraire les informations indispensables pour ces deux médias émotionnels. Le

bébé possède la capacité d'intégration intermodale dans la mesure où il a une préférence pour les stimuli visuels dont la dimension émotionnelle correspond aux stimuli auditifs.

Cette faculté se développe durant les premiers mois du bébé. Les données montrent que les bébés de 2 mois fixent exclusivement les expressions de joie même si l'expression sonore est différente. Quelques mois plus tard (à 4 mois), les bébés augmentent leur temps de fixation lorsque l'expression de joie est concominante tant sur le plan visuel que auditif. Le temps de fixation est plus long lorsque l'expression vocale correspond à l'expression faciale. Cette compétence est accessible pour les bébés de 5 à 7 mois.

Cependant, toutes les études décrites, ne nous renseignent pas sur comment nous comprenons et reconnaissons les stimulus émotionnels venant d'un partenaire. Elles nous informent exclusivement sur la capacité à produire, discriminer et catégoriser. Discriminer consiste à différencier (percevoir les différences) tandis que catégoriser, c'est trouver le même à travers les différences, trouver la caractéristique qui permet de regrouper en un ensemble des éléments perçus à la fois comme différents et comme équivalent quant à cette caractéristique commune.

Ces capacités sont indispensables pour comprendre l'expression émotionnelle. Les premières expérience montrant une compréhension de l'information émotionnelle arrivent vers l'âge de 12 mois avec les aspects de référencement social (Feinman, 1982) que nous décrirons plus précisément dans la troisième partie consacrée aux interactions triadiques. Le référencement social montre la capacité du bébé à adapter son comportement en fonction de l'expressivité faciale maternelle. (Trevarthen, 1993) postule que ces données montrent la capacité du jeune enfant à ressentir et partager l'état émotionnel de l'autre.

2.3.3 L'enfant d'un an et plus

Entre 2 et 6 ans, en parallèle de l'acquisition du langage (lexique émotionnel), le jeune enfant acquiert un certain nombre de facultés qui semblent indispensables pour un développement normal de la vie sociale. Notamment, l'enfant est capable de décrire ses propres expériences émotionnelles ainsi que celles des autres. Dans cette même période, il est également capable de comprendre que sa propre expérience émotionnelle peut être différente des autres personnes. Autrement dit, il possède à présent la faculté à différencier sa propre expérience émotionnelle.

Des données indiquent que l'identification verbale de ses propres émotions et de celles d'autrui s'organise progressivement en termes hédoniques et que l'identification verbale des émotions positives précèdent les négatives (Bretherton et al., 1986). Entre 4 et 5 ans apparaît des différenciations entre les émotions négatives (Bullock and Russell, 1986). A la même période, les enfants nomment les expressions de joie, de colère, de tristesse et le visage neutre. La catégorisation sémantique est acquise vers l'âge de 5 ans (Bullock and Russell, 1984; Gosselin et al., 1995; Gross and Ballif, 1991). Cependant ce n'est qu'entre 6 et 8 ans que l'enfant nomme la surprise et le dégoût (Gosselin et al., 1995). Ces résultats montrent que l'identification verbale des expressions émotionnelles chez le jeune enfant ne se développe pas à la même vitesse pour toutes les expressions. Les positives semblent être les premières nommées alors que les négatives sont plus tardives. De plus, l'identification verbale des expressions comme la surprise est encore plus lente. Ces données montrent que la discrimination, la reconnaissance, et l'identification verbale des expressions émotionnelles se développent à des vitesses différentes mais également que les différentes expressions émotionnelles se développent à des vitesses différentes, laissant supposer des structures corticales différentes.

Quand est-il de la production des expressions faciales? La tâche de Loveland (Loveland et al., 1994) montrent que la compréhension des émotions peut être jugée vis-à-vis de la production motrice. Cette tâche est très simple, on demande aux enfants d'imiter une expression faciale qu'il perçoit ou de produire une expression que l'on dicte. Dans ce paradigme expérimental plusieurs aspects sont testés notamment la discrimination, la catégorisation, et la production d'expressions émotionnelles. Les enfants de 5 ans imitent les expressions faciales émotionnelles (Brun, 1998) et la production suivant l'évocation des émotions est acquise vers 6 ans. Les travaux de Field et Walden (Field and Walden, 1982) montrent que l'enfant exprime une émotion particulière par une simple demande verbale. Le résultat sous-jacent montre que le jeune enfant a plus de faciliter à exprimer une expression faciale lorsqu'il l'a voit que sur évocation verbale. Est-ce une question d'empathie? ou de résonance émotionnelle?

Tous les résultats énoncés jusqu'à présent dépendent exclusivement du stimulus d'entrée. Il est important de souligner que le stimulus peut être statique (photos) mais également dynamiques (situations réelles). Malheureusement, ces expériences ne mettent pas en avant le faît que le stimulus (autre personne) peut avoir des réactions et que ces réactions peuvent influencer la décision de l'enfant. Autrement dit, le contexte social peut amener l'enfant à exprimer des émotions en fonction du partenaire sociale ce qui implique que l'enfant peut être amené à prédire la réaction du partenaire. La réactivité émotionnelle peut dans ce cas être modulée par la prédiction de l'enfant. Saarni montre que l'enfant est capable d'exprimer des émotions en respectant certaines règles sociales (Saarni, 1999). L'enfant âgés de 3 à 6 ans reste souriant même s'il reçoit un cadeau non désiré. Montrant ainsi qu'il est capable de faire la différence entre les états émotionnels observables et les états émotionnels qui sont ressentis (Harris, 1989). Ces résultats soulignent que l'enfant peut décrire, produire une expression émotionnelle sans la ressentir et par conséquent qu'il est capable de se mettre à la place de l'autre.

2.4 Résonance émotionnelle

Une définition de la résonance émotionnelle serait une réponse spontanée (expression faciale) vis à vis de stimuli ayant une forte connotation émotionnelle.

2.4.1 Comment mesurer la résonance émotionnelle?

Le visage humain est composé de plusieurs muscles permettant une grande diversité des mouvements faciaux. Cependant, le réseau musculaire du visage est très complexe par conséquent il est parfois difficile de dire avec précision quels muscles sont contractés. Différentes méthodes d'analyses ont été développées permettant l'étude des mouvements faciaux pendant la production d'expressions faciales.

L'utilisation de l'EMG (électromyographie) est une méthode fréquement utilisée pour analyser la contractions des muscles. Néanmoins, la pose de plusieurs électrodes est nécessaire pour augmenter la précision. De nombreux auteurs ont utilisés cette méthode (Schwartz et al., 1976; Dimberg, 1982), analysant la résonnance motrice devant des photos de joie, de colère ... Généralement, les étrectrodes étaient placées sur les zones stratégiques par exemple le "zigomatus major" (sourire) ou "corrugator supercilii" (froncement des sourcils). Les électrodes ont des activités mêmes si la contraction n'est pas visible à l'oeil nu ce qui peut être due au bruit.

Le FACS (Facial Action Coding System) a été introduit par Ekman et Friesen (Ekman and Friesen, 1978). Il permet d'analyser l'ensemble des mouvements faciaux observables à l'oeil nu. (Ekman and Friesen, 1978) ont décidé d'utiliser les unités d'actions (UAs) comme unité de base pour décrire les mouvements faciaux. Les UAs correspondent à la contraction de 1 à plusieurs muscles. Le baby FACS est utilisé pour analyser les mouvements faciaux des jeunes enfants car ils ont des morphologies différentes vis à vis des proportions et des structures osseuses. Notons dans le même temps, Izard (Izard, 1979) crée le MAX (Maximally Discrimination Facial Movement Coding System) basé sur 68 codes représentant chacun une expression faciale.

Différentes méthodes sont à la disposition des chercheurs pour mesurer les mouvement faciaux et par conséquent la résonnance motrices que l'humain vis à vis des stimuli extérieur qu'ils soient statiques, dynamiques, biologiques ou non biologiques.

2.4.2 Différents travaux sur la résonance émotionnelle

La résonance émotionnelle est étudiée suivant que le stimulus soit de nature biologique ou non biologique. Cependant, la majorité des recherches s'intéressent davantage aux stimuli biologiques. Les études de Dimberg (Dimberg, 1982) montrent que les adultes ont cette faculté de résonance motrice émotionnelle. Les stimuli utilisés sont des expressions faciales intenses ("Pictures of facial affect" de Ekman). Des travaux récents (de Wied et al., 2006) montrent que des enfants de 10 ans expriment spontanémment des contractions musculaires devant des expressions de joie et de colère. Ces auteurs quand eux utilisaient des vidéos d'expressions faciales dynamiques et les résultats sont observés à l'aide des techniques EMG où les muscles "zygomaticus major" et "corrugator supercilii" sont respectivement plus activés lors de l'expression faciale de joie et de colère. De plus, les recherches de (Magnée et al., 2007) s'intéressent à la résonance motrice émotionnelle vis à vis des expressions de joie et de peur. Leurs résultats montrent que les sujets ont des contractions musculaires spontanées vis de ces deux expressions mais ils soulignent également que la contraction du muscle "corrugator supercilii" n'est pas spécifique à l'expression de colère.

D'autre part, la nature des stimuli peut influencer la résonance émotionnelle. Par exemple, l'intensité de l'expression perçue peut être une dimension non négligeable dans le processus. D'autres aspects doivent avoir leur importance dans le déclenchement de ce phénomène par exemple un stimulus statique ou dynamique. (Hess and Blairy, 2001) ont utilisé des expressions faciales moins intenses venant d'un enregistrement dynamiques. Ces résultats montrent malgré une plus faible intensité des stimuli et l'aspect dynamique de ce dernier (vidéo), des activités musculaires spécifiques sont détectées pour les expressions faciales correspondantes. Les résultats de (Sato and Yoshikawa, 2007) quand à eux, montrent que les réponses spontanées sont plus importantes lorsqu'on présente aux sujets des stimuli dynamiques que statiques (cette comparaison utilise le FACS).

Très peu d'études se sont intéressées à la résonance émotionnelle vis à vis de stimuli non biologiques. Nous citerons une première étude réalisé chez l'adulte où les auteurs utilisent des expressions faciales statiques et dynamiques d'un avatar. Leurs observations mettent en évidence que la résonance émotionnelle est plus prononcée lors d'une présentation dynamique que statique (Weyers et al., 2006). Enfin, nous finirons par l'étude de Simon et Nadel (Simon et al., 2007) montrant l'enfant face à des expressions robotiques et humaines. Les résultats montrent que les expressions humaines sont mieux reconnues que les expressions robotiques et les performances s'améliorent avec l'âge. Cette étude montre que les enfants de 3 à 5 ans résonent

émotionnellement devant un partenaire non humain mais expressif (i.e une tête de robot similaire à celle que j'utiliserais dans mes expériences).

2.5 Interactions triadiques: le référencement social

Les stimuli émotionnels sont en permanence utilisés pour permettre une adaptation à l'environnement social. Selon Darwin, les émotions possèdent une valeur adaptative qui est d'origine phylogénétique (Darwin, 1965). L'idée sous-jacente vient de ses propres travaux sur l'évolution biologique des espèces pour comprendre les expressions émotionnelles ainsi le principe de théorie de l'évolution peut être appliqué pour les expressions émotionnelles. Ceci pourrait expliquer l'universalité des expressions émotionnelles et leur présence chez tous les hommes, indépendamment de leur culture. Ces travaux soulignent que les émotions jouent un rôle important chez beaucoup d'espéces pour communiquer des états émotionnels et que l'évolution les auraient façonnées.

Un grand nombre d'études en psychologies (Campos, 1983; Campos et al., 1989; Izard, 1990; Klinnert et al., 1983b) mettent en évidence que les expressions émotionnelles sont essentielles aux relations avec l'environnement physique et social. Elles constituent une forme de communication non-verbale, permettant de communiquer ses émotions mais également de comprendre celles des autres. Les émotions jouent un rôle prédominant dans l'adaptation de l'individu à son milieu social.

Wallon est l'un des premiers à proposer l'idée que les émotions sont régulatrices du comportement (Wallon, 1987). Elles ont une fonction de régulation posturale. Dans sa théorie du développement, le stade émotionnel (3 à 9 mois) est le deuxième stade où les gestes deviennent utiles et expriment des besoins. Les émotions deviennent organisatrices du comportement de l'enfant.

L'émotion est un processus permettant l'adaptation des organismes résultant par exemple de la compréhension d'une expression émotionnelle. Ceci laisse supposer la capacité à discriminer et catégoriser les émotions exprimées par un visage (Nelson and Dolgin, 1985; Serrano et al., 1992; Kestenbaum and Nelson, 1990; Farroni et al.,) ou la compréhension sémantique des labels émotionnels (Bullock and Russell, 1986).

2.5.1 Les émotions un moyen d'adapter le comportement

A l'âge de 1 an environ, le bébé montre des aptitudes reflétant la compréhension des émotions transmises par un autre. Cette faculté se nomme le référencement social et se définit comme un processus permettant à un individu d'utiliser les informations émotionnelles fournis par d'autres individus lors d'une situation ambigue et ceci dans le but d'adapter son propre comportement à cette nouvelle situation (Campos and Steinberg, 1981; Feinman, 1982).

D'un point de vue développemental, le jeune enfant acquiert très tôt des capacités émotionnelles allant d'une discrimination des expressions faciales émotionnelles jusqu'à la catégorisation sémantique ou bien même encore la perception intermodale de stimuli émotionnels. Ces facultés ont plusieurs finalités mais l'une des plus importantes doit sans aucun doute être la capacité à moduler son comportement vis à vis d'une nouvelle situation. Le référencement social est composé de plusieurs capacités, supposant toujours une interaction avec un autre individu:

- l'orientation du regard entre la source d'information et le stimulus. La source d'information peut correspondre à un parent et le stimulus à un objet.
- des réactions motrices: comportement d'approche et de recul par exemple.

– des modifications émotionnelles de l'individu cherchant l'information (modification de son expression faciale)

Le référencement social doit être l'une des premières manifestations d'un décodage explicite des émotions puisqu'il est défini comme étant le résultat de l'assimilation d'une information émotionnelle d'un autre individu dans le but de faciliter sa propre adaptation comportementale. Les capacités acquises durant les 6 premiers mois de vie vont être utilisées dans le référencement social (Nelson, 1987) (expressions faciales, sonores ...).

Dans les expériences construites par Klinnert (Klinnert, 1984), on présente à des bébés agés de 12 et 18 mois des jouets ayant une connotation négative comme par exemple une araignée. On demande simultanément aux mères d'exprimer de la joie, de la peur ou un visage neutre. Les observations montrent que les bébés sont capables de discriminer les expressions faciales émotionnelles car le comportement des bébés est corrélé avec l'expression produite par la mère. Dans le cas où la mère exprime de la joie, le bébé explore davantage les objets. Cependant, lorsque la mère produit une expression de peur, le bébé adapte son comportement en évitant l'objet et en se rapprochant de la mère. (Gunnar and Stone, 1984) font les mêmes observations. Ils montrent que les bébés de 1 an ont plus de réactions positives face à un robot mécanique lorsque la mère exprime des visages souriants. (Sigman and Kasari, 1994) utilisent eux aussi le robot comme stimulus ambigu. Les observations montrent que des enfants de 8 mois à 2 ans ont tendance à se référer davantage à l'adulte face à l'arrivée d'un robot bruyant ainsi qu'à la production de la peur par l'adulte (expression faciale, sonore, gestuelle). D'autres auteurs essayent de généraliser ce concept en montrant que d'une part l'expression faciale n'est pas le seul moyen de communiquer avec le bébé et que d'autre part l'information émotionnelle transmise peut désambiguiser non seulement un objet inquiétant mais également une situation contextuelle nouvelle. Par exemple, (Feinman and Lewis, 1983) montrent que des bébés de 10 mois acceptent davantage un individu étranger dans la mesure où le discours maternel renseignant sur cet individu a une tonalité joyeuse. Toutes ces expériences soulignent combien le décodage émotionnel est en lien direct avec l'adaptation du comportement du bébé. (Walden and Ogan, 1988) mettent même en évidence que l'information émotionnelle est préférée à d'autres sources d'informations telles que les contacts physiques. L'attirance pour la dimension émotionnelle augmente notamment pour les émotions négatives des jeunes enfants agés de 6 mois à environ 2 ans.

Fig. 2-4 – Référencement social: falaise visuelle.

L'expérience de Sorce (Sorce et al., 1985) montre la réaction d'un enfant face à une falaise visuelle (la figure 2-4 montre une illustration de cette expérience), créée grâce à un dispositif expérimental original. La falaise visuelle à la taille d'une table dont la première partie est opaque et la seconde est vitrée. Cette expérience est idéale pour que l'enfant puisse se mouvoir dans un environnement parfois ambigu. Le but de l'expérience est de comprendre si le comportement de l'enfant s'adapte aux informations émotionnelles transmises par la mère. Lorsque l'expérience débute, les mères sont placées du côté vitré de la falaise tandis que leur enfant se situe du coté opaque de la falaise. Les mères ont pour objectif d'inciter les enfants à se rapprocher ou à éviter le vide selon leurs expressions émotionnelles. Au moment où l'enfant se trouve confronter au vide, la mère envoie des informations émotionnelles en exprimant de la joie, de la peur, de l'intérêt, de la colère ou de la tristesse. Les résultats montrent que si les signaux émotionnels transmis par la mère sont positifs, par exemple l'expression de joie, alors 14 enfants sur 19 traversent la falaise visuelle et 11 enfants sur 15 traversent lors que les mères expriment de l'intérêt. Cependant, aucun des enfants ne traversent la falaise lors que les mères expriment de la peur, 2 enfants sur 18 traversent lorsque les mères expriment de la colère et 6 enfants sur 18 traversent lorsque la mère exprime de la tristesse. Ces résultats reflètent que l'enfant possède la capacité à discriminer les émotions positives et négatives par le simple fait qu'il soit capable d'adapter des comportements spécifiques en fonction des informations émotionnelles qui lui sont émis par la mère.

(Sorce and Emde, 1981) construisent une expérience montrant que la présence active de la mère influence le comportement d'exploration et de jeu chez des enfants agés de 15 mois. Dans cette étude, les auteurs ont à leur disposition 40 couples mère-enfant qui sont scindés en deux groupes. Le premier groupe dispose des mères dites disponibles, c'est à dire ayant le droit d'interagir avec leur enfant par l'intermédiaire de l'expressivité émotionnelle et d'un second groupe contenant des mères dites non disponibles, autrement dit des mères ne changeant pas leur activité malgré les sollicitations de leur enfant. Dans cette expérience, le paradigme expérimental est le suivant: les mères non disponibles lisent un magazine durant 15 minutes et ne répondent pas aux demandes des enfants. Un petit robot ainsi qu'un individu étranger rajoute une certaine ambiguité à la situation. Au début de l'expérience, le bébé joue au centre de la pièce en toute liberté alors que la mère non disponible et l'individu étranger se situent chacun dans un coin de la pièce. Dans un second temps, l'individu étranger essaie d'interagir avec l'enfant, puis un petit robot mécanique entre dans la scène. L'expérience reste identique pour les mères disponibles excepté qu'elles ne lisent pas de magazine et que contrairement aux mères non-disponibles elles ont l'autorisation de communiquer par l'intermédiaire d'expressions émotionnelles. L'objectif de cette expérience est de montrer la réaction de l'enfant lors d'une situation ambigue. Les auteurs comparent les résultats expérimentaux d'un groupe d'enfants recevant des signaux émotionnels de leur mère versus un groupe d'enfants dont les mères n'interagissent pas avec eux. Cette expérience montre que l'enfant adopte des comportements différents suivant les deux situations:

- Le comportement émotionnel: les enfants sont moins joyeux dans le cas où la mère n'est pas disponible.
- Le comportement d'exploration (recherche de signaux): à partir du moment où la mère est activement présente alors l'enfant cherche à s'adapter à la situation. L'exploration est davantage présente lorsque la mère est disponible.
- Le comportement de l'enfant est fonction de l'expressivité émotionnelle de la mère vis à vis de la nouvelle situation.

2.5.2 La mère n'est pas la seule source d'information

Certains auteurs comme Dickstein et Parke (1988) (Dickstein and Parke, 1988) montrent que les bébés agés de 11 mois utilisent comme source de référencement le père aussi bien que la mère. L'étude souligne qu'ils se réfèrent aux expressions émotionnelles de leur père lors de l'apparition d'une personne étrangère. Cette étude montre qu'il n'y a pas de différence entre la source d'information mère ou père.

De même (Hirshberg and Svejda, 1990) montrent que la mère aussi bien que le père servent de source de référencement en observant des triades père, mère et enfant agé de 12 mois. L'expérience est très simple: des stimuli sonores et menaçants sont montrés aux bébés comme par exemple une peluche ou un robot menaçant tandis que les parents envoient des informations émotionnelles. Notons que les parents peuvent produire exclusivement les expressions faciales de joie, de colère et de peur. L'étude montre que les informations émotionnelles sont d'une part décodées correctement et utilisées par l'enfant quelque soit le parent.

Une étude (Klinnert et al., 1986) montre que des bébés d'environ 1 an sont capables de se référer à un individu familier (personne étrangére) en présence de la mère. La mère durant toute la durée de l'expérience reste neutre tandis que les expérimentateurs peuvent exprimer soit un visage de joie soit un visage de peur suivant la situation. L'enfant et la personne étrangère jouent pendant quelques minutes ensemble, un jouet se déplaçant en direction de l'enfant apporte un contexte ambigu. Dans cette situation, l'expérimentateur émet des signaux émotionnels (expressions de joie ou de peur) pouvant influencer le comportement de l'enfant. Les résultats montrent que les jeunes enfants tendent à utiliser l'expressivité de l'expérimentateur pour moduler leur propre comportement (83% des enfants se réfèrent à l'expérimentateur). La présence de la mère n'influence pas l'enfant car elle n'affiche qu'un visage neutre et de plus l'enfant se réfère davantage à l'expérimentateur étranger qu'à la mère. Les résultats montrent que les signaux expressifs sont correctement utilisés par les enfants. Les enfants se rapprochent, touchent et manipulent plus longtemps le robot lorsque l'expérimentateur exprime de la joie que lorsque l'expérimentateur affiche de la peur. Par conséquent, cette étude met en évidence plusieurs éléments qui semblent importants:

- Les enfants sont capables d'utiliser les signaux émotionnels émis par des humains étrangers.
- La présence passive de la mère n'influence pas le comportement de l'enfant.
- Le comportement de l'enfant vis à vis d'un objet s'adapte en fonction de l'émotion exprimée par un adulte même si celui ci n'est pas un proche.

2.5.3 L'attention conjointe comme moyen de communication non émotionnelle

Un des autres aspects du référencement social qui joue un grand rôle, est sans doute la capacité du nourrisson à regarder là où l'autre regarde. Cette compétence montre clairement une forme de prémice de contact social qui doit certainement être à la base d'autres formes de coordination sociale. Certaines études montrent l'importance du regard chez le nourrisson. Par exemple (Murray and Trevarthen, 1985) montrent que les enfants de 2 mois réagissent négativement lors d'une présentation d'un film où le regard et les vocalisations de leur mère ne sont pas coordonnées avec les leurs. Ces résultats mettent en évidence que les regards mutuels sont une dimension importante pour le maintient d'une interaction harmonieuse entre le nourrisson et la mère. Les regards mutuels sont un mécanisme primordial notamment pour le maintient du contact visuel. L'attention mutuelle est un échange bi-directionnel entre 2 agents (diade) permettant le maintien d'une interaction. L'attention conjointe quand à elle, permet

une interaction triadique (enfant-adulte-objet), dont l'objectif est d'orienter le regard de l'enfant en fonction du regard de l'autre en s'attardant sur le stimulus que l'autre est entrain de regarder. Trevarthen postule que dès les premiers mois, l'enfant possède des prédispositions vis à vis de l'attention mutuelle et que l'attention conjointe est une capacité innée (Trevarthen, 1993). Cependant d'autres auteurs comme (Tomasello et al., 2005) défendent l'idée que l'attention conjointe est l'aboutissement d'un apprentissage social. Nous proposerons dans les chapitres suivants un modèle allant dans ce sens.

A partir de 6 mois, l'enfant commence à suivre le regard de la mère (Butterworth, 1991). Entre 6 et 12 mois, l'enfant suit le regard de l'adulte qui porte son attention vers un objet (attention conjointe). Les enfants commencent par suivre le regard de l'autre aboutissant sur la cible que l'adulte regarde. L'adulte est une source d'information, son regard est un point de passage pour la découverte d'objets potentiellement intéressants. Une fois encore, l'adulte est à la disposition de l'enfant pour lui faire partager son expérience. D'après (Conein, 1998), les

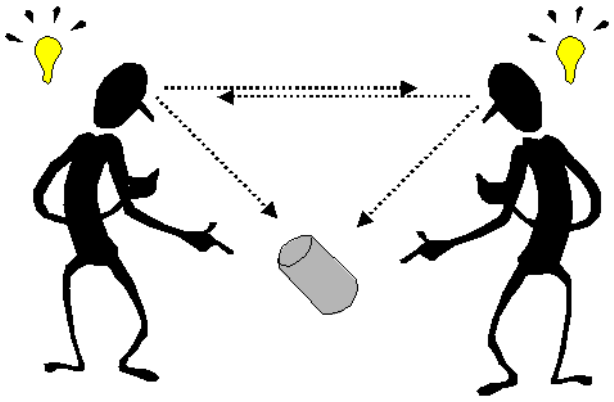


Fig. 2-5 – Attention conjointe.

phases de développement de l'attention conjointe sont les suivantes:

- à 2 mois: l'attention mutuelle permet à l'enfant d'ajuster son regard avec le regard de l'autre pour maintenir une expérience partagée.
- à 6 mois: les prémices de l'attention conjointe sont présents, l'enfant peut suivre le regard de l'autre.
- à 12 mois: l'attention conjointe est possible et permet à l'enfant de suivre le foyer d'attention de l'autre et sa focalisation (le stimulus).
- à 18 mois: l'attention contrôlée est acquise. Elle permet à l'enfant de controler la direction du regard de l'autre au moyen de son propre regard en l'orientant vers une cible (manipuler l'autre).

2.6 Conclusion

Les différents modèles théoriques des émotions ont tous de nombreux intêrets. Pour nous roboticiens, il semble clair que chacun des modèles possède des caractéritiques que nous aimerions introduire pour construire un robot doté "d'émotions". Cependant, nous pensons qu'en partant d'expériences robotiques élémentaires, nous construirons des modèles émotionnels minimaux, qui pourront discuter de l'aspect de certains paramètres (ou composantes) des différents modèles.

De notre point de vue, les expériences robotiques que nous réalisons, permettent d'avoir une nouvelle vision sur les émotions.

Les gens se posent souvent la question: est-ce qu'un robot peut "ressentir" des émotions? La réponse à cette question est délicate mais si elle est positive comment le juger? Forcément, le comportement du robot doit nous faire ressentir, le fait qu'il "ressent" lui-même des émotions. Une interaction avec le monde physique, social et ses propres états internes induit des comportements qui pourront nous montrer la compréhension de certains états émotionnels vis à vis de certains stimuli (figure 2-6). Les émotions induises des actions motrices afin de s'adapter à l'environnement.

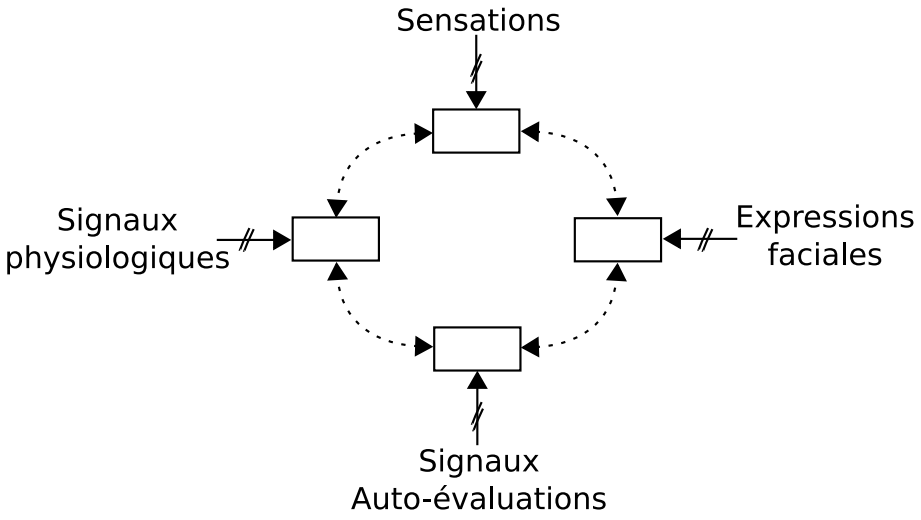


Fig. 2-6 – Les émotions sont vues comme des interactions entre les états internes, l'environnement physique et l'environnement social.

Le référencement social est défini par les psychologues comme la capacité à utiliser des messages venant d'un expérimentateur extérieur notamment des messages émotionnels dans certaines situations ambigues. Cependant tous les auteurs ne s'entendent pas sur cette définition car certains affirment que l'individu doit chercher l'information (Rosen et al., 1992) tandis que d'autres comme (Campos, 1983; Walden and Baxter, 1989) ajoutent que l'individu doit soit chercher l'information soit l'information lui est imposée. Les modèles que nous allons développer dans les chapitres suivants ne discutent pas sur la définition du référencement social. Ils montrent simplement qu'il est possible de réaliser cette tâche en utilisant des messages émotionnels expressifs. Dans ce cadre, l'expression faciale est un excellent de commnunication non-verbale permettant la régulation du comportement.

Dans la littérature, l'âge de développement de cette capacité est souvent discuté. (Feinman and Lewis, 1983; Repacholi, 1998) estiment que cette faculté émerge vers 10 mois tandis que la majorité des études indiquent que cette capacité est présente à environ 1 an (Baldwin and L.J., 1996; Mumme et al., 1996). Ces études nous montrent très clairement que cette capacité n'est pas innée chez l'enfant mais qu'elle se développe. L'enfant doit acquérir plusieurs capacités avant de pouvoir adapter son comportement selon les messages émotionnels que lui envoie l'expérimentateur. Le référencement social relie plusieurs composantes fondamentales:

- La capacité à analyser la situation sociale: nouvelle situation, situation ambigue, ...
- La capacité à décoder l'information sociale: l'enfant agé d'un an doit être capable de

discriminer et de catégoriser les expressions émotionnelles aussi bien de la mère que celles d'un individu étranger.

– La capacité à réguler son comportement: utiliser l'information sociale dans le but de produire un comportement approprié.

Les chercheurs sont unanimes pour dire que le référencement social est une fonction de régulation du comportement (Klinnert et al., 1983a; Walden and Baxter, 1989; Feinman et al., 1992; Rosen et al., 1992; Mumme et al., 1996; Baldwin and L.J., 1996). La régulation du comportement est spécifique à la situation ou au contexte environnemental dans lequel l'enfant est immergé. Des recherches ont été réalisées montrant que le référencement social et la modification de l'humeur sont des phénomènes pouvant être séparés. Par exemple, (Stenberg and Hagekull, 1997) étudient des mères utilisant des signaux émotionnels lors de la présence d'un nouvel objet. Les résultats soulignent que le comportement des enfants change vis à vis de l'objet alors que l'humeur globale de l'enfant n'est pas affectée. Une étude de Hornick en 1987 (Hornik et al., 1987) montre également qu'en présence d'une mère effectuant une expression négative vis à vis d'un nouveau jouet, les enfants continuent à jouer avec tous les jouets qui sont dans la chambre excepté le nouveau jouet. Cette étude montre que l'humeur globale des enfants n'a pas été affectée par l'expression de la mère et que les enfants ont compris que le message émotionnel était spécifique à cet objet particulier. Nous soutenons l'idée que le référencement social sert de fonction de régulation des comportements spécifiques à une situation ou un contexte donné.

Le dernier point que nous désirons discuter est la notion de précablage biologique du référencement social. Certains comme (Campos, 1983) soutiennent que ce processus est précablé biologiquement tandis que d'autres auteurs comme (Gewirtz and Pelaez-Nogueras, 1992) mettent en évidence que le référencement social peut émerger d'un apprentissage opérant. (Gewirtz and Pelaez-Nogueras, 1992) ont mis en évidence que la référenciation sociale est un processus pouvant être appris. Par l'intermédiaire de l'interaction, cette tâche peut émerger d'un apprentissage opérant. Ces auteurs semblent dire que ce processus n'est pas le simple résultat d'une association avec les expressions émotionnelles, les gestes corporelles ou encore des sons. Mais au contraire, ils soutiennent l'idée que le référencement social peut être le résultat d'une association entre un motif visuel, par exemple mettre la main devant le visage pour attribuer un signal positif ou par opposition mettre les mains autour du visage pour assigner un signal négatif. Dans leur expérience, le jeune enfant est capable de prendre en compte des signaux non triviaux et de les utiliser pour la régulation de son comportment. Cette étude est relativement intéressante à plusieurs niveaux. Dans un premier temps, elle modifie légèrement la définition du référencement social car dans la littérature le message envoyé à l'enfant doit être émotionnel. Cependant, on peut penser que les signaux envoyés sont toujours de nature émotionnelle car ils ont été conditionnés. Dans un second temps, l'expérience montre que ce processus n'est pas précablé car d'autres signaux peuvent conditionner le comportement de l'enfant. Et enfin, l'enfant est capable d'apprendre à réagir à certains stimuli comme des expressions particulières du visage, montrant ainsi l'immense plasticité du comportement de l'enfant vis à vis de son environnement.

Chapitre 3

Réseaux de neurones et formalisme des systèmes cognitifs

Ce chapitre présente différents "outils" utiles pour le développement d'architecture neuronale. Ces outils se situent à plusieurs niveaux: neurone formel, type de réseau de neurone et architectures neuronales. Nous présenterons dans un premier temps les neurones formels ainsi que différents réseaux de neurones sans oublier des architectures neuronal plus complexes. Enfin, nous présenterons un formalisme pour les systèmes cognitives qui a été développé pour décrire simplement les architectures de contrôle.

3.1 Les réseaux de neurones

Dans notre cadre théorique, l'approche connexionniste semble avoir des propriétés intéressantes pour le développement d'architecture de contrôle complexe. L'utilisation de réseaux de neurones présente un certain nombre d'intérêts qui semblent adéquates pour la robustesse des modèles développées. Les avantages sont multiples: l'adaptabilité à des environnements complexes (naturel) et les non linéarités introduites garantissent une robustesse au bruit de l'environnement (Rumelhart and al. (D.E), 1986) qui sont des propriétés émergentes aux réseaux de neurones. Un intérêt grandissant concerne le coût computationnel, les architectures deviennent de plus en plus complexes est le besoin d'avoir un calcul distribué se fait sentir (Lagarde et al., 2008). Les calculs étant locaux, la répartition des architectures neuronales sur différentes machines est faisable. La réalisation et la compréhension d'une architecture de contrôle a pour conséquence une compréhension détaillée des relations à établir entre les différents éléments constituants l'architecture ceci étant dû à l'aspect local des réseaux de neurones.

3.1.1 Neurones biologiques et modèles

Le système nerveux compte plus de 100 milliards de neurones interconnectés qui transmettent de l'information sous forme de potentiel électrique et de codage chimique. Leurs formes et certaines caractéristiques peuvent varier d'un neurone à l'autre suivant leur type. En 1952, Hodgkin et Huxley (Hodgkin and Huxley, 1952) ont proposé un modèle biophysique de la génération du potentiel d'action dans l'axone du calmar, qui est la base de la plupart des modèles biophysiques actuels. Le neurone est composé:

– un corps cellulaire contenant le noyau du neurone ainsi que la machine biochimique

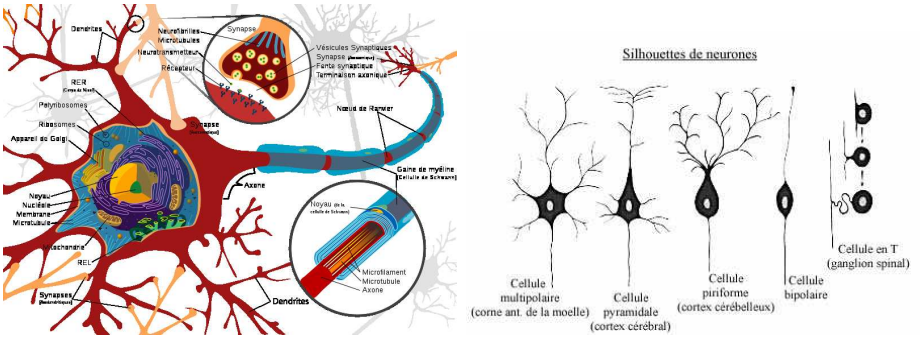


Fig. 3-1 – Schéma d'un neurone biologique.

nécessaire à la synthèse des enzymes. Ce corps cellulaire est de forme sphérique ou pyramidale contenant les molécules nécessaires au développement de la cellule.

- les dendrites sont des fines extensions tubulaires se ramifiant autour du neurone pour former une vaste arborescence. Les signaux envoyés au neurone sont réceptionnés par les dendrites.
- l'axone est plus long que les dendrites, se ramifiant autour de son extrémité pour se connecter aux dendrites des autres neurones. La fonction de l'axone est de réceptionner le signal venant d'autres neurones. La forme de l'arbre dendritique a une influence sur la fonction du neurone (Yuste and Tank, 1996)
- La synapse est l'interconnexion entre deux neurones (émetteur et récepteur), ils forment la connexion entre l'axone de l'émetteur et la dendrite du récepteur. Notons qu'il peut exister des jonctions axo-axonales et des synapses électriques.

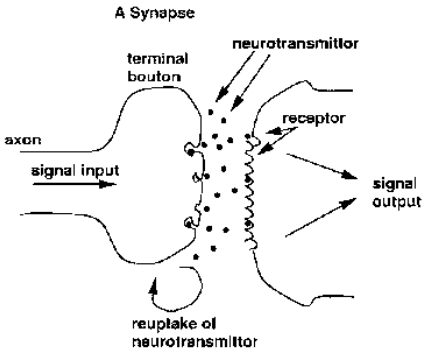


Fig. 3-2 – Schéma d'une synapse

Du point de vue fonctionnel, le neurone doit être regardé comme une entité polarisée car l'information est transmise dans un seul sens des dendrites vers l'axone. Le neurone reçoit par l'intermédiaire de ses dendrites des informations venant d'autres neurones. Toutes ces informations sont sommées au niveau du corps cellulaire. Un potentiel d'action se forme (signal électrique) pour être ensuite transmis le long de l'axone jusqu'aux terminaisons synaptiques. L'influx nerveux (l'information) lors de son arrivée, va libérer des neurotransmetteurs (médiateurs chimiques) dans la fente synaptiques grâce aux vésicules synaptiques fusionnant avec la mem-

brane cellulaire. Les neurotransmetteurs rendent possible le passage du flux nerveux entre les neurones car le signal électrique ne peut pas passer la synapse dans le cas précis d'une synapse chimique. Il existe deux types de neurotransmetteurs (excitateur ou inhibiteur), ils excitent ou inhibent le neurone auquel ils sont connectés et par conséquent peuvent autoriser ou interdire la propagation de l'information à d'autres neurones. Sur la membrane dendritique au niveau post-synaptique se trouvent des récepteurs pour les neurotransmetteurs. En fonction du type de neurotransmetteur et du type de récepteurs, l'activité du neurone (excitation) augmente ou diminue pour propager ou non l'information. Les synapses possèdent une "mémoire" permettant d'ajuster leur fonctionnement, par exemple la co-activation répétée entre deux neurones modifie les connexions synaptiques qui les relient (règle de Hebb (Hebb, 1949)). Ces connexions se modifient au fur et à mesure des flux nerveux pour faciliter ou non leur passage. Cette plasticité est à l'origine des mécanismes d'apprentissage. Le modification synaptique peut intervenir sous trois formes principales:

- augmentation ou diminution du nombre de récepteurs sur le neurone post-synaptique.
- augmentation ou diminution du nombre de neurotransmetteurs émis.
- augmentation ou diminution de l'efficacité d'un neurotransmetteur pour la génération d'un potentiel d'action.

L'étude du neurone biologique a permis le développement de modèles formels. La connaissance approfondie des principes fonctionnels du neurone ainsi que ces propriétés permettent des stimulations computationnelles (Rumelhart and al. (D.E), 1986). Mc. Culloch & Pitts (McCulloch and Pitts, 1943) ont développé un modèle de neurone formel que nous allons considérer pour les architectures de contrôle. Ce modèle ne tient pas compte de l'aspect temporel des décharges du neurone, il code simplement une activité analogique comprise dans l'intervalle $[0, 1]$. Le neurone formelle de Mc. Culloch & Pitts se composent de :

- un vecteur de connexions W définissant les poids des synapses reliant les neurones d'entrées et le neurone considéré. W est l'efficacité des connexions synaptiques.
- un potentiel d'action Pot avec $Pot = W^T X$ où X est le vecteur d'entrée.
- un seuil θ permettant une réponse du neurone si le Pot est suffisamment élevé.
- une fonction f dit fonction d'activation permettant le calcul de l'activité de sortie du neurone $s = f(Pot - \theta)$

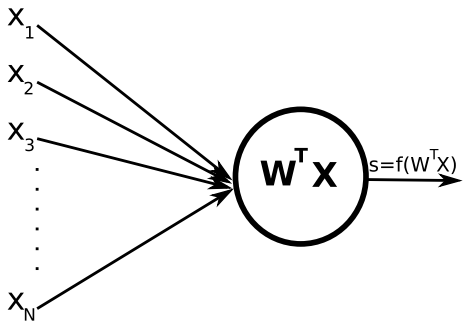


Fig. 3-3 – Schéma d'un neurone formel. Neurone formel de Mc. Culloch et Pitts

Pot étant le produit entre les entrées et les poids peut s'écrire sous la forme:

$$Pot = \sum_i w_i x_i \tag{3.1}$$

Le potentiel d'action Pot est un simple produit scalaire pouvant être vue comme une mesure de similarité si les vecteurs sont normés. Les poids W pouvant variés au cours du temps, cette variation des poids est à l'origine de la capacité de plasticité synaptique. Cette plasticité est modélisée par une loi de variation également appelée loi d'apprentissage. La fonction d'activation f introduit des non linéarités permettant l'utilisation des réseaux de neurones pour l'apprentissage de classes (catégories). Cette fonction a comme propriété de limiter la valeur du Pot (valeur minimale et maximale) permettant une stabilité des poids synaptique ainsi qu'une convergence pour l'apprentissage des catégories.

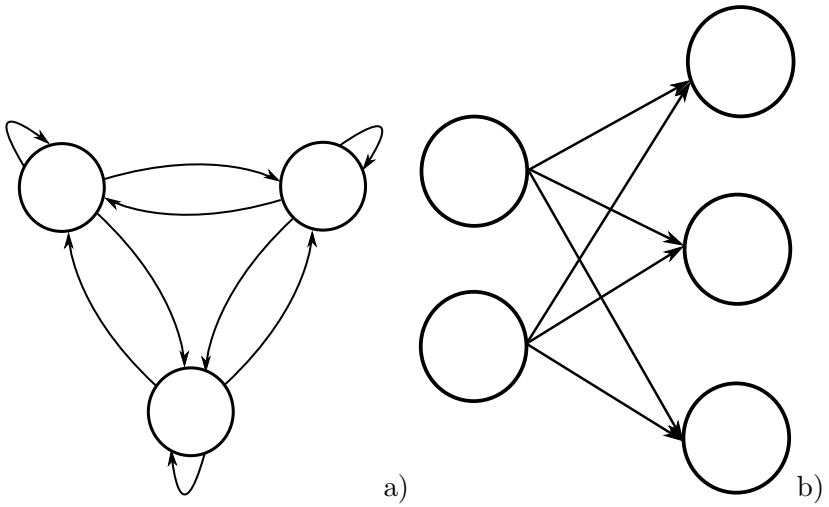


Fig. 3-4 – Schéma de 2 réseaux de neurones différents: a) montre un réseau de neurone complètement connecté. b) montre un réseau de neurones à 2 couches de type feedforward.

Le modèle de Mc. Culloch & Pitts (McCulloch and Pitts, 1943) est le modèle que nous utiliserons dans la suite de nos développements. Il est compatible avec des architectures de contrôle temps réelles constituées de grandes populations de neurones (quelques milliers). Contrairement aux modèles de neurones impulsionnels introduit par Lapicque (Lapicque, 1907) qui eux sont des modèles plus précis prenant en compte la modélisation de l'aspect temporel des décharges. Leur intérêt est grandissant car leurs simulations peuvent être comparées à des enregistrements neurobiologiques fins.

Le modèle formel Mc. Culloch & Pitts (McCulloch and Pitts, 1943) décrit sur la figure 3-3 reste un modèle cohérent pour l'étude de système cognitif ainsi que ces interactions avec des environnements complexes. L'interconnexion entre plusieurs neurones formels décrit un réseau de neurone. Nous allons décrire un certain nombre de réseau de neurones se basant sur la notion de neurone formel. Cette sélection est non exhaustive montrant les principaux modèles existants sur lesquels se basent nos développements.

3.1.2 Réseaux de neurones non supervisés

3.1.2.1 Règle de Hebb

La règle d'apprentissage de Hebb montre que la co-activation répétée d'un neurone pré-synaptique et d'un neurone post-synaptique lors de stimulation produit des modifications synaptiques faisant croître l'efficacité de la transmission. Cette d'apprentissage est d'inspiration biologique (Hebb, 1949). La figure 3-5 illustre le fonctionnement de la règle d'apprentissage de

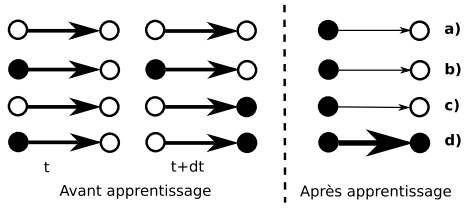


Fig. 3-5 – Principe de renforcement des poids pour la règle de Hebb. Différents cas sont montrés: a) montre aucune co-activation pré-synaptique et post-synaptique avant l'apprentissage n'impliquant pas d'activation post-synaptique lors d'un stimulus pré-synaptique. b) montre une activation pré-synaptique mais aucune activation post-synaptique avant l'apprentissage impliquant après apprentissage aucune activation post-synaptique lors d'un stimulus pré-synaptique. c) montre une activation post-synaptique mais aucune activation pré-synaptique avant l'apprentissage impliquant après apprentissage aucune activation post-synaptique lors d'un stimulus pré-synaptique. d) montre une co-activation pré et post synaptique avant l'apprentissage impliquant une activation post-synaptique lors d'un stimulus pré-synaptique. La largeur de la flèche représente le renforcement des poids synaptiques (flèche fine correspond à une diminution du poid, trait épais correspond à une augmentation du poid).

la règle de Hebb suivante:

$$\delta w_{ij} = \epsilon x_i yi - \lambda w_{ij} - \lambda^{'} w_{ij} x_i \tag{3.2}$$

w_{ij} étant les poids synaptique en 2 neurones i et j ayant respectivement x_i et y_j comme activité, les termes λw_{ij} et $\lambda^{'} w_{ij} x_i$ correspondent respectivement à un oubli passif et un oubli actif. La règle de Hebb correspond à un apprentissage associatif entre un neurone d'entrée et de sortie. Les poids synaptiques augmentent lorsqu'il y a une co-activation entre le neurone post-synaptique x_i et y_j. Elle permet d'apprendre des corrélations et peut être vue également comme l'apprentissage d'un "OU" logique entre les différentes entrées. Cependant la règle 3.2 reste instable car les poids $w_i j$ peuvent augmenter indéfiniment par conséquent des mécanismes sont introduits pour palier à ce problème par exemple l'oubli passif ou la normalisation des poids synaptiques. La figure 3-5 montre qu'une co-activation des neurones pendant la phase d'apprentissage induit un renforcement des poids synaptiques. Ainsi après apprentissage l'activité du neurone pré-synaptique déclenche une activité du neurone post-synaptique. Dans les autres cas traités, les neurones ne sont pas co-activités en même temps par conséquent les poids n'évoluent pas n'impliquant après apprentissage aucune activité du neurone post-synaptique lors de l'activation du neurone pré-synaptique.

3.1.2.2 Winner Takes all

Le WTA (Winner Takes All) simule les mécanismes de compétition existant entre neurones ou populations de neurones. Le modèle général est constitué de neurones formels interconnectés dont l'apprentissage peut être fixé par la règle de Hebb. Des liaisons inhibitrices latérales sont ajoutées permettant de simuler le processus de compétition (Figure 3-6). Le neurone vainqueur

est celui ayant la plus forte activité, tous les autres neurones sont inhibés après convergence (Rumelhart and Zipser, 1985; Lippman, 1987). Par exemple, le réseau MAXNET a ces propriétés. Ce type de réseau de neurone catégorise les vecteurs présentés en entrée. La répétition des formes (présentation des catégories un nombre suffisant de fois) permet une sensibilité de chaque neurone composant le réseau. Cependant, un WTA permet de classifier correctement des données entrantes si l'orthogonalisation de ces données reste simple.

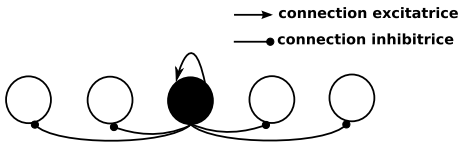


Fig. 3-6 – Schéma représentant un mécanisme de compétition sur un réseau à une couche de type WTA.

Le modèle Instar de Grossberg utilise un mécanisme de compétition sur deux couches (Grossberg, 1973) correspondant à un Winner Take All modifié. L'instar est un mécanisme de réhaussement de contraste aidant à normaliser des données d'entrées (image).

3.1.2.3 Carte de Kohonen

Grâce à certaines expériences de neurobiologie, on a pu observer dans certaines aires corticales que des colonnes voisines ont tendance à répondre à des entrées proches. Par exemple, dans les aires visuelles, deux colonnes proches sont en correspondance avec deux cellules proches de la rétine (Hubel and Wiesel, 1977). Ces résultats ont pu être observés aussei dans le bulbe olfactif ou l'appareil auditif (Knudsen and Konishi, 1979). Ces observations ont abouti à la proposition de modèles de carte topologique auto-adaptative permettant le codage de motif tout en respectant la topologie de l'espace d'entrée. Les cartes Kohonen (Kohonen, 1989a) sont des réseaux permettant ce type d'apprentissage non supervisé en respectant la topologie des données entrantes. Des entrées voisines sont codées par des neurones voisins. Des réseaux $1D$ ou $2D$ sont souvent utilisés cependant la topologie des données d'entrées peut être rompue à cause d'une mauvaise convergence. Ces réseaux permettent tout de même une quantification de l'espace d'entrée. Les cartes de Kohonen suivent le principe suivant:

- Un vecteur X est présenté et les activités de l'ensemble du réseau sont calculées
- Le neurone ayant la plus petite distance au vecteur d'entrée sera sélectionné après compétition
- Le neurone vainqueur ainsi que ses voisins voient leurs poids synaptiques se modifier suivant la règle d'apprentissage suivant: $w_j(t+1) = w_j(t) + \mu(t,j)(x_i(t) - w_i(t))$

μ est la vitesse d'apprentissage dépendant du temps et de la position des neurones sur la carte, définissant ainsi la taille du voisinage d'interaction du neurone gagnant. Le voisinage est défini généralement par une fonction ayant la forme d'une DOG (Différence de Gaussienne). Ce type de compétition mettant en interaction les neurones d'un voisinage permet de regrouper les motifs similaires et d'éloigner les neurones codant pour des motifs lointains grâce aux mécanismes d'inhibition. La taille du voisinage ainsi que la valeur μ diminue au fur et à mesure de l'apprentissage, le but étant de stabiliser le réseau.

3.1.3 Réseaux de neurones supervisés

Le contrôle de l'apprentissage est effectué par l'intermédiaire d'une supervision extérieur

3.1.3.1 LVQ

Les LVQ (Learning Vector Quantification) est un modèle de réseau de neurone provenant des cartes de kohonen. Ces réseaux permettent une catégorisation des entrées en fonction d'une supervision extérieure. Les LVQ cherchent la meilleure frontière de séparation des classes alors que les cartes de kohonen minimisent une erreur de quantification. Le réseau LVQ effectue un apprentissage suivant la règle suivante:

$$\begin{cases} w_q(t+1) = w_q(t+1) + \epsilon.(X - W) & \text{si le neurone q code la classe X} \\ w_q(t+1) = w_q(t+1) - \epsilon.(X - W) & sinon \end{cases}$$

Il existe plusieurs type de LVQ, les LVQ1 et LVQ2 introduisent des raffinements en fonction de conditions supplémentaires comme par exemple la sélection de deux neurones vainqueurs pour se rapprocher de la frontière optimal.

3.1.3.2 Perceptron et règle des moindres carrés

Rosenblatt (Rosenblatt, 1958) invente l'un des premiers réseaux de neurones capable de classifier des motifs de manière itérative, le perceptron simple couche. Sa règle d'apprentissage permet la séparation de classe par simple correction d'erreurs. Des hyperplans séparateurs sont trouvés (frontières de séparation), cependant le perceptron n'est capable de séparer que des entrées linéairement séparables (Figure 3-7). La règle d'apprentissage est la suivante:

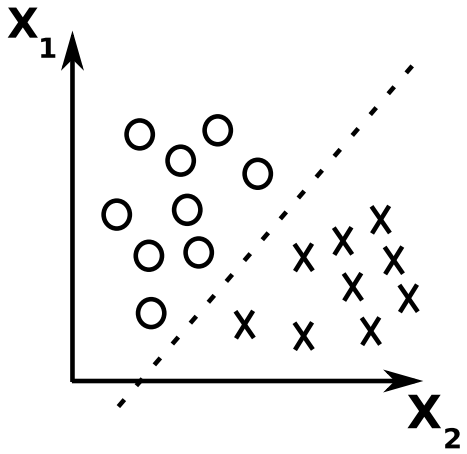


Fig. 3-7 – Problème linéairement séparable. Les motifs "croix" et "rond" sont caractérisés par les données X_1 et X_2. L'exemple montre un problème de classification linéairement séparable où un hyperplan séparateur (ligne pointillée) est trouvé pour résoudre le problème.

$$w_{ij}(t+1) = w_{ij}(t) + \epsilon(y_j^d - y_i)x_i(t) \tag{3.3}$$

x est le vecteur d'entrée appelé aussi "rétine", y_i est la valeur de sortie du j^{ieme} neurone du réseau, y_j^d est la sortie désirée du j^{ieme} neurone (la classe d'appartenance du motif) et ϵ est l vitesse d'apprentissage du réseau. L'activité y_j^d du j^{ieme} est généralement binaire $1,0$ ou $-1,1$. Cette règle d'apprentissage n'est pas suffisante lorsque le problème n'est pas linéairement séparable. Ce modèle conduit à des oscillations des sorties du perceptron. L'hyperplan séparateur

ne peut se stabiliser sur une solution optimale, la non linéarité des entrées induit une impossibilité de la convergence du réseau.

Une règle plus évoluée a été introduite par Widrow et Hoff (Widrow and Hoff, 1960) où la méthode des moindres carrées a été adaptée aux réseaux de neurones reposant une variante de la règle d'apprentissage du perceptron. Widrow se base sur la minimisation de l'erreur quadratique où l'erreur à minimiser est définie par:

$$\xi = E[\frac{1}{2}(Y^d - W^T X)^2] \tag{3.4}$$

L'apprentissage consiste ainsi à trouver le minimum global en fonction des poids synaptiques. La méthode utilisée pour atteindre à ce minimum est une descente des gradients où les poids sont modifiés de manière itérative comme suit:

$$W(t+1) = W(t) - \epsilon \Delta \xi \tag{3.5}$$

Δ symbolise le gradient. La règle d'apprentissage obtenue est équivalente à celle du perceptron mais contrairement à celui-ci les poids synaptique sont modifiés avant la fonction d'activation des neurones. Même quand le problème n'est pas linéairement séparable, la règle d'apprentissage de Widrow converge asymptotiquement vers une solution linéaire la plus proche de l'optimalité. La figure 3-8 illustre la convergence de la règle de Widrow également appelé LMS.

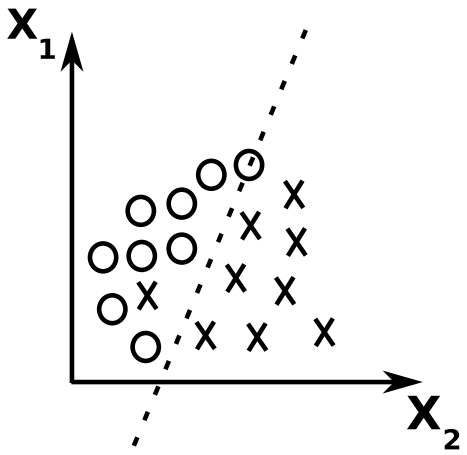


Fig. 3-8 – Hyperplan séparant deux classes dans un problème non linéaire. La règle d'apprentissage de Widrow est capable de déterminer cette frontière dans le cas d'un problème non linéaire. La frontière s'est adaptée en fonction de la non linéarité introduite par les entrées.

3.1.3.3 Réseaux multi-couches

Les réseaux de neurones multicouches ont été introduits pour résoudre les problèmes de non linéarités. L'apprentissage de ces réseaux de neurones multicouches est réalisé par l'intermédiaire de la rétro-propagation du gradient. Ces réseaux de neurones trouvent les hyperplans permettant la séparation des différentes catégories (Rumelhart and al. (D.E), 1986) (Figure 3-9). La technique de rétro-propagation du gradient ou "backpropagation" permet de calculer le gradient de l'erreur pour chaque neurone du réseau, de la dernière couche vers la première. Les poids

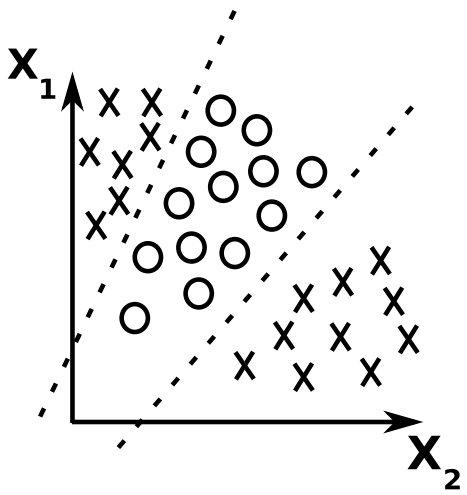


Fig. 3-9 – Hyperplans séparant deux classes dans un problème non linéaire. Le perceptron multicouche est capable de trouver plusieurs frontières à l'aide des couches cachées lorsque les motifs d'entrées sont non linéaires. Dans cet exemple, le MLP trouve deux hyperplans (ligne pointillé) séparant les deux classes.

synaptiques qui contribuent à engendrer une erreur importante se verront modifiés de manière plus significative que les poids qui ont engendré une erreur marginale. Ce principe est à la base des réseaux de neurones multicouches comme le perceptron multicouches.

3.1.3.4 Réseaux à base radiale

Les réseaux à base radiale sont constitués de fonctions d'activation radiale de la forme:

$$\phi_j(x) = \phi(\|x - \mu_j, \rho_j) \tag{3.6}$$

avec $\phi(u) = e^{\frac{r^2}{2\rho^2}}$ étant l'une des fonctions radiales les plus utilisées, μ_j et ρ_j étant respectivement la moyenne et l'écart type de la j^{eme} fonction radiale.

Un réseau à base radiale est une combinaison linéaire de fonctions radiales de centre:

$$y(x) = \sum_j w_j \phi(\|x - \mu_j, \rho_j) \tag{3.7}$$

Il est constitué de deux couches:

- la première couche (couche cachée) possède des neurones qui répondent spécifiquement pour certaines régions de l'espace d'entrée. Les fonctions à base radiale effectuent ce découpage de l'espace. Cette couche peut être déterminée par un apprentissage de type kohonen.
- la deuxième couche étant la couche de sortie, réalise une combinaison linéaire des sorties de la couche cachée. Un apprentissage de type perceptron permet l'évolution des poids des neurones de la couche de sortie.

Il existe d'autres méthodes (Moody and Darken, 1989) pour estimer les paramètres w_j, μ_j et ρ_j. Les paramètres μ_j et ρ_j peuvent être estimés à l'aide d'algorithme non supervisée de type k-moyennes et une fois ces paramètres fixés, il est possible de calculer les poids w_j optimaux par régression linéaire.

Les non linéarités introduites par les fonctions radiales permet d'obtenir une notion de distance, l'activité d'un neurone de la première couche est la même pour toutes les entrées situées à égale distance du centre de la fonction.

3.2 Les architectures neuronales

Les architectures neuronales que nous allons présenter, sont basées sur les réseaux de neurones précédemment décrit.

3.2.1 adaptative Resonnance Theory

Carpenter et Grossberg (Grossberg and Mingolla, 1985; Carpenter and Grossberg, 1987; Grossberg, 1988; Grossberg and Somers, 1991) ont créé un système d'apprentissage autonome capable de résoudre le compromis entre l'adaptation et la stabilité (compromis entre plasticité et rigidité synaptique): l'ART (adaptative resonnance theory) est décrite sur la figure 3-10. Un système adaptatif peut facilement diverger si la plasticité synaptique est trop importante, en apprenant des informations qui ne sont pas pertinentes ou en oubliant des informations apprises. Ces conséquences impliquent un manque de stabilité. D'autre part, un système dépourvu d'adaptation (les synapses n'évoluent pas) condamne tout apprentissage. Le compris entre plasticité et stabilité est un problème centrale pour être capable d'apprendre de manière autonome. L'ART propose une solution pour concilier ce compromis, introduisant un facteur de vigilance qui module la plasticité du système. L'ART est constitué d'une première couche F1 permet-

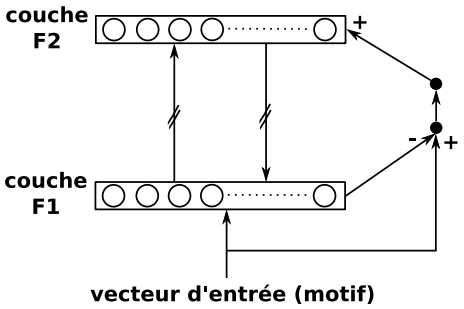


Fig. 3-10 – Schéma simplifié de l'architecture ART.

tant la normalisation des activités des motifs d'entrée et d'une seconde couche F2 effectuant la catégorisation des motifs. Des liens bidirectionnels relient F1 et F2. Leur modification permet un apprentissage long terme. En effet, les poids synaptiques reliant F2 à F1 permettent d'apprendre à reconstruire les motifs catégorisés. F1 possède une mémoire à court terme (STM) permettant une résonance si le motif a déjà été appris. La principale nouveauté de cette architecture est l'introduction d'un facteur de vigilance permettant de déterminer si un nouveau neurone doit être recruté pour coder le motif d'entrée ou si ce dernier est suffisamment proche d'un motif déjà appris. En effet, la vigilance permet de décider si la forme présentée est nouvelle pour être apprise. Un calcul de similarité est effectué permettant de décider si le motif est nouveau. Cette similarité est une comparaison de l'erreur de reconstruction de la forme mémorisé avec le motif actuel. La vigilance permet de régler la plasticité du système.

La première version de l'ART est consacré à des vecteurs d'entrées binaire {0,1} cependant d'autres versions ART2, ART3 ont été développée pour permettre l'apprentissage de

formes analogiques (Carpenter et al., 1991a). D'autres versions plus complexe comme l'ART-MAP ou le fuzzy-ARTMAP ont été développé permettant d'obtenir une notion de topologie a postériori (Carpenter et al., 1991b; Carpenter et al., 1992).

3.2.2 Counter Propagation

La Counter Propagation est une architecture de contrôle multicouches développée par Hecht-Nielsen (Hecht-Nielsen, 1987). Cette architecture est décrite figure 3-11, elle permet l'apprentissage de vecteurs d'entrées non linéaires combinant une couche neuronale auto-organisatrice (type WTA ou kohonen) et une couche neuronale dont l'apprentissage est supervisé. La première

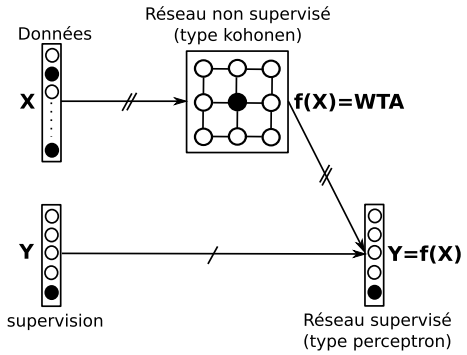


Fig. 3-11 – Schéma de la counter propagation. Architecture neuronale à deux couches. Une première couche servant de quantification vectorielle (réseau de neurone de type kohonen). Une seconde servant à catégoriser les classes déjà prédéfinies (réseau de neurone de type perceptron)

couche quantifie (vectorise) les données d'entrées servant à l'orthogonalisation des données d'entrées. Cette première couche simplifie donc les motifs d'entrées (peut être vue comme simple un changement de référentiel). La deuxième couche permet de catégoriser les vecteurs simplifiés de la première couche. La seconde couche est un réseau de neurone de type perceptron réalisant l'association entre des catégories déjà prédéfinie et la sortie de carte de kohonen.

3.2.3 Architecture PerAc

L'architecture PerAc (perception/action) est une architecture sensori-moteur où les sensations et les actions sont étroitement liées (Figure 3-12). Cette architecture s'inspirant de la counter propagation, a été développée pour le contrôle de robot interagissant avec leur environnement naturel. Elle a été introduite par Gaussier et Zrehen (Gaussier and Cocquerez, 1991; Gaussier and Zrehen, 1995) et étendue par (Joulain et al., 1997). PerAc permet des conditionnements sensori-moteurs grâce à un apprentissage de type associatif entre sensations et actions. Elle est constituée d'une première voie de bas niveau générant des comportements réflexes par exemple de l'évitement d'obstacle. Des capteurs (infrarouges, tactiles, vision de bas niveau ...) traitent les données venant de l'environnement déclenchant ainsi des actions réflexes. Cependant le câblage entre les comportements réflexes et les capteurs (données sensorielles frustres) est souvent ad-hoc. Cette voie réflexe est par conséquent idéale pour la survie du robot par exemple l'évitement d'obstacles. Une autre dynamique se greffe en parallèle de cette première voie permettant ainsi l'anticipation des comportements réflexes par l'intermédiaire de l'apprentissage. Cet apprentissage réalise une association entre des informations sensorielles de plus haut niveau et des comportements réflexes permettant ainsi la création de comportements complexes comme

la perception d'objet (Maillard et al., 2005a) ou le retour vers une source (Giovannangeli and Gaussier, 2008). PerAc est basé sur une architecture sensori-motrice où les actions modifient les

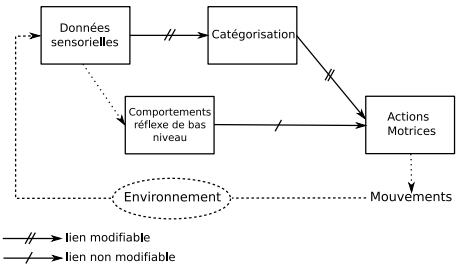


Fig. 3-12 – Schèma de la l'architecture PerAC (Perception/Action).

signaux sensoriels que le robot a de l'environnement. Les actions imposent donc une modification des sensations, ainsi une boucle sensori-motrice est créée et est fermée par l'intermédiaire de l'environnement. Le problème de l'ancrage des symboles ("symboles grouding problem") de l'IA classique (Harnad, 1990) est évité car l'architecture PerAC puise directement les informations venant des capteurs et de l'environnement sans besoin de caractériser la situation (pas de symbole unique), montrant ainsi que ce type d'architecture n'est pas figé à une tâche prédéterminée.

3.3 Formalisme des systèmes cognitifs

3.3.1 Pourquoi un formalisme pour les systèmes cognitifs?

Les architectures robotiques sont de plus en plus complexe par conséquent le besoin de pouvoir comparer des modèles différents pour résoudre des problèmes identiques commence à se faire sentir. Evidemment, des critères de performances sont mis en place pour juger le fonctionnement d'une tâche donnée dans un cadre de test relativement contrôlé mais dans ce cas il est effectué qu'une simple mesure de l'optimalité du système robotique face à la tâche considérée. Par conséquent, la comparaison de modèles reste difficile (points communs et différences) car aucun formalisme existant rend compte de la totalité du problème (Steels, 1994; Ikegami, 1993). Autrement dit, la détermination des briques élémentaires nécessaires à l'émergence des systèmes cognitifs n'est pas encore évidente aujourd'hui.

Différents formalismes existent, Smither propose par exemple des mesures du comportement d'un robot dans son environnement (interaction agent-robot est prise en compte) en utilisant les mesures empiriques venant des systèmes dynamiques (exposant de Lyapounov, dimension de corrélation ...) cependant les mesures sont lourdes et les applications sont simples (Smithers, 1995). De plus, l'analyse des réseaux de neurones artificiels restent difficile notamment la comparaison avec des systèmes biologiques. Les comparaisons sont possibles quand les réseaux sont de petites tailles et sans rebouclages. Très souvent, les problèmes complexes que nous traitons demandent des réseaux neuronaux qui sont de grandes tailles. Dans cette section, nous résumerons le formalisme introduit par Gaussier (Gaussier, 2001a; Maillard et al., 2005a; Prepin et al., 2005) pour l'étude des systèmes cognitifs qui a l'avantage de décrire de manière concise et synthétique les architectures de contrôles développées, contrairement à la figure 3-13 qui montre une architecture compliquée et difficile à analyser.

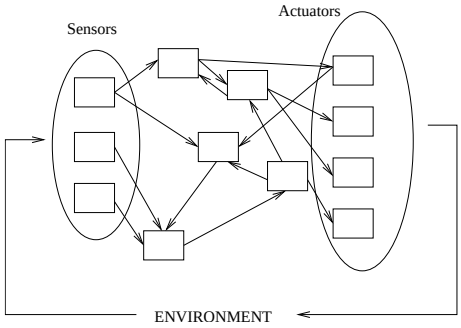


Fig. 3-13 – Architecture pouvant être manipulée par notre formalisme.

3.3.2 Représentation graphique

Le formalisme des systèmes cognitifs est un graphe orientée où un noeud (groupe de neurone) est considéré comme une fonctionnalité. Une fonctionnalité peut être une simple opération logique, un filtrage, réseau de neurone ... Tous les noeuds possèdent un vecteur d'entrée $|x>$ et un vecteur de sortie $|y>$ (Figure 3-14), les vecteurs sont constitués exclusivement de neurones dans les modèles que nous développerons. Les noeuds sont reliés par des connexions synaptiques

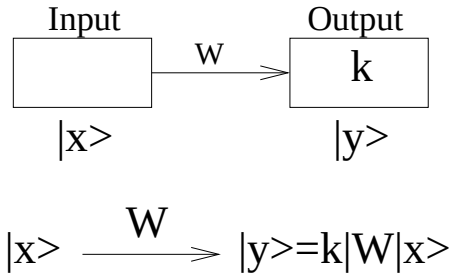


Fig. 3-14 – Schéma entrée-sortie entre deux noeuds reliés par des connections

(matrice de poids W) qui peuvent être de différents types:

- connexions de 1 vers 1: relient 1 à 1 les neurones de deux noeuds. Il n'y pas d'apprentissage sur ces connexions, elle est imposée par le concepteur pour obtenir des comportements réflexe.
- connexions de 1 vers tous: relient tous les neurones d'un groupe donnée avec l'ensemble des neurones d'un autre groupe. Ces connexions ont des capacités d'adaptation. Elles évoluent au fur et à mesure des interactions avec l'environnement induit des capacités d'apprentissage.
- connexions de 1 vers un voisinage: leur fonctionnalité peut être identique à la précédente seule la connectivité est réduite ou peut servir de projection.

La fonctionnalité d'un groupe de neurone et des connexions est régie par un opérateur k. Cet opérateur défini la règle de mise des neurones en fonction des entrées et des poids synaptiques, la règle d'évolution des poids ainsi la compétition entre les neurones d'un groupe. La figure 3-15 montre la représentation graphique du formalisme.

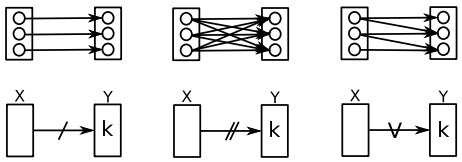


Fig. 3-15 – Schéma formalisant les 3 types de connexions entre noeuds. De gauche à droite: les connexions de 1 vers 1 (lien réflexe), les connexions de 1 vers tous (lien apprenant), les connexions de 1 vers un voisinage (lien pouvant apprendre ou projection).

3.3.3 Ecriture formelle

Les entrées et les sorties des systèmes cognitifs sont représentées sous forme de vecteurs en utilisant la notation "bra-ket" très utilisée en mécanique quantique. Un vecteur x est noté $|x>$ et son transposé $<x|$. $|x>$ représente l'activité d'un groupe de neurones de dimension m avec $|x> \in R^{+m}$. $<x|x>$ est un scalaire représentant la norme L^2 de $|x>$.

Soit un groupe de neurone Y d'activité $|y>$ ayant une fonctionnalité k, si Y est relié à un groupe de neurone X d'activité $|x>$ alors $|y>$ est déterminé par la relation:

$$|y> = k(W|x>) \tag{3.8}$$

k est l'opérateur indiquant comment appliquer la matrice des poids W au vecteur $|x>$. k peut avoir une fonctionnalité très simple comme le produit matricielle entre une matrice de poids W et un vecteur d'entrée $|x>$ ceci correspondrait à un neurone formel de Mc Culloch et Pitts. Cependant k peut avoir d'autres fonctionnalités par exemple un calcul de distance entre les colonnes de W et le vecteur $|x>$.

3.3.4 Cas d'un conditionnement: le perceptron

Le perceptron est un réseau de neurone se formalisant très bien à l'aide de l'algèbre des systèmes cognitifs (figure 3-16) dans le but de réaliser des conditionnements pouvant s'intégrer dans les architectures PerAc que nous développerons dans la suite. Les liens de "1 vers 1" décrits

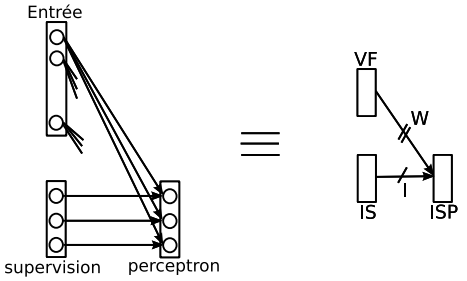


Fig. 3-16 – Exemple d'une architecture neuronale formalisée. Le réseau de neurone de gauche montre l'apprentissage d'un conditionnement et le schéma de droite montre l'écriture formalisée équivalente.

par la matrice I et reliant IS à ISP sont des liens réflexes dépourvus d'apprentissage permettant la supervision du réseau. Les liens de "1 vers tous" décrits par la matrice de connexions W reliant VF et ISP sont des connexions modifiables permettant l'apprentissage. Les sorties de

ISP s'expriment de la manière suivante:

$$|ISP >= D_{perceptron}(I|IS >, W|VF >) \tag{3.9}$$

$D_{perceptron}$ défini la règle d'activation des neurones ainsi que la règle d'apprentissage du perceptron.

3.3.5 Règles de simplifications

Des règles de simplifications peuvent être obtenues à partir de ces équations, ces simplifications reposent sur l'analyse des fonctionnalités des réseaux de neurones. Par exemple la figure 3-17 illustre des simplifications dans le cas d'un conditionnement, la dynamique du réseau avant apprentissage est très frustre, le réseau se comporte comme si seule la voie réflexe existait. Au contraire après apprentissage, le réseau se comporte comme si seule la voie conditionnelle existait. Il existe d'autres simplifiations qui sont relatives à l'enchainement des groupes de neu-

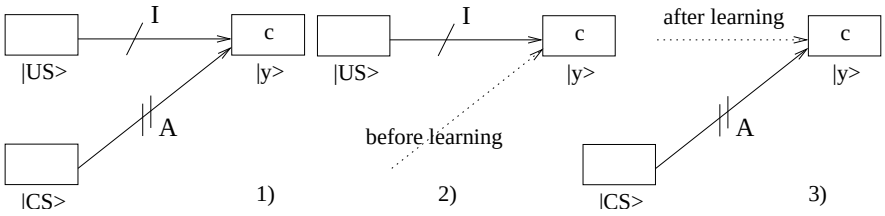


Fig. 3-17 – Les différentes phases de simplifications pour un conditionnement. 1) est la représentation graphique du conditionnement. 2) est la représentation graphique d'une simplification "avant apprentissage". 3) est la représentation graphique d'une simplification "après apprentissage".

rones reliés par des groupes de "1 vers 1" ou de "1 vers tous" sont réalisables sous certaines conditions (Figure 3-18). Ces règles de simplifications permettent d'une part de simplifier les

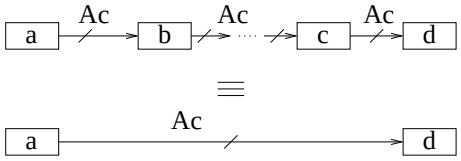


Fig. 3-18 – Equivalence d'une chaîne de liens inconditionnels. Chaque groupe de neurone effectue une compétition tout en étant relié au suivant par un lien de "1 vers 1".

architectures de contrôle et d'autre part de comparer les différentes architectures. L'ensemble de ce formalisme permet d'avoir une vision claire des éléments nécessaire pour la conception de systèmes cognitifs.

3.4 Conclusion

Les réseaux de neurones ainsi que l'utilisation d'un formalisme cognitif semblent indispensables pour la construction de modèle essayent de mimer des comportements adaptatifs. Les réseaux de neurones seront utilisés car ils possèdent des capacités d'apprentissage permettant l'adaptation à l'environnement. L'architecture PerAc sera à la base de nos développements car

nous soutenons l'idée que les sensations sont des éléments précurseurs de l'action. Enfin, le formalisme cognitif que nous avons décrit, permet d'avoir une vue d'ensemble du problème et d'obtenir une conception de systèmes cognitifs minimal. Les éléments constituants l'architecture sont élémentaires.

Deuxième partie

Interaction diadique: Homme - "Bébé robot"

Chapitre 4

Apprentissage et reconnaissance des expressions faciales émotionnelles à travers un jeu d'imitation

4.1 Introduction

Depuis plusieurs années, le thème de l'interaction homme/robot est devenu un important domaine de recherches. Les architectures proposées utilisent principalement des stratégies d'ingénierie ad-hoc permettant de montrer des résultats impressionnants. Cependant même si des techniques d'apprentissage sont utilisées, la plupart d'entre elles ne correspondent pas à la simulation d'un comportement adaptatif. Nous pensons que le développement d'architectures robotiques doit être compris dans une perspective développementale sinon le problème du symbol grounding (Harnad, 1990) pourrait apparaitre, révèlant le manque d'autonomie au système.

Dans ce chapitre, nous montrons qu'une tête robotique n'ayant au départ aucune connaissance du monde, peut apprendre à reconnaître des expressions faciales au fur et à mesure des interactions avec son environnement. L'interaction a une place prépondérante pour le développement de capacité cognitive. La question abordée dans ce chapitre est : comment une tête de robot peut-elle apprendre des expressions faciales de manière autonome? Cette question est primordiale pour le développement d'architectures plus complexes comme le référencement social que nous expliciterons dans les chapitres suivants. Dans ce chapitre, nous nous intéressons à comprendre comment un bébé apprend à reconnaitre des expressions faciales sans avoir de signal de supervision permettant d'associer par exemple la vision d'une expression de joie avec son propre état émotionnel interne de joie (G. Gergely, 1999). Notre point de départ a été motivé par la question de comment un système "naif" peut apprendre à répondre correctement à l'expression faciale d'une autre personne durant une interaction naturelle. "Naturelle" ici signifie que l'interaction devrait être la moins contraignante possible, sans signal de supervision explicite, mécanisme de détection ad-hoc ou sans technique d'enseignement formaté. Dans ce cas, une inspiration est donnée par l'interaction mère-bébé où, le nouveau né ou le très jeune bébé, a un ensemble d'expressions liées à ses propres émotions. Au tout début de sa vie, le lien avec les expressions de l'autre n'est pas encore construit, on se demande alors comment ce lien entre ses propres émotions et l'expressions de l'autre peut émerger d'une interaction non-verbale?

L'utilisation de conditions expérimentales minimales (Fig. 4-1) est importante pour éviter

les problèmes liés à ce qu'on appelle l'uncanny valley (vallée de l'étrange) (Masahiro, 1970). La vallée de l'étrange est un phénomène courant en robotique humanoïde où lorsque l'on cherche à améliorer la ressemblance du robot avec l'homme, le robot peut induire un sentiment de rejet et poser des problèmes pour la construction d'interactions homme/robot. Certaines rétissances peuvent apparaître lorsque le robot a quasiment forme humaine (il peut être pris pour un zombie). Un compromis peut être fait en utilisant des robots n'ayant pas une ressemblance humaine mais pouvant réaliser des "tâches cognitives". La tête expressive que nous utilisons est suffisamment élémentaire pour éviter ce problème.

De plus, l'avantage d'utiliser une tête robotique peut être justifie par:

- résonance émotionnelle: Les humains reproduisent "involontairement" les expressions faciales d'une tête de robot (Nadel et al., 2006b)
- Le robot facilite la communication car l'homme et le robot interagissent dans le même environnement ("présence" du robot). l'aspect actif et communicatif est agréable pour un expérimentateur humain.
- Le robot a besoin de diriger son regard vers des objets d'intêret. Il est pratique que ce dispositif puisse aussi permettre à l'humain de savoir dans quelle direction regarde le robot et ainsi servir de moyen de communication implicite. De la même manière, une réaction de retrait peut être vue comme de la peur.

Nous considérons dans nos développements que le robot est un jeune enfant et le partenaire humain interagissant avec lui correspond à un parent. De manière générale, le robot ne connait rien sur son environnement. Il ne posséde que des comportements réflexes (expressions faciales prototypiques). Il commence à apprendre grâce à l'interaction avec le monde extérieur: ici le partenaire humain constitue une grande partie de l'environnement.

En utilisant l'algèbre des systèmes cognitifs de (Gaussier, 2001b), nous avons montré qu'une simple architecture sensori-motrice basé sur un simple paradigme de conditionnement classique pouvait apprendre en ligne à reconnaitre des expressions faciales si et seulement si nous supposons que le "bébé robot" produit le premier des expressions faciales selon son propre état émotionnel interne et que l'expérimentateur l'imite permettant en retour l'association de son propre état interne avec la sensation visuelle de l'expressivité du partenaire humain (P. Gaussier, 2004). Des expériences de psychologies (Nadel et al., 2006b) ont montré que les humains 'reproduisent' involontairement les expressions faciales de notre robot. Cela met en évidence un bas niveau de résonance émotionnelle pouvant être considéré comme un amorçage pour l'apprentissage du robot [1].

Nous verrons dans la suite de ce chapitre un résumé du modèle formel pour l'apprentissage en ligne des expressions faciales, ensuite l'implémentation du modèle théorique sans détection de visage ainsi que les contraintes dues à l'apprentissage en ligne et enfin plusieurs résultats expérimentaux.

1. Il est difficile ici de statuer sur quel est l'état émotionnel de l'humain en face du robot. Dans une version optimiste, on peut considérer que certains humains ont de l'empathie pour la tête de robotique. De manière plus prosaïque, on détecte que les expressions faciales des sujets humains miment de manière involontaire l'expression de la tête de robotique traduisant la mise en oeuvre d'un mécanisme de bas niveau qui pourrait être impliqué dans le processus de résonance émotionnelle.

4.2 Matériel & Méthode: l'apprentissage en ligne

4.2.1 Description de la tête robotique

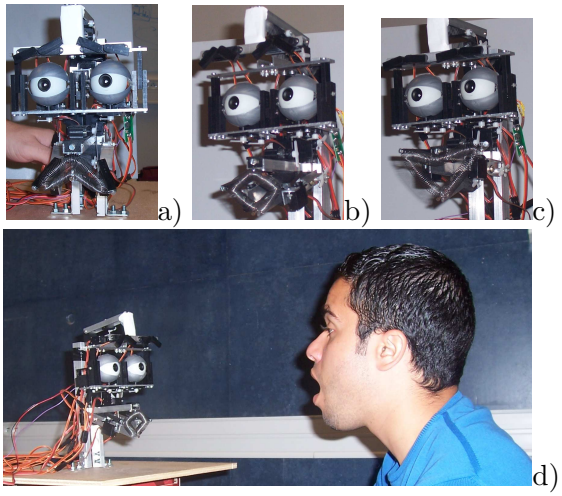


Fig. 4-1 – Exemples d'expressions faciales que le robot est capable de reproduire: a) tristesse, b) surprise, c) joie. d) situation typique du robot et de l'humain dans le cas d'un jeu d'interaction.

Une tête robotique très simple a été développée comme un outil pour des recherches en sciences cognitives impliquant à la fois des expériences psychologiques et des modèles computationnels (Nadel et al., 2006a). La tête a été conçue pour être un système minimum permettant l'affichage de certaines expressions faciales basiques et prototypiques FACS (Izard, 1971; Ekman et al., 1972; Ekman and Friesen, 1978; Plutchick, 1980). Ce dispositif n'a pas une réelle ressemblance humaine (dépourvue de peau), toutes les articulations sont visibles ainsi que le cablâge électrique. Cependant, les yeux, les sourcils, la bouche et le front sont suffisamment apparants pour que l'humain distingue les expressions faciales du robot (visage actif sans ressemblance avec le visage humain).

Dans ce travail, nous utiliserons simplement les expressions du visage suivantes: la joie, la tristesse, la colère, la surprise (Figure 4-1) ainsi que le visage neutre (inexpressif). La validité de ce choix peut être discutée (spécialement pour la surprise) (Jennifer M. Jenkis, 1998). Cependant, ce choix n'est pas crucial pour notre étude. Il suffit en fait d'un petit ensemble d'émotions, chacune peuvant être associée à des signaux internes présents chez l'humain et l'animal.

Notre tête de robot a été conçue par Pierre Canet au centre émotion de la Pitié Salpétrière dans le cadre d'un projet avec l'équipe neurocybernétique du laboratoire ETIS. Cette tête expressive est composé de 13 servos moteurs qui sont contrôlés par une mini carte SSC3 permettant de maintenir les servos moteurs dans une position donnée (contrôle en position). Ainsi différentes parties du visage sont mobile, permettant de créer une tête de robot expressive: 4 moteurs controlent les 2 sourcils, 1 moteur contrôle le front (mouvement du haut de la tête), 5 autres moteurs permettent le contrôle de la bouche (ouverture et écartement de la bouche artificielle), enfin 3 servos moteurs contrôlent l'orientation des 2 caméras localisées dans les yeux du robot avec 1 servo contrôlant le plan verticale (mouvement tilt) et 2 servos-moteurs contrôlant le plan horizontal (un moteur pour chaque caméra). Une caméra standard PAL fournie des images couleur qui sont ensuite transformées en noir et blanc. La tête de robot a été programmée pour

afficher 4 expressions faciales plus un visage neutre. Chaque expression faciale est produite par un positionnement précis de chaque servo moteur, a été contrôlée par des experts FACS (Ekman and Friesen, 1978). Le programme contrôlant la tête de robot permet à tous les servos moteurs de bouger en parallèle et chaque unité exécute la commande en position donnée par le contrôleur. L'expressivité du robot résulte en un processus dynamique et homogène où toutes les parties du visage se modifient pour former l'expression voulue. Le passage d'une expression à l'autre s'effectue entre 200-400 ms dépendant fortement de la distance entre l'expression courante et celle à effectuer (distance dans l'espace moteur). Grâce à la dynamique des servos-moteurs, le robot est capable de produire un grand nombre d'expressions faciales. Néanmoins, pour les expériences décrites dans ce chapitre, nous avons figé ce nombre à 5 prototypes expressifs.

4.2.2 Modèle théorique

Nous considérons un système simple composé de 2 agents interagissant dans un environnement initialement neutre: les 2 agents peuvent exprimer des expressions faciales (positionnement d'éléments de leur visage) en fonction d'un signal émotionnel interne et peuvent voir dans leur champ de vision le visage de l'autre (Fig. 4-2). Un agent est supposé être un adulte avec une

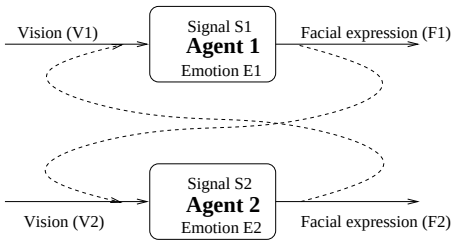


Fig. 4-2 – Le système bidirectionnel dynamique étudié. Les deux agents sont l'un en face de l'autre. L'agent 1 est considéré comme un nouveau né et l'agent 2 comme un adulte pouvant imiter les expressions faciales du nourrisson. Les 2 agents sont dirigés par des signaux internes qui peuvent induire des 'émotions' particulières.

parfaite reconnaissance de l'expression émotionnelle des autres et des capacités de reproduction parfaites. Le second agent est considéré comme un nouveau né sans connaissance du monde extérieur et sans apprentissage du rôle social des émotions. Comment le second agent peut-il acquérir des connaissances? L'interaction avec l'environnement peut être une solution pour la découverte de soi et de l'autre. Nous supposons simplement que l'apprentissage des expressions faciales est possible si et seulement si l'agent 2 produit des expressions et que l'agent 1 l'imite. Par conséquent, si l'agent 2 est doté de mécanisme sensori-moteur alors il sera capable d'associer des sensations avec des actions. Sans langage, sans connaissance de l'autre, l'agent 2 (bébé) pourra reproduire des expressions vues sur le visage d'un expérimentateur. Formellement, l'agent bébé peut être décrit comme sur la figure 4-3. Nous supposons que nos agents reçoivent un signal visuel (V_i vision de l'agent i). Cela peut être appris et reconnu par le groupe de neurone VF_i (caractéristiques visuelles de l'agent i), VF_i pouvant être le résultat d'un apprentissage non supervisé tel que WTA (Rumelhart and Zipser, 1985) (Winner Take All), réseau ART (Grossberg, 1987) ou une carte de Kohonen (Kohonen, 1989b). Ainsi, la présence d'un visage exprimant une expression particulière déclenchera l'activation de plusieurs neurones correspondant à une certaine caractéristique faciale dans le groupe VF_i:

$$VF_i = c(A_{i1}.V_i) \tag{4.1}$$

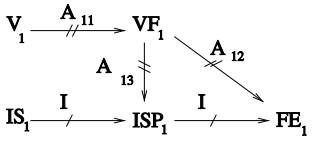


Fig. 4-3 – Représentation schématique d'un agent qui peut exprimer et reconnaitre des "émotions". Les flèches avec un trait représentent des connexions réflexes de "un vers un". Les flèches avec les deux traits en parallèle représentent des connexions de "un vers tous" modifiables. V: Stimulus visuel, VF: reconnaissance de vue locale (caractéristiques visuelles), IS: état interne, ISP: prédiction de l'état interne (état émotionnel), FE: expression facial (command motrice).

où c est un mécanisme de compétition, A_{i1} représente la matrice de poids des neurones dans le groupe de reconnaissance de l'agent i permettant une quantification des caractéristiques (bouche, sourcils ...). Par exemple, si l'agent bébé perçoit un visage exprimant de la joie, la caractéristique première caractérisant cette expression faciale est la bouche. La bouche pourra être apprise par le groupe VF_i (un neurone s'activera en présence d'un sourire), une fois appris les poids A_{i1} reliant l'entrée visuelle au réseau de neurones sont corrélés avec la bouche exprimant un sourire. D'autre part, nos agents sont également affectés par la sensation de leur milieu interne (faim, peur...). IS_i (état interne de l'agent i) est lié avec les entrées physiologiques tel que la faim, la peur, le plaisir... La détection d'un état interne particulier déclenchera un état émotionnel ISP_i. Nous supposons aussi que ISP_i dépend de la reconnaissance visuelle VF_i (visual features) qui elle dépend du signal visuel V_i.

De plus, les agents peuvent exprimer des commandes motrices FE_i (facial expression) correspondant aux différentes expressions faciales. Si l'un des deux agents peut agir comme un adulte, il doit avoir la capacité de "ressentir" les émotions des autres (empathie). Les connexions entre la reconnaissance visuelle et le groupe de reconnaissance ISP_i représentant l'état émotionnel du robot doivent exister. Pour afficher une expression émotionnelle, nous devons aussi supposer qu'il y a des connections allant des signaux internes vers le contrôle des expressions faciales. Les connections entre l'état interne et la commande motrice peut se faire directement (le signal circule sans interférence) ou au travers d'un groupe intermédiaire dévoué à la reconnaissance des expressions émotionnelles ISP_i. Par soucis d'homogénéité, nous supposons que le signal interne active par l'intermédiaire de lien inconditionnel, le groupe de reconnaissance de l'émotion qui lui même active également par des connexions inconditionnelles l'affichage des expressions faciales. Ceci est équivalent à une activation directe de FE_i par IS_i (chapitre 3). Donc, la somme de ces deux flux d'informations est:

$$ISP_i = c\,(I.IS_i + A_{i3}.VF_i) \tag{4.2}$$

Enfin, nous pouvons aussi supposer que l'agent enseignant peut également exprimer une émotion sans la "ressentir" (simplement par un comportement imitatif obtenu par reconnaissance de l'expression de l'autre). Cela peut être l'effet d'une résonnance émotionnelle ou d'une simple résonnance motrice. La sortie motrice des 2 agents dépend à la fois de l'expression faciale reconnue et de la volonté d'exprimer un état interne particulier:

$$FE_i = c\,(I.ISP_i + A_{i2}.VF_i) \tag{4.3}$$

Nous pouvons étudier les conditions minimales permettant la fabrication d'un attracteur comportemental global (pour imiter et comprendre les expressions faciales de l'autre). La figure 4-4 représente le système complet avec les deux agents en interaction. Ce système est considéré comme un réseau virtuel qui peut être étudié de la même manière qu'une archi-

tecture isolée. L'architecture est viable si et seulement si l'agent enseignant a les connaissances nécessaires pour que les dynamiques du systéme globale soient stables (boucle de sensation/action). La compréhension de ce genre de système n'est possible qu'à partir du moment où les deux agents (enseignant/apprenant) ne forment qu'une seule entité indissociable. Après

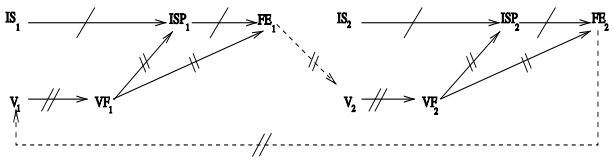


Fig. 4-4 – Représentation schématique du réseau global représentant les interactions émotionnelles entre les 2 agents. Les liens en pointillés représentent les connections venant de l'observation de l'autre individu. L'un des deux agents expriment une expression pendant que l'autre la voit.

simplifications (P. Gaussier, 2004), nous obtenons finalement le réseau de la figure 4-5a. On

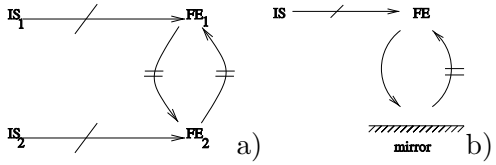


Fig. 4-5 – a) simplification finale du réseau représentant l'interaction entre nos deux agents (capacité à exprimer et comprendre des "émotions"). b) Architecture minimale permettant à l'agent d'apprendre les associations entre les "états internes" et l'expression faciale.

peut alors simplement en extraire la condition de stabilité durant la période d'apprentissage. Si les deux agents expriment leur état émotionnel interne IS_1 et IS_2 de façon indépendante (IS_1 et IS_2 sont différents n'impliquant pas de corrélation) alors l'apprentissage est impossible. L'apprentissage ne pouvant pas se stabiliser dans ces conditions et en supposant que l'état interne du bébé n'est pas contrôlable alors la seule solution est que l'agent "enseignant" imite ou résonne (Nadel et al., 2006b) aux expressions faciales du bébé, permettant ainsi une corrélation explicite entre l'état interne du bébé et l'entrée visuelle (l'expression qu'exprime l'agent "enseignant"). Le parent n'est alors rien de plus qu'un miroir.

Si cette condition est vérifiée, le système peut apprendre. L'agent 1 (nouveau né) apprend à associer la reconnaissance visuelle des expressions faciales prototypiques avec ses propres états émotionnels internes (ISP_1). L'agent apprend à connecter son "ressenti" lié aux actions, avec la vision des mouvements non "ressentie" de l'autre, ne pouvant voir ses propres actions. Ce paradigme met en évidence que l'agent nouveau né apprend la corrélation entre ce qu'il ressent et ce qu'il voit.

4.2.3 Protocole expérimental

Pour tester notre modèle, nous proposons de développer une architecture neuronale qui sera testée dans le protocole expérimental suivant: lors de la phase d'apprentissage, le robot produit des expressions faciales de manière aléatoires parmi les expressions de surprise, joie, colère, tristesse et sans oublier le visage neutre (chacune durant 2 secondes). Entre chaque expression faciale le robot repasse par le visage neutre pour éviter les mauvaises interprétations par un

observateur. Une procédure équivalente est utilisée en psychologie expérimentale permettant d'éliminer le biais expérimental suite à une nouvelle expression du robot. On demande au sujet humain d'imiter chaque expression du robot. Après cette première phase qui dure entre 2 et 3 minutes selon la 'patience' du sujet, le générateur d'état émotionnel aléatoire est stoppé. La deuxième phase d'interaction débute alors. On demande à l'humain de produire une expression pendant quelques secondes et on teste la capacité du robot à mimer à son tour l'expression faciale du partenaire humain.

Dans la section suivante, nous allons présenter l'architecture computationnelle (figure 4.6) per-

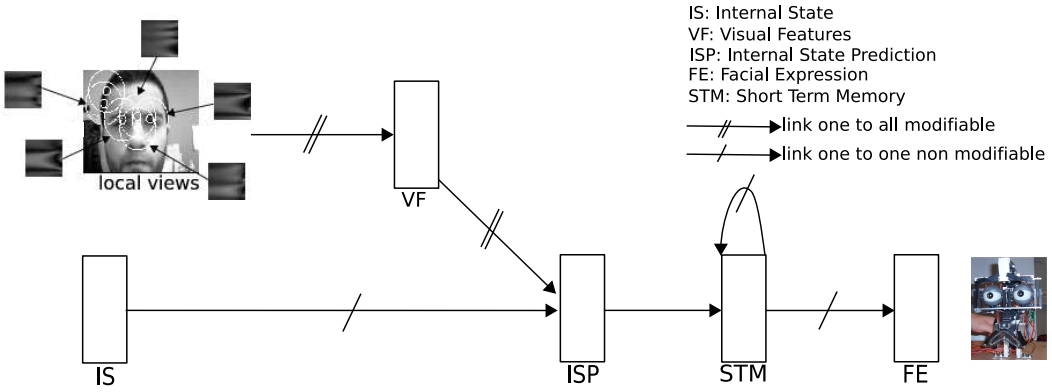


Fig. 4.6 – Architecture imitant les expressions faciales. Processus visuel permettant d'extraire les vues locales de manière séquentielle. La prédiction de l'état interne (groupe ISP) apprend l'association entre les vues locales et l'état interne (groupe IS).

met de reconnaître certaines caractéristiques faciales du partenaire humain et de les apprendre si ces caractéristiques sont corrélées avec ses propres expressions faciales.

4.3 La reconnaissance de l'expression faciale

4.3.1 Introduction

La partie vision et reconnaissance d'expressions faciales à tout d'abord était traitée de façon classique, à savoir plusieurs étapes qui se succédent: (1) localisation du visage utilisant par exemple (R.L Hsu, 2002) or (Viola and Jones, 2001), puis (2) cadrage du visage, et (3) reconnaissance de l'expression faciale de l'image normalisée. Dans ce cas, la qualité des résultats est hautement dépendante de la précision du cadrage du visage. La capacité de généralisation peut être affectée par cette précision. De plus, le robot ne peut pas réellement être autonome car l'apprentissage de la discrimination de visage/non visage est réalisé hors ligne. Très étrangement, l'apprentissage en ligne de la reconnaissance visage/non visage n'est pas aussi évident que l'apprentissage en ligne des expressions faciales. L'apprentissage de la discrimination visage/non visage est plus complexe à obtenir que celui des expressions faciales car il faudrait trouver un signal moteur lié à une activité "visage/non visage" qui puisse être émise par le robot et auquel l'humain réagirait en rendant visible ou non son visage de manière à ce que la discrimination visage/non visage puisse émerger. Il est bien évident qu'en situation naturelle cela semble on ne peut plus compliqué surtout que la palette des visages affichés par l'humain devrait être suffisament générique pour permettre une bonne généralisation. L'affichage d'une expression spécifique et la réaction par l'affichage en miroir de la même expression semble être la seule solution pour

disposer des données nécessaires à la discrimination des expressions mais aussi des visages en général.

Dans la suite de ce chapitre, nous détaillerons notre modèle sur la reconnaissance des expressions faciales qui ne suit pas les 3 étapes décritent. Dans le chapitre 5, nous décrirons l'architecture permettant de détecter la présence ou l'absence d'un visage grâce à l'interaction émotionnelle.

4.3.2 Traitement visuelle

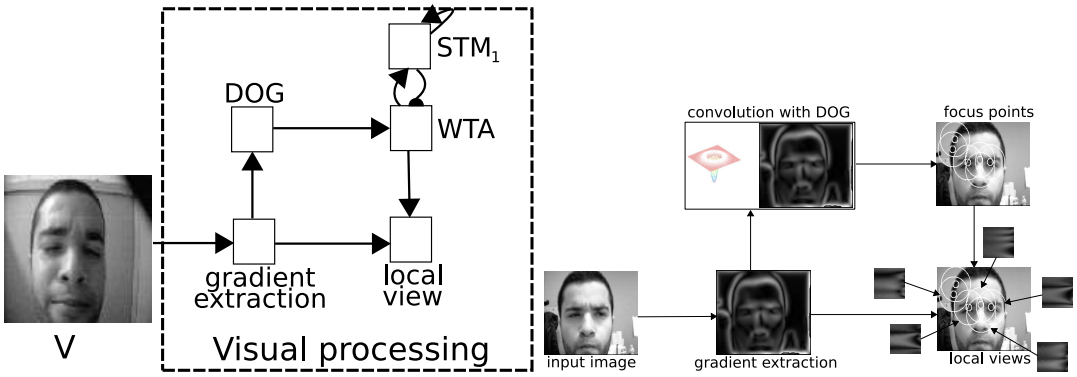


Fig. 4-7 – Processus visuel: Notre modèle est basé sur une exploration séquentielle des points de focalisations de l'image. La taille de l'image d'entrée est de 256x192 pixels. Une convolution de l'image avec une différence de gaussienne (DOG) est réalisée permettant d'extraire les points de focalisations. Pour finir, des vues locales sont extraites autour de chaque point de focalisation. Les vues locales constituent la signature de l'image.

Dans la perspective d'un apprentissage autonome, nous cherchons à éviter des mécanismes ad-hoc non compatible d'un point de vue développemental tels que le cadrage du visage. Nous proposons donc une alternative basée sur une exploration des points de focalisations de la scéne (Figure 4-7). Les points de focalisations sont les maxima locaux d'une convolution entre un filtre de différence de gaussiennes (DOG) et l'image du gradient (obtenue par un filtre de Canny-Derriche). Ce traitement visuel est déjà utilisé dans plusieurs travaux notamment pour la création de cellules de lieux et pour la reconnaissance d'objets (Giovannangeli et al., 2006; Maillard et al., 2005b). Ce mécanisme visuel permet au système de focaliser plus sur les coins, les fins de lignes d'une image (sourcils, coins de bouche, etc). Ces principaux avantages par rapport à des méthodes de type SIFT (Scale Invariant Feature Transform) (Lowe, 2004) sont: un coût computationnel réduit, très peu de points extraits et un algorithme neuromimétique (les filtres DOG modélisent bien le comportement de certaines cellules visuelles). Une fois la convolution calculée, l'intensité du point est directement liée au niveau d'intérêt.

L'un après l'autre, les points de focalisation les plus actifs sont utilisés pour extraire une caractéristique locale ou vue locale, soit par l'intermédiaire de la transformée log-polaire soit par le passage de filtres de Gabor (Figure 4-8). La transformation log-polaire locale augmente la robustesse des vues locales extraites par rapport aux petites rotations et aux changements d'échelles. Cette transformations est centrée sur le point de focalisation et prend en compte une zone autour de celui-ci de taille 20x20 correspondant à un cercle de R pixels de rayon dans l'image d'entrée. Le vecteur de signature est donc de dimension 400.

Les filtres de Gabor ont également été utilisés car ils conduisent à une caractérisation plus

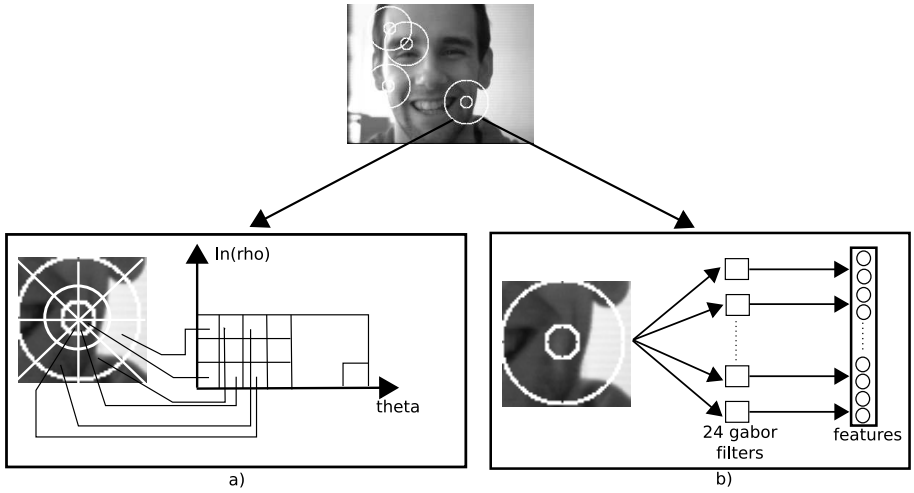


Fig. 4-8 – Caractéristiques visuelles: a) la transformée polaire locale augmente la robustesse des vues locales extraites aux petites rotations et changements d'échelles. Cette transformation est centrée autour du point de focalisation en prenant un rayon de 20 pixels autour de celui-ci. Cette transformation consiste simplement à changer de référentiel (coordonnée log-polaire). b) Les filtres de Gabor sont appliqués pour obtenir une signature plus robuste que la transformée log-polaire (les filtres de Gabor sont de tailles 60x60). La signature extraite par l'intermédiaire de cette convolution est la moyenne et l'écart-type des différents filtres de Gabor.

robuste aux différentes transformations (rotations, variations de la distance). Les filtres de Gabor suivent l'équation suivante:

$$G_{\theta,f}(x,y) = \frac{1}{2\pi\sigma}\exp(-\frac{x^2+y^2}{2\sigma^2})\cos(2\pi f(x\cos(\theta)+y\sin(\theta))) \tag{4.4}$$

$$f = \frac{1}{\sigma\gamma} \tag{4.5}$$

Les paramètres importants pour la construction de notre banc de filtre sont θ et f qui sont respectivement l'orientation et la fréquence. Ces paramètres sont utiles pour identifier différentes textures locales dans l'image. La figure 4-10 illustre le traitement visuel utilisé pour les filtres de Gabor: la construction des différents filtres ayant une fréquence et une orientation fixées (nous avons utilisé 24 filtres), la convolution avec ces filtres capable de faire ressortir les textures de l'image et enfin l'extraction de caractéristiques. Les caractéristiques extraites pour chaque image issue de la décomposition de Gabor sont la moyenne et l'écart-type. Le vecteur de signature est de dimensions 48. L'utilisation de la décomposition en Gabor induit un vecteur de caractéristique 8 fois plus petit que pour le traitement log-polaire, un gain significatif de calcul et une économie de synapses est à constater pour les réseaux de neurones traitant ces informations.

4.3.3 Architecture de contrôle neuronale

Les vues locales sont apprises par recrutement de nouveaux neurones, cette apprentissage est réalisé par le groupe de neurone VF (caractéristiques visuelles) en utilisant une variante de l'algorithme k-means permettant un apprentissage en ligne et un fonctionnement temps réel. Le réseau de neurone est comparable au développement de (Kanungo et al., 2002) mais avec

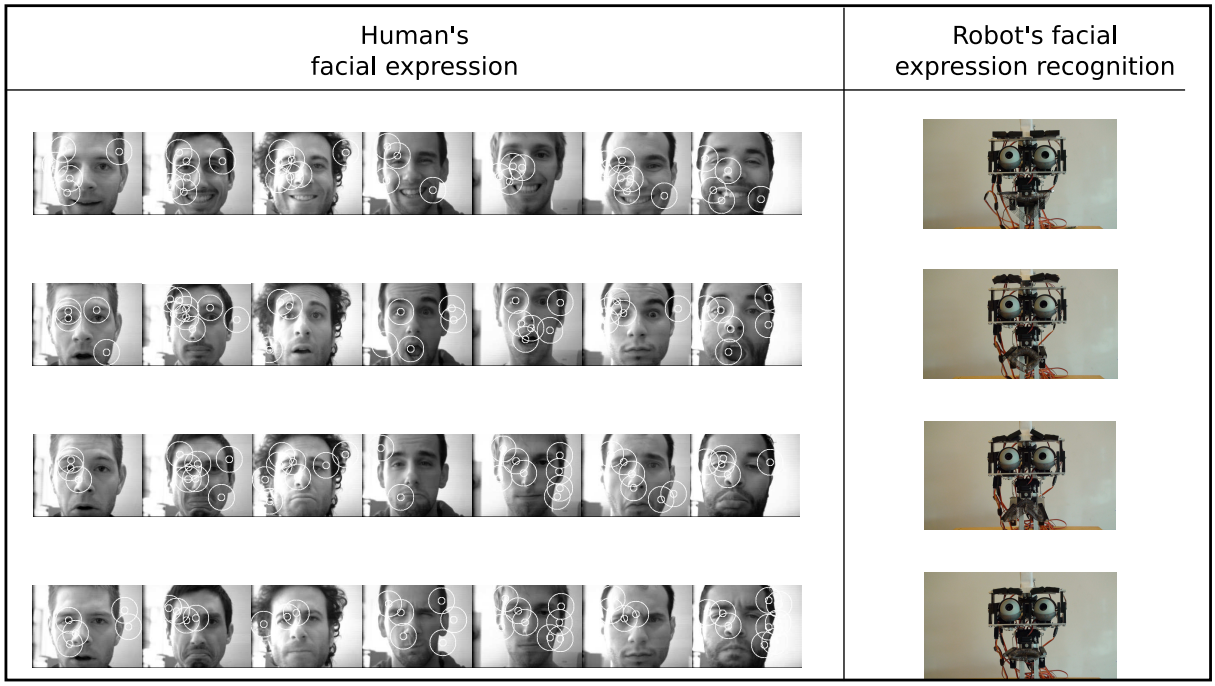


Fig. 4-9 – Etude après l'apprentissage des expressions faciales. Les cercles sont centrés sur les points de focalisations qui ont été associés aux expressions de joie, surprise, tristesse, colère. Les points de focalisations qui votent pour l'expression sont trouvés dans les mêmes zones suivant les différents individus. Par exemple, pour l'expression de joie, les points de focalisations sont principalement trouvés autour de la bouche (coin de bouche) contrairement à l'expression de colère où les points sont généralement trouvés autour des sourcils.

un apprentissage très rapide parfois en un coup et un temps d'exécution réduit. De plus les équations sont neuronales et ont certaines similitudes avec la règle de Kohonen. Nous avons nommé ce réseau de neurone SAW (Self Adaptatif Winner) et sa règle de mise à jour est:

$$VF_j = net_j . H_{max(\gamma, \overline{net}+\sigma_{net})}(net_j) \tag{4.6}$$

$$net_j = 1 - \frac{1}{N} \sum_{i=1}^{N} |W_{ij} - I_i| \tag{4.7}$$

avec VF_j l'activité du neurone j dans le groupe VF, $H_\theta(x)$ la fonction de Heaviside[2], γ la vigilance qui équivaut à un certain seuil de reconnaissance. $\overline{net}$ et σ_{net} sont respectivement la moyenne et l'écart-type des neurones de sorties VF_j. Son fonctionnement est le suivant: si la reconnaissance du prototype est en dessous de γ alors un nouveau neurone est recruté sinon les poids du neurone le plus actif sont moyennés.

La règle d'apprentissage permet à la fois un apprentissage en un coup et un moyennage des prototypes dans le temps. Cette règle mélange deux notions qui sont d'apprendre les choses nouvelles très rapidement et pouvoir s'adapter dans le temps. La modification des poids synaptiques

2. Fonction de Heaviside:

$$H_\theta(x) = \begin{cases} 1 & \text{si } \theta < x \\ 0 & \text{sinon} \end{cases}$$

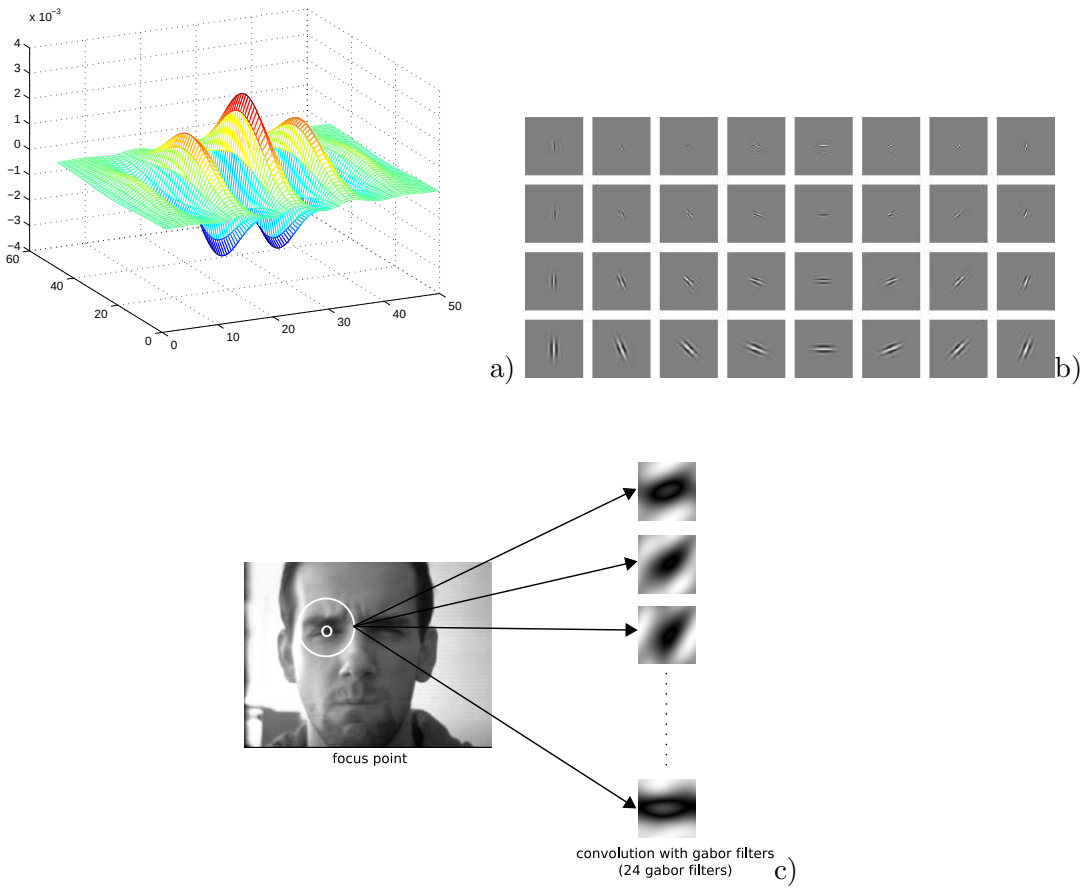


Fig. 4-10 – Visualisation du processus utilisant les filtres de Gabor. a) montre un filtre de Gabor avec les paramètres suivant $\sigma = 8$, $\gamma = 3$, $\theta = \pi/3$. Visualisation d'une Gabor dans un espace 3 dimensions. b) montre différents filtres de Gabor avec différentes orientations et fréquences. Enfin c) montre le resultat d'une convolution avec les 24 filtres de Gabor utilisés pour un point caractéristique donné (au niveau de l'oeil).

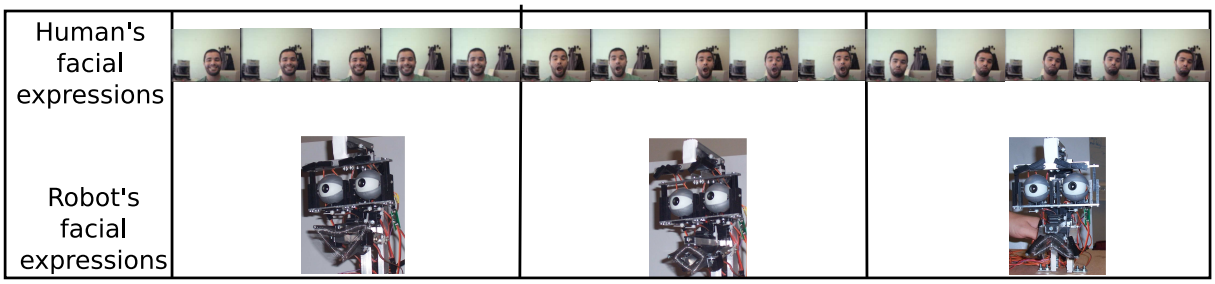


Fig. 4-11 – Test de robustesse qualitatif. Le robot est capable de reconnaitre l'expression faciale lorsque le partenaire humain est à une distance plus grande du robot. L'humain et le robot interagissent à environ 2 mètres de distance.

est calculée comme suit:

$$\Delta W_{ij} = \delta_j{}^k(a_j(t)I_i + \epsilon(I_i - W_{ij})(1 - VF_j)) \tag{4.8}$$

avec $k = ArgMax(a_j)$, $a_j(t) = 1$ seulement quand un nouveau neurone est recruté sinon $a_j(t) = 0$, ${\delta_j}^k$ est le symbole de Kronecker[3], I_i est l'entrée visuelle, W_{ij} le poid synaptique reliant l'entrée I_i avec le j_{eme} neurone et ε est la vitesse d'apprentissage (moyennage). Lorsqu'un nouveau neurone est recruté, les poids sont modifiés pour être corrélés avec l'entrée (terme $a_j(t)I_i$).

L'autre partie de la règle d'apprentissage $\varepsilon(I_i - W_{ij})(1 - VF_j)$ moyenne les prototypes anciennement appris. Plus les entrés sont proches des poids, moins les poids sont modifiés. Au contraire, moins les entrés sont proches des poids, plus les poids vont être modifiés par l'intermédiaire du moyennage. La qualité des résultats dépend de la valeur ε comme montré sur la figure 4-12. Si ε est choisi trop petit cela implique un impact faible sur la généralisation. Au

.ϵ	sadness	neutral	happy	anger	surprised
1	49%	48%	70%	97%	66%
0.1	58%	50%	67%	81%	84%
0.01	68%	55%	67%	90%	83%
0.001	76%	50%	62%	92%	89%
0.0001	78%	46%	70%	86%	68%

Fig. 4-12 – Taux de succès hors ligne pour la reconnaissance de l'expression faciale. Taux de succès de la reconnaissance de l'expression faciale avec comme processus visuel la transformée log-polaire. Réponse à choix forcé grâce au Winner Take All. Les taux de succès sont calculés sur les 4 expressions faciales et sur le visage neutre. Effet du facteur d'adaptation ε (moyennage de la vue locale) sur les performances globales du système avec une vigilance γ constante et γ=0.92. La base de donné est constituée de 4 individus pour chacun de ces individus, nous avons 32 images par expression faciale.

contraire, quand ε est trop grand, les prototypes anciennement appris peuvent être désappris. Grâce à cette règle d'apprentissage, les neurones du groupe VF apprennent les caractéristiques du visage (par exemple, le moyennage du coin de la bouche pour l'expression de joie). Autrement dit, chaque neurone de ce groupe représente une caractéristique du visage.

Il n'y a aucune contrainte dans la sélection des vues locales dans le sens où l'humain ne met pas de connaissance dans le système. Par exemple, l'humain ne dit pas au système que les points qui sont importants pour l'expression de colère sont ceux autour des sourcils, signifiant que des distracteurs peuvent être présents. Les distracteurs sont généralement pris sur le fond de la scène ou bien sur des parties non expressives du partenaire comme les cheveux. Par conséquent, tous ces distracteurs sont appris par le groupe de neurone VF, impliquant que des neurones seront activés en présence de certaines caractéristiques non-émotionnelles.

Néanmoins, l'ensemble du modèle renforcera exclusivement les caractéristiques expressives du visage (Figure 4-6). En effet, dans des situations de face à face, les distracteurs seront présents pour toutes les expressions faciales, ils ne seront donc pas associés à un état émotionnel spécifique. Leur activité tendra forcément vers zéro dû à la règle d'apprentissage suivante:

$$\Delta w_{ij} = \epsilon.Vf_i.(IS_j - ISP_j) \tag{4.9}$$

ISP correspond à la prédiction de l'état interne (état "émotionnel") qui associe l'activité de VF avec l'état interne courant IS du robot. L'association se fait par simple mécanisme de

3. fonction de Kronecker:

$${\delta_j}^k = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise} \end{cases}$$

conditionnement (Widrow and Hoff, 1960). j étant l'indice des neurones ainsi ISP_j correspond au j^{eme} état émotionnel ($0 < j \leq 5$).

STM_2 est une mémoire à court terme utilisée pour sommer et filtrer les états émotionnels $ISP_i(t)$ associés avec chaque vue locale explorée:

$$STM_{2,i}(t+1) = \frac{1}{N}.ISP_i(t+1) + \frac{N-1}{N}STM_{2,i}(t) \tag{4.10}$$

La STM_2 accumule les vues locales sélectionnées par les DOGs.

Enfin, le dernier groupe de neurone FE déclenche l'expression du robot, seule l'activité la plus importante FE_i déclenche la i^{eme} expression faciale (mécanisme de Winner Take All). Pour accroitre la robustesse et la stabilité du système, FE posséde une mémoire à court terme, donnant plus d'importance au présent qu'au passé, son principal avantage réside dans la stabilité du système physique (le robot ne passe pas d'une expression à l'autre). Les paramètres de cette mémoire à court terme sont $\beta = 1$ et $\alpha < 1$ ($\alpha = 0.8$):

$$FE_i(t+1) = \beta.STM_{2,i}(t+1) + \alpha.FE_i(t) \tag{4.11}$$

Arbitrairement, nous avons limité le temps d'analyse d'une image pour garder un système suffisamment réactif. Ce temps dépend exclusivement du nombre de points de focalisations traités par image. Nous avons fait le choix d'analyser seulement 10 points de focalisations par image. Analyser moins de points de focalisations peut rendre le système plus instable. Mais il ne faut pas croire qu'analyser plus de points, pourrait résoudre d'une manière optimale la convergence du système. Bien au contraire, nous perdrons à deux niveaux: le premier serait d'introduire plus d'ambiguité vis à vis de l'apprentissage car les vues locales seraient de moins en moins robustes et le deuxième est le temps de réaction du système car le robot ne réagirait plus aussi rapidement. Parmi ces 10 vues locales, nous remarquons que 4 ou 5 vues sont généralement prises sur des zones importantes du visage permettant de caractériser une expression faciale (bouche, sourcils). Cela suffit à la convergence du modèle et à maintenir une interaction temps réelle (5 images et actions par seconde).

4.3.4 Résultats

Après l'apprentissage, les associations entre VF et ISP sont suffisamment fortes pour bypasser l'activité réflexe de bas niveau venant de l'état interne IS (voir la section 4.2). Dans ce cas, l'expression faciale FE correspond à l'intégration temporelle de l'état émotionnel associé aux différentes caractéristiques visuelles analysées par le système. Ces caractéristiques ont une valeur émotionnelle, si elles ont été corrélées avec les caractéristiques expressives du visage humain. La figure 4-9 montre la tête expressive imitant les différentes expressions du partenaire humain ainsi que les points de focalisations qui sont associés à l'expression. Nous pouvons voir que chaque expression est principalement caractérisée par un ensemble spécifique de points de focalisations correspondant aux zones locales qui sont importantes pour la reconnaissance des expressions spécifiques. Par exemple, certaines vues autour de la bouche caractérisent la joie, certaines autres autour des sourcils caractérisent la colère. Nous remarquons que ces vues locales sont prises dans les mêmes zones quelque soit l'individu, laissant supposer que le mécanisme visuel de bas niveau focalise sur les mêmes caractérisques expressives quelque soit l'expérimentateur.

Après l'apprentissage, la figure 4-11 montre que le robot reconnait les expressions faciales même lorsque les interactions se font à des distances de 2 mètres. Dans cette expérience, le

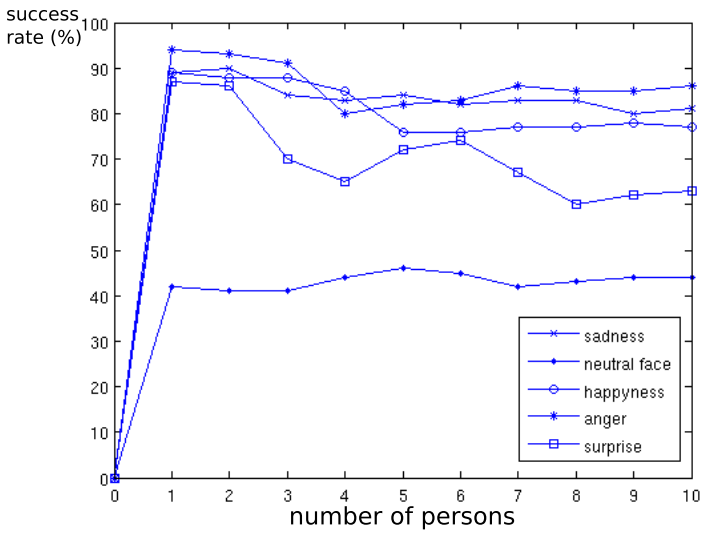


Fig. 4-13 – Le taux de succès pour chaque expression faciale (tristesse, joie, colère, surprise et visage neutre) avec la transformée log-polaire. Ces résultats sont obtenus durant une interaction naturelle avec la tête robotique. 10 personnes ont interagi avec le dispositif robotique (32 images par expressions par individus). Durant la phase d'apprentissage (2 minutes d'interaction), les individus doivent imiter le robot et une fois cette phase d'apprentissage finie l'on bascule le système en mode reproduction où le robot dans ce cas doit imiter l'humain. Pour fabriquer les statistiques, chaque image a été annotée en fonction de la réponse du robot. Les images annotées ont été analysées pour vérifier s'il y avait correspondance entre l'expression réalisée par le robot et celle réellement exprimée par l'humain.

protocole est le suivant: l'humain imite le robot à une distance de 2 mètres et lors de l'inversion des rôles, l'humain reste à une distance d'environ 2 mètres (plus ou moins quelques centimètres). Cette expérience met en évidence que les paramètres du modèle n'ont pas été réglés pour fonctionner à une distance fixée. Sans modification de l'architecture, le robot est toujours capable d'imiter même à des distances qui ne sont pas naturelles pour des interactions avec une tête de robot. Dans ce genre d'interaction, la deuxième chose qui est mise en avant est que le système est capable d'apprendre à discriminer les informations du fond de la scène par rapport aux caractéristiques visuelles du visage. Le système renforce d'avantage les caractéristiques expressives (vues locales associées à une valeur émotionnelle).

De manière générale, les figures 4-12, 4-13, 4-14, 4-16 montrent que notre architecture est capable de reconnaître les expressions faciales de plusieurs partenaires sociaux qui interagissent avec le dispositif robotique sans aucune supervision extérieure et sans aucun à priori vis à vis des connaissances que le "bébé" robot doit avoir à sa disposition (pas de modèle du visage, pas de reconnaissance de visage).

La figure 4-12 montre l'influence de la vitesse d'apprentissage du groupe de neurone VF (apprentissage des vues locales) pour l'ensemble de l'architecture. Nous constatons que plus cette vitesse d'adaptation est grande moins les performances seront élevées car les anciens prototypes appris seront trop modifiés et n'appartiendront plus à la même classe. Au contraire moins la vitesse d'adaptation est grande plus le système aura la chance de trouver des prototypes moyens stables.

human's facial expressions	robot's facial expressions: sadness	neutral face	joy	anger	surprised
sadness	**27%**	18%	23%	24%	8%
neutral face	27%	**26%**	23%	19%	5%
joy	18%	9%	**62%**	9%	2%
anger	23%	11%	10%	**52%**	4%
surprised	26%	12%	13%	14%	**35%**

a)

human's facial expressions	robot's facial expressions: sadness	neutral face	joy	anger	surprised
sadness	4%	18%	20%	**38%**	21%
neutral face	7%	16%	27%	**36%**	15%
joy	0.08%	9%	**70%**	18%	4%
anger	0%	0.08%	14%	**80%**	2%
surprised	6%	14%	23%	13%	**44%**

b)

human's facial expressions	robot's facial expressions: sadness	neutral face	joy	anger	surprised
sadness	4%	32%	3%	**43%**	17%
neutral face	7%	**56%**	6%	26%	5%
joy	4%	17%	**65%**	14%	0%
anger	3%	11%	10%	**73%**	3%
surprised	5%	33%	6%	9%	**47%**

c)

Fig. 4-14 – Test de généralisation du système obtenue durant une interaction naturel avec le robot: Après que 20 personnes ont interagit avec la tête de robot durant la phase d'apprentissage (les partenaires doivent imiter le dispositif robotique), l'on inverse les rôles et le robot doit imiter de nouvelles personnes avec lesquelles il n'a jamais interagit durant l'apprentissage. Ces tableaux montrent le taux de succès ans les taux de confusions pour chaque expression faciale ainsi que le visage neutre. a) montre les taux de confusions lors de l'utilisation de la transformé log-polaire, b) montre les taux de confusions lors de l'utilisation des filtres de gabor et c) montre les taux de confusion lors de l'utilisation de la fusion de ces deux caractéristiques. Ces statiques sont obtenues grâce à la méthode de cross validation.

La figure 4-13 montre qu'une interaction avec le dispositif robotique durant 2 minutes est suffisante pour apprendre les expressions faciales. Cet apprentissage incrémental est robuste bien que le nombre d'individus interagissant avec la tête de robot augmente conjointement avec la variabilité expressive des individus (les partenaires humains n'expriment pas identiquement toutes les expressions faciales, par exemple la tristesse, voir figure 4-16). Ce résultat renforce l'idée que le nombre d'individus interagissant avec le dispositif n'affecte pas les résultats de la reconnaissance des expressions faciales. Ces résultats suggèrent également que plus on augmente le

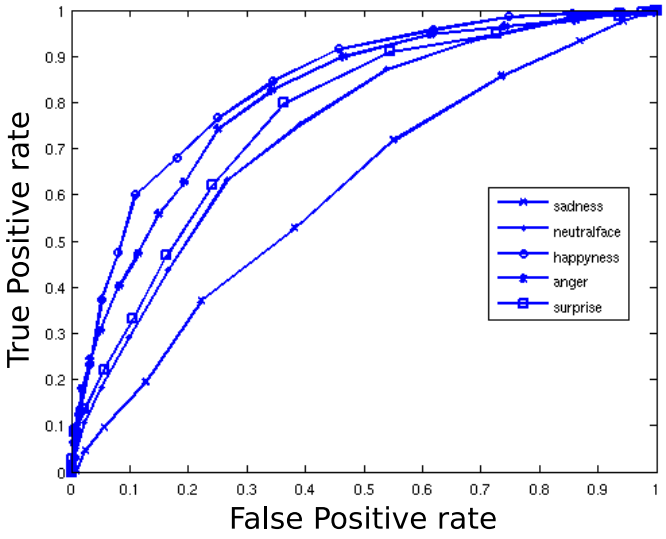


Fig. 4-15 – Taux de faux positifs et taux de vrai positif de chaque expression faciale (surprise, joie, colère, surprise et visage neutre) avec la fusion des informations visuelles (transformée log-polaire et filtres de Gabor). Dans ce cas de figure, nous ne pouvons plus utiliser une compétition dure avec les mécanismes de compétition du Winner Take All mais nous utilisons une fonction à seuil qui active tous les neurones au dessus du seuil.

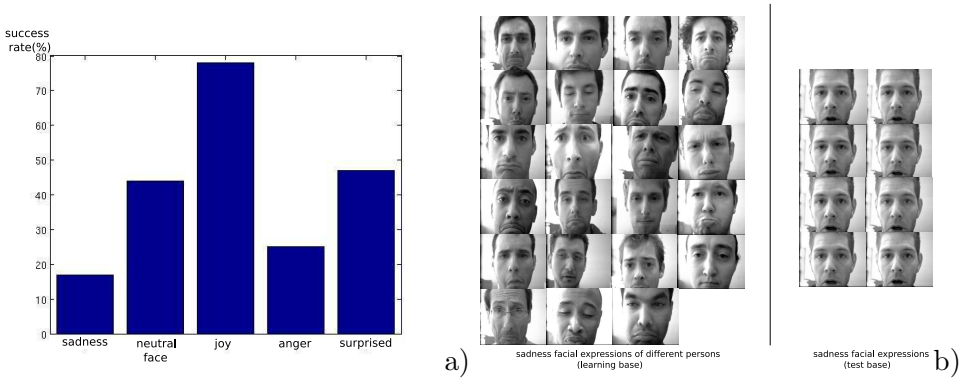


Fig. 4-16 – Etude d'un sujet spécifique. a) Le taux de succès pour chaque expression faciale pour un individu A, ces résultats sont obtenues durant une interaction naturelle avec le dispositif robotique. Le protocole expérimental est le suivant: 20 personnes interagissent avec le robot durant l'apprentissage (2 minutes d'interaction par personne) ensuite la tête robotique doit imiter l'individu A qui n'a encore jamais interagi avec le dispositif, autrement dit les expressions faciales de l'individu A n'ont pas été apprises par le robot. Les résultats mettent en évidence que les expressions de joie, de surprise et du visage neutre sont bien reconnues par le système contrairement à la tristesse. b) Plusieurs exemples d'expression de tristesse que les différents individus expriment, nous remarquons que les individus expriment la tristesse de différentes manières, d'un individu à l'autre la tristesse n'est pas réellement identique.

nombre d'individus durant la phase d'apprentissage plus on aura de chances d'avoir des résultats intéressants en généralisation sur de nouveaux individus.

La figure 4-14 et 4-15 montre que notre modèle est capable de généraliser à des individus

qui ne sont pas présents durant la phase d'apprentissage. Les taux de succès pour les différents traitements visuels sont très bons pour la joie et la colère, corrects pour la surprise et le visage neutre, mauvais pour la tristesse.

Nous pouvons remarquer dès à présent que la fusion des primitives visuelles apporte une très nette amélioration. La moyenne de la reconnaissance de l'expression faciale est de 40.4% lors de l'utilisation de la transformée log-polaire uniquement, de 42.8% lors de l'utilisation des filtres de Gabor et de 49% lors de la fusion de ces deux caractéristiques. Ces résultats montrent que notre modèle a la capacité à généraliser sur de nouveaux individus tout en sachant que les différents individus sont positionnés à des distances différentes, dans des conditions de luminosité différentes et sans aucun contrôle sur l'expressivité de l'humain (on n'élimine pas les humains qui ne "s'expriment" pas comme les autres).

En observant la base d'apprentissage, on remarque assez aisément que les individus ont des difficultés à exprimer certaines expressions faciales. L'absence d'une réelle mise en situation fait que le partenaire humain ne produit pas vraiment l'expression désirée. La tristesse par exemple est difficile à mimer sans contexte réel. L'analyse de la base d'apprentissage montre qu'il y a autant d'expressions faciales de tristesse que d'individus (voir figure 4.16b).

Nous supposons que certains partenaires éprouvent des difficultés à exprimer à la demande des expressions émotionnelles devant notre tête robotique. Pour justifier notre hypothèse, nous demandons à des humains non-experts d'annoter les images de la base de donnée (figure 4.17). Nous constatons très clairement que les humains non-experts ont énormément de difficultés à reconnaître certaines expressions des sujets enregistrés dans des sessions d'imitation de la tête robotique, en particulier l'expression de tristesse.

Pour vérifier l'impact de ces variations, nous avons enlevé de la base de données ces expressions faciales (nettoyage la base d'apprentissage). Nous avons demandé à 7 personnes d'annoter et de supprimer les expressions qui ne correspondent pour eux à aucune des catégories expressives (tristesse, joie, colère, surprise et visage neutre). Seules les images d'expressions faciales annotées de manière identique par les sujets chargés de la correction de la base sont conservées. Avec cette nouvelle base de donné, les résultats sont améliorés de 12%. Toutes ces expériences mettent en évidence le fait que travailler dans des environnements naturels, n'est pas aussi trivial que de travailler avec des bases de données déjà fabriquées par des experts où toutes les images sont labellisées correctement.

4.3.5 Contraintes d'une interaction en ligne

Néanmoins, l'apprentissage en ligne implique des problèmes dus à la dynamique de l'interaction entre l'homme et le robot. Il semble clair que la réaction à un stimulus n'est pas immédiate. Le temps de réaction de l'humain vis à vis des réactions expressives du robot n'est pas instantané. Un certain temps d'adaptation est indispensable à la perception visuelle de l'expression, à sa reconnaissance et à son imitation (Fig. 4.18a). Selon les travaux de Simon Thorpe (Thorpe et al., 1996), 150 ms sont requises pour la reconnaissance d'un objet ou la discrimination animal/non animal. Ce temps peut être inférieur ou supérieur à 150 ms si l'on imagine que ce temps de traitement de la reconnaissance d'expressions faciales fait intervenir des processus plus ou moins rapides que celle de la reconnaissance d'un objet. Le temps minimal T d'une boucle d'interaction entre l'homme et le robot est la somme de t_1 correspondant au délai du robot pour former l'expression faciale complète, t_2 correspondant au délai de l'humain pour reconnaître l'expression de la tête expressive et t_3 correspondant au délai du sujet humain pour mimer l'expression reconnue

human's facial expressions	facial expressions annotated by no-expert humans				
	sadness	neutral face	joy	anger	surprised
sadness	**45%** ± 2.2	10% ± 1.1	0% ± 0	19% ± 0.55	26% ± 1.6
neutral face	4% ± 0	**89%** ± 0.36	3% ± 0.13	1% ± 0	3% ± 0.13
joy	0% ± 0	6% ± 0.10	**90%** ± 0.43	1% ± 0	3% ± 0.19
anger	9% ± 0.24	15% ± 3.86	10% ± 0	**46%** ± 2.46	20% ± 1.43
surprised	2% ± 0.12	12% ± 0.78	5% ± 0	6% ± 0.19	**75%** ± 0.64

Fig. 4-17 – Humains non-experts annotant notre base de donnée d'humain imitant la tête expressive (choix forcé). Ce tableau montre le taux de confusion et l'écart type de chaque expression faciale. Le protocole est constitué de deux phases. La première phase fait intervenir un premier groupe d'individus devant imiter le robot expressif, leurs images sont enregistrées pour constituer une base de donnée d'expressions faciales. La deuxième phase fait intervenir un second groupe d'individus naïfs devant annoter la base d'images d'expressions faciales du premier groupe d'individus.

($T = t_1 + t_2 + t_3$). Lorsque le robot est seulement un automate produisant les expressions faciales les unes à la suite des autres, nous mesurons un temps T d'environ 800 millisecondes pour des sujet experts (individus ayant connaissance du dispositif robotique expressif) et un temps de 1.6 secondes pour des sujets naïfs (individu n'ayant jamais interagi avec la tête expressive). Ce délai perturbe bien évidement le processus d'apprentissage. Ce délai introduit une forte contrainte dûe à cette dynamique d'interaction car les premières images vues par le robot et qui vont être apprises pour l'expression courante, appartiendront à l'expression précédente (figure 4-18). Au cours de ce jeu d'imitation, nous aurons forcément un désapprentissage des premières images donc une perte de la qualité d'apprentissage. Pour y remédier, le temps de présentation d'une expression faciale doit être suffisamment élevé pour pouvoir négliger les premières images appartenant à l'expression précédente. La figure 4-18.b montre l'activtée neuronale durant la phase où le robot expressif reproduit les expressions du partenaire humain. Dans cette séquence de reproduction, nous remarquons que la tête robotique réagit correctement à toutes les expressions excepté le visage neutre. Cette figure met en évidence l'effet de la mémoire à court terme sur les images. L'activité à la sortie est lissée par l'utilisation de cette dernière. Cette mémoire a deux effets indissociables: le robot aura un temps de réaction un peu plus grand mais par conséquent il se stabilisera plus nettement sur un bassin attractif expressif (il ne passera pas d'une expression à l'autre).

Dans ce chapitre, nous avons montré que le dispositif robotique constitué d'une simple tête expressive est capable d'apprendre et reconnaître de manière autonome les expressions faciales si et seulement si durant l'apprentissage le "bébé" robot réalise des expressions émotionnelles et le partenaire social l'imite. Le modèle computationnel associe simplement les vues locales prises sur le visage de l'humain avec son propre état interne: association de ce que le robot fait avec ce qu'il voit. Notre modèle met en évidence que les expressions faciales peuvent être apprises sans détecteur de visage ad-hoc précédent l'analyse des expressions et par conséquent sans modèle de visage. Malgré tout, le robot peut apprendre des associations non pertinentes si personne n'interagit avec lui car le robot apprend sans distinction toutes les vues locales. L'apprentissage

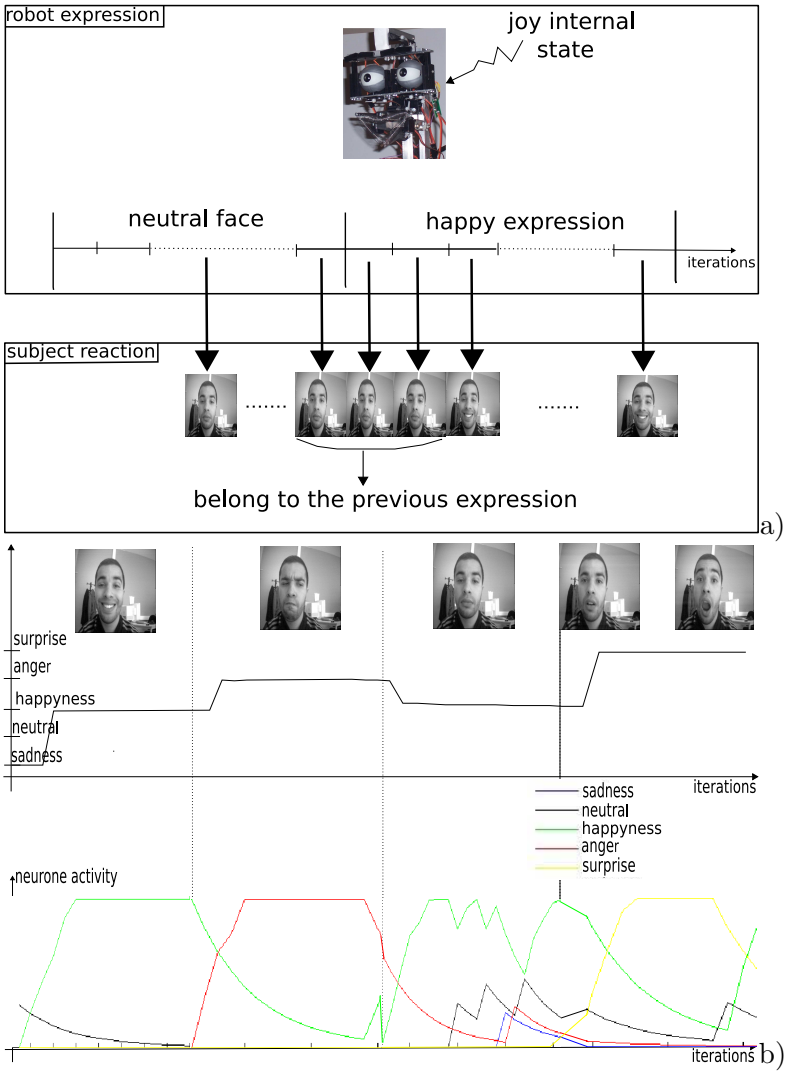


Fig. 4-18 – a) Phase de transition entre l'expression de l'humain et l'expression faciale du robot durant le jeu d'imitation (phase d'apprentissage: l'humain imite). b) Activité temporelle des neurones associée aux différentes expressions faciales (phase de test: le robot doit imiter).

devrait être modulé en fonction de la présence ou de l'absence d'un partenaire humain. Dans le chapitre 5, nous nous focaliserons sur l'ajout d'un mécanisme permettant de moduler la vitesse d'apprentissage en fonction de la situation interaction/non interaction et par la même de faire la discrimination entre un visage d'un non visage.

4.4 Discussion et Conclusion

Beaucoup de recherches se sont focalisées sur la construction de systèmes robustes pour la reconnaissance des expressions faciales. A notre connaissance, il n'existe aucun système s'intéressant à comprendre comment cette capacité peut se développer de manière autonome. Pour les architectures traditionnelles, l'accent est mis sur comment sélectionner le meilleur ensemble de caractéristiques et comment obtenir le meilleur algorithme. Très souvent, ces archi-

tectures sont performantes sur des bases de données déjà construites mais les résultats sur des images réelles sont très peu étudiés.

Certaines méthodes sont basées sur l'analyse en composante principale (PCA) par exemple le LLE (Locally Linear Embedding) (Liang et al., 2005) qui réalise une réduction des vecteurs d'entrés. Des architectures neuronales ont également été développées pour la reconnaissance des expressions faciales, par exemple le réseau de Franco and Treves (Franco and Treves, 2001). Ce dernier utilise un réseau de neurones multicouches utilisant une règle d'apprentissage supervisée classique. Cependant, l'architecte du réseau détermine le nombre de neurones associés aux différentes expressions selon leur complexité. D'autres méthodes sont basées sur des modèles du visages comme par exemple les modèles d'appartenance (B. Abboud, 2004). Yu (Yu and Bhanu, 2006) utilise un support vecteur machine (SVM) pour catégoriser les expressions, son intérêt étant d'obtenir les meilleurs frontières de séparation. Wiskott (Wiskott, 1991) utilise des ondelettes de gabor pour le codage des caractéristiques faciales appelées "jets", ces caractéristiques sont insérées dans un graphe labellisé où les noeuds sont des "jets" et les liens entre les différents noeuds sont des distances (distances entre la bouche et un oeil par exemple). La reconnaissance se traduit par une comparaison de graphe ("graphe matching"). Toutes ces techniques et beaucoup d'autres encore utilisent un apprentissage hors ligne ainsi qu'un grand nombre d'à priori pour améliorer les performances. Notons également que toutes ces méthodes ont besoin d'accéder à l'ensemble de la base d'apprentissage ce qui rend impossible la comparaison avec un processus développemental. Ces modèles ne peuvent pas mimer des processus biologiques par leurs manques d'adaptation aux différents environnements. Soulignons finalement que ces architectures ne proposent pas d'explications sur comment un nourrisson peut apprendre des expressions émotionnelles.

D'autres laboratoires comme le MIT se rapprochent davantage de nos travaux de recherches. Cynthia Breazeal (Breazeal et al., 2005), la créatrice de Kizmet, possède une tête robotisée pouvant reproduire les expressions faciales d'un humain (Figure 4-19). Ce robot ressemble beaucoup à une peluche capable de mimer un ensemble d'expressions. Par l'intermédiaire d'un jeu d'imitation entre l'humain et la peluche expressive, Kismet apprend à mimer les expressions faciales du partenaire humain. Cependant dans ces travaux de forts apriori sont introduits rendant l'architecture plus fiable. Notamment, l'aspect visage est traité grâce à un module qui lui est dédié. Les zones importantes à analyser tel que les yeux, la bouche, les sourcils, le nez sont spécifiées par une connaissance apriori du système. Ces hypothèses rendent le modèle peu autonome, moins adaptative et plus rigide aux interactions naturelles. Contrairement à notre modèle qui a la capacité d'apprendre sans connaissance du monde dans lequel il interagit et agit. Dans notre perspective développementale, c'est grâce à ce type d'interaction que le robot sera capable de découvrir son environnement. Il développera par la suite des capacités de reconnaissances de l'autres (détection d'un visage, d'un oeil, d'une bouche ...)

D'autres tête robotiques comme le robot Einstein (Wu et al., 2009) explorent le processus d'apprentissage auto-guidé (figure 4-20). Leurs travaux s'intéressent à comprendre comment apprendre à faire les expressions faciales. Ils disposent d'une tête de robot à 31 degrés de liberté où les paramètres des moteurs expressifs sont appris en utilisant un feedback venant de la reconnaissance temps réel d'expressions faciales.

Notre modèle nous a permis de montrer qu'il est possible d'apprendre et de reconnaitre des expressions faciales de manière autonome. L'apprenant doit produire les expressions et l'enseignant doit l'imiter. Ce petit jeu d'imitation est suffisant pour qu'après 2 minutes d'apprentissage

Fig. 4-19 – Kismet. Le robot développé au MIT dans l'équipe de C. Breazeal.

Fig. 4-20 – Robot Einstein (Wu et al., 2009)

le robot soit capable de reproduire les expressions de l'expérimentateur. Sans l'empathie[4] de

4. ou sa capacité à résoner avec l'expression du robot

l'expérimentateur, rien n'est possible. L'empathie ou la résonance émotionnelle sont la clé de voûte de notre paradigme expérimentale.

Les difficultés de notre approche sont multiples contrairement aux autres méthodes. La question que nous venons de traiter dans ce chapitre n'est pas centrale pour les méthodes concurrentes. La difficulté de notre approche est due à la contrainte d'autonomie de l'apprentissage car la solution implique un jeu d'imitation en ligne non contrôlé et donc une interaction naturelle. Cette interaction peut être vue à la fois sur le court terme comme une contrainte mais en réalité sur le long terme ce jeu d'imitation est un point fort permettant de s'adapter à l'environnement. Un point important est que les expressions faciales des différents expérimentateurs sont relativement diversifiées (certains manquant d'expressivité Figure 4-16b). Contrairement aux autres méthodes[5], la base de données pour l'apprentissage n'est pas contrôlée.

Notre paradigme expérimental peut paraître pauvre car il demande à l'humain d'imiter la tête expressive sans contexte environnemental réel, sans situation réelle d'interaction. D'ou la difficulté de certains expérimentateurs à afficher certaines expressions comme la surprise ou la tristesse. L'empathie reste essentielle pour garder une bonne interaction. Pour justifier cet argument, nous avons demandé à des humains naïfs d'annoter la base de donné pour obtenir leurs taux de réussite. Le taux de reconnaissance des humains montrent qu'ils ont également des difficultés pour reconnaître les expressions faciales (Figure 4-14, 4-17). Nos résultats soulignent que les taux de reconnaissance des humains et du robot sont assez similaires pour certaines expressions comme la tristesse.

Le dernier point que nous discuterons est un problème récurrent aux interactions Homme/Robot. L'apprentissage en ligne induit des problèmes dus aux temps de réaction de l'humain. Ce délai perturbe la phase d'apprentissage car lorsque le robot change d'expression faciale, les premières images capturées correspondent toujours à l'ancienne expression. A chaque fois que le robot exprime une nouvelle expression, les premières images apprises correspondent à l'expression précédente. Il est crucial de considérer ce problème pour l'amélioration des résultats car dans les systèmes classiques d'images, ce problème est évité puisque les bases de données sont labellisées par des experts.

La stratégie que nous employons dans ce chapitre basée exclusivement sur une reconnaissance locale de certaines zones de l'image correspond à une analyse séquentielle et précise de l'image. Cette stratégie peut être vue comme une implémentation simple de la voie thalamo-cortico-amygdala chez les mammifères (LeDoux, 1996). Dans des travaux précédents (Gaussier et al., 2007), nous avions implémenté, une architecture simple et rapide basée sur l'analyse de l'image globale. Cette dernière pourrait correspondre à une voie de traitement courte, la voie thalamo-amygdala (Papez, 1937; LeDoux, 1996) impliquée dans les réactions émotionnelles rapides.

Dans une perspectif développementale, il serait intéressant de vérifier le couplage de ces deux approches (Figure 4-21). Autrement dit, l'influence des deux stratégies car le réseau thalamo-cortico-amygdala peut être une manière de contrôler l'apprentissage du réseau thalamo-amygdala permettant à la fois une reconnaissance rapide et robustesse des expressions faciales.

5. Habituellement, les bases de données sont construites en cadrant bien les visages, en faisant attention aux conditions d'éclairages et aux différents types de bruits qui pourraient perturber l'apprentissage. Un soin tout particulier est porté à la qualité des labels associés aux visages

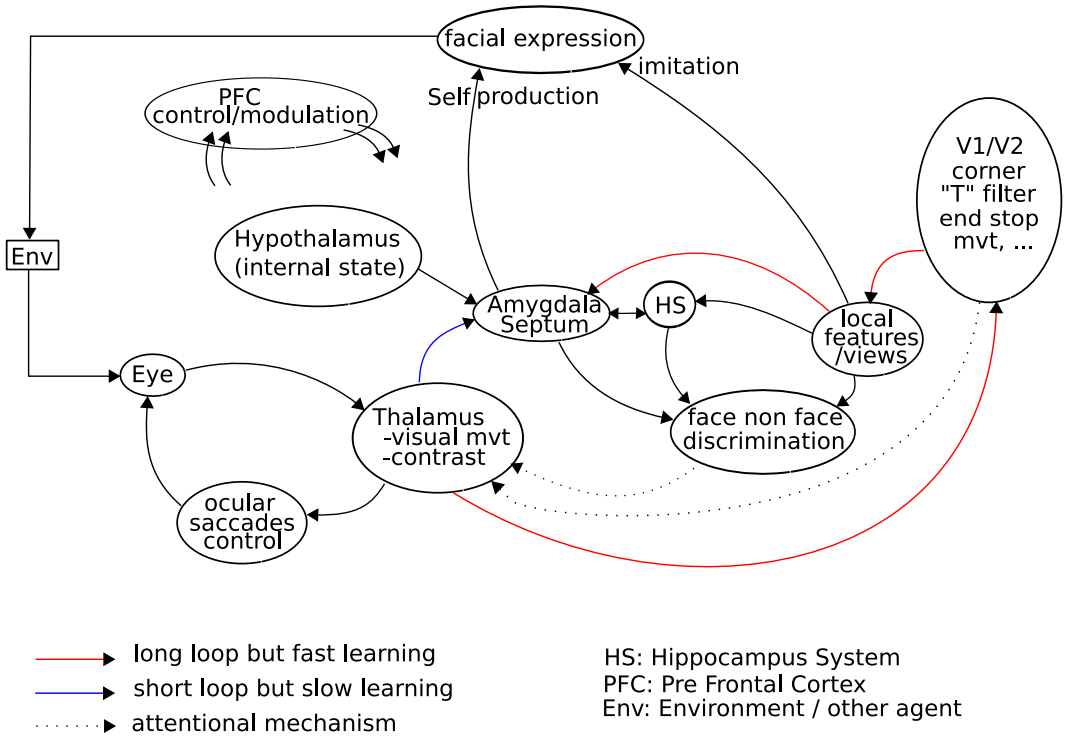


Fig. 4-21 – Circuit neuronale développé: la voie thalamo-cortico-amygdala et la voie thalamo-amygdala dans le cerveau d'un mammifère.

Chapitre 5

Apprentissage autonome de la détection de visage

5.1 Introduction

Les applications de l'analyse de visages concernent aujourd'hui un grand nombre de domaines allant de la biométrie à l'indexation d'images en passant par les nouvelles interfaces Homme-Machine. L'analyse de visages est un vaste domaine de recherche où plusieurs questions théoriques restent encore sans réponses définitives. Dans le domaine des neurosciences les recherches s'accentuent sur les différents mécanismes mise en oeuvre pour la perception des visages sachant que les nouveaux nés ont des préférences pour les stimuli de visage. Dans le domaine des sciences informatiques les thématiques sont classiques:

- la détection: un visage est-il présent dans l'image?
- la localisation: où se situe le visage dans l'image?
- la reconnaissance: A qui appartient le visage dans l'image?

Cependant, les systèmes informatiques manquent de robustesse dans les environnements complexes et naturels en raison de difficultés multiples. Les visages sont des structures déformables (l'expression faciale, l'orientation du visage, structure, ...) induisant un éventaille de paramètres à considérer. Des paramètres extérieurs sont également contraignants (luminosité, ombres, occultations).

De manière générale la détection de visage est la première étape dans des processus comme la reconnaissance des expressions faciales, la reconnaissance de l'identité, la détection de motifs faciaux ... Cependant, il est communément admis que sa performance influence grandement le processus de reconnaissance. Nos travaux s'intéressent à répondre à plusieurs questions qui sont à la frontière de plusieurs domaines:

- Quels sont les mécanismes à introduire pour développer la capacité de détection de visage?
- Comment un "bébé robot" peut-il développer cette capacité de manière autonome?
- Les interactions émotionnelles peuvent-elles résoudre une partie du problème?

Dans ce chapitre, nous présenterons un bref état de l'art concernant la détection de visage mettant en évidence certaines difficultés. Nous verrons ensuite comment les interactions émotionnelles peuvent influencer l'apprentissage autonome de visage/ non visage par la construction d'un signal implicite émergeant de l'interaction.

5.2 Etat de l'art

De nombreuses méthodes de détection de visage existent s'intéressant exclusivement à l'amélioration des performances. Cependant, aucune recherche à notre connaissance ne porte sur les mécanismes à introduire pour une meilleure compréhension d'un système aussi efficace que l'être humain. Nous avons choisi de diviser les méthodes existantes en trois grandes familles: modèles avec connaissance de la structure du visage, modèles globales, modèles hybrides. Cependant, ce découpage n'est pas unique, on aurait pu aussi choisir de distinguer les modèles globaux et locaux.

5.2.1 Introduction d'information a priori sur le visage

Ces techniques sont basées sur la connaissance de la structure d'un visage à savoir les éléments qui le constituent. Plusieurs méthodes concurrentes existent ayant toutes le même point commun : l'ajout d'une connaissance extérieure et experte au problème à traiter.

Certaines approches considèrent les parties caractéristiques comme le nez, la bouche et les yeux pour ensuite appliquer un système à base de règles sur ces caractéristiques. Notamment (Kotropoulos and Pitas, 1997) utilise cette méthode où la localisation des différentes parties du visage est effectuée par l'intermédiaire de la méthode de Kanade (Kanade, 1973) (les caractéristiques sont trouvées à l'aide de projections horizontales et verticales). L'inconvénient majeur de cette méthode est son manque de robustesse face à des fonds texturés.

L'approche basée sur des "template-matching" est peut être la plus utilisée pour la détection de visage. L'idée sous-jacente est de calculer la corrélation entre l'image originale et le ou les "templates" (image(s) prototypique(s)). Cette méthode très intuitive manque de robustesse vis à vis des changements de luminosité, des changements d'échelle ainsi que de la pose du visage (orientation). Pour palier ces faiblesses (Sinha, 1994) introduit un ensemble d'invariants décrivant le modèle du visage ainsi qu'une relation entre les différentes éléments le constituant (relation basée sur un rapport de luminosité). D'autres recherches se sont intéressées à l'utilisation de templates paramètriques (Yuille et al., 1992).

Viola et Jones (Viola and Jones, 2001; Viola and Jones, 2004) proposent un détecteur ayant des capacités temps réel et de très bonnes performances. Il est basé sur l'utilisation de classifieurs (filtres de Haar) montés en cascade (algorithme Adaboost), permettant d'obtenir un classifieur composé de structures élémentaires capable d'éliminer au fur et à mesure les zones de l'image non significatives. D'autre part, ce modèle possède une généricité permettant d'appliquer le principe à d'autres objets. Certaines études se sont intéressées à l'amélioration de cette technique (invariance à la rotation par exemple) (c. Huang et al., 2004; Li and Zhang, 2004; Kim et al., 2005).

Enfin, un grand nombre de modèles basés exclusivement sur la couleur de la peau ont pu être développés (Vezhnevets et al., 2003; Yang et al., 1997; Chai and Ngan, 1999; Garcia and Tziritas, 1999). Ces méthodes sont rapides et restent efficaces puisqu'elles sont robustes aux rotations, changement d'échelle et occlusions. Cependant, le choix de l'espace couleur reste un problème important que ce soit en RGB, HSL ... lorsque les changements d'éclairage sont trop importants. Certains modèles neuronaux utilisent deux réseaux de Kohonen pour l'apprentissage de la peau et de la non peau (Brown et al., 2001).

5.2.2 Les modèles globaux

Les méthodes globales sont rapides à mettre en oeuvre et les temps de calculs sont généralement réduits ce qui en fait des candidats privilégiés pour des applications temps réelles et embarquées. Cependant, elles semblent être sensibles aux conditions d'éclairage, de pose et d'expression faciale.

Ces modèles se basent soit sur des méthodes statistiques soit sur des mécanismes d'apprentissages neuronaux. L'apprentissage est toujours réalisé hors ligne avec un ensemble d'individus suffisamment représentatif de la variabilité des visages. L'une des techniques les plus connues est celle des "eigenfaces" (Turk and Pentland, 1991) consistant à projeter l'image dans un espace de plus petite dimension et à calculer la distance entre l'image et sa projection. Plus la distance est grande moins l'image correspond à un visage.

D'autres méthodes basées sur des réseaux de neurones ont vu également le jour. Rowley (Rowley et al., 1998) propose un modèle neuronal constitué de deux couches: une première couche réalise un simple pré-traitement normalisant les images d'entrées (égalisation d'histogramme). Une seconde couche étant un réseau de neurone balayant l'image pour détecter la présence ou non de visage. Une analyse multi-résolution est intégrée permettant la détection de visage de différentes tailles. D'autres modèles neuronaux existent ainsi (Garcia and Delakis, 2004; Farrugia et al., 2009) utilisent un réseau de neurones multicouches convolutionel où chaque couche traite une résolution, l'extraction des caractéristiques est réalisée grâce à un produit de convolution. A l'inverse (Feraud et al., 2001) utilisent un réseau de neurones multicouches auto-associatif entièrement connecté. Cependant, les résultats montrent que la méthode de (Garcia and Delakis, 2004) semble plus robuste que (Feraud et al., 2001) sur la base de donnée CMU. (Yang et al., 2003) implémentent un modèle de localisation et de reconnaissance de visages en temps réel avec un réseau de neurones de type RBF. Leur modèle a été implanté sur des architectures électroniques (ZISC et FPGA) soulignant une adéquation alogorithme-architecture.

5.2.3 Les modèles hybrides

Les modèles hybrides essaient de tirer partie à la fois des approches globales et locales. (Yang and Huang, 1994) proposent un modèle hiérarchique de détection de visage. Il commence par une analyse basse résolution induisant la sélection d'un ensemble de candidats. La sélection se fait par l'intermédiaire d'un système à base de règles permettant de trouver les régions homogènes dans l'image correspondant peut-être à un visage. Une fois cette présélection effectuée, les résolutions supérieures (hautes fréquences) sont utilisées pour chercher l'existence de traits faciaux. L'idée de ce modèle repose sur le traitement d'une information globale basée sur l'analyse basse fréquence pour se diriger vers des traitements locaux pour obtenir des informations plus détaillées.

Des modèles moins élégants sont apparus (Silva et al., 1995) cherchant des caractéristiques du haut du visage pour ensuite utiliser un template flexible. Cependant, ces techniques manquent de robustesse lorsque le visage possède des lunettes ou si les cheveux recouvrent le front.

D'autres méthodes (Yow and Cipolla, 1996) commencent par chercher dans l'image des caractéristiques locales (points caractéristiques) qui sont généralement des zones de courbures pour ensuite les regrouper formant ainsi un concept de plus haut niveau.

5.2.4 Discussion

Malgré, les efforts de toutes ces recherches, il n'existe toujours pas de modèle capable d'intégrer toutes les difficultés. Certaines sont spécialisées pour l'invariance à la luminosité, d'autres pour les changements d'échelles, d'autres pour les occlusions et d'autres encore pour les différentes orientations ou expressivités du visage. A notre connaissance, il n'existe encore aucun modèle capable de résoudre toutes ces contraintes simultanément en environnement réel même si certains essayent de fusionner ces différentes techniques permettant ainsi d'obtenir de meilleurs résultats (Belaroussi et al., 2005).

Cependant ces travaux ne s'intéressent pas à construire un modèle adaptatif capable d'apprendre en ligne. Le modèle que nous allons présenter posséde cette capacité. Le prix à payer est l'utilisation de concepts s'intégrant dans un processus développemental.

5.3 Prédiction du rythme de l'interaction

Jusqu'à présent, nous avons tenté de répondre à la question: comment un robot peut-il apprendre des expressions faciales émotionnelles de façon autonome? La condition nécessaire pour réaliser cette apprentissage, est le mimétisme de l'expérimentateur. Dans ces conditions, l'expérimentateur est l'élément indispensable pour le développement du robot. Par conséquent, le robot est capable d'associer ce qu'il voit avec ce qu'il fait. Néanmoins, le robot ne posséde encore aucune représentation de qui est l'autre.

Dans ce chapitre, nous nous intéresserons à faire la distinction entre un individu "social" (individu interagissant avec lui) ou un "objet" inanimé. En d'autres termes, nous proposerons de construire un modèle permettant à notre robot d'acquérir la faculté de détecter un congénère. Dans notre cadre expérimental, nous profiterons de l'interaction émotionnelle pour réaliser une discrimination visage/non visage. La question de l'autonomie du robot passe forcément par la construction d'un signal permettant de réaliser cette discrimination. La construction de ce signal ajoute un nouveau noyau à l'architecture, montrant ainsi que la détection de visage semble plus complexe au moins en termes d'éléments élémentaires à introduire.

Dans un premier temps, nous proposons un modèle capable de construire ce signal de supervision émergeant d'une interaction entre 2 agents (Homme et Robot): modèle détectant une rythmicité entre un "apprenant" et un "enseignant". Dans un second temps, nous proposons une architecture couplant la reconnaissance des expressions faciales et la détection de la rythmicité pour faire émerger de la capacité à détecter un visage.

5.3.1 La rythmicité et son apprentissage

Dans cette partie, nous discuterons de la notion de rythme pouvant être définie comme la fréquence d'un événement périodique. Par exemple, le rythme cardiaque est un phénoméne physiologique périodique. La notion de rythme que nous emploierons, s'inscrira dans un cadre d'interaction où 2 agents communiquent. En d'autres termes, l'aspect "rythme d'interaction" signifie un échange périodique entre 2 agents.

Nous essayerons de faire émerger un rythme d'interaction entre le partenaire humain et le dispositif robotique. Nous nous inspirerons des études de neurobiologie, notamment des travaux réalisés sur l'hippocampe et le cervelet. Ces deux structures ont la capacité à encoder des séquences d'évènements ainsi que des timing d'événements.

– le cervelet: il est connu pour son rôle dans le contrôle moteur, et plus particulièrement pour

sa fonction d'apprentissage du timing des événements moteurs (Bullock et al., 1994). Une lésion des parties latérales du cervelet entraîne des difficultés au niveau de la coordination motrice.

– l'hippocampe: l'hippocampe est connu pour le rôle qu'il joue dans l'apprentissage de séquences, dans la mémoire, et dans le conditionnement. Certaines études menées sur la pathologie de la mémoire, tendent à montrer l'importance de l'hippocampe dans l'apprentissage et la détection d'événements ou de séquences nouvelles (Grossberg and Merrill, 1992). Mais il semble que l'hippocampe ne soit pas la structure qui mémorise pour du long terme.

5.3.2 Un réseau de neurones pour l'apprentissage du rythme

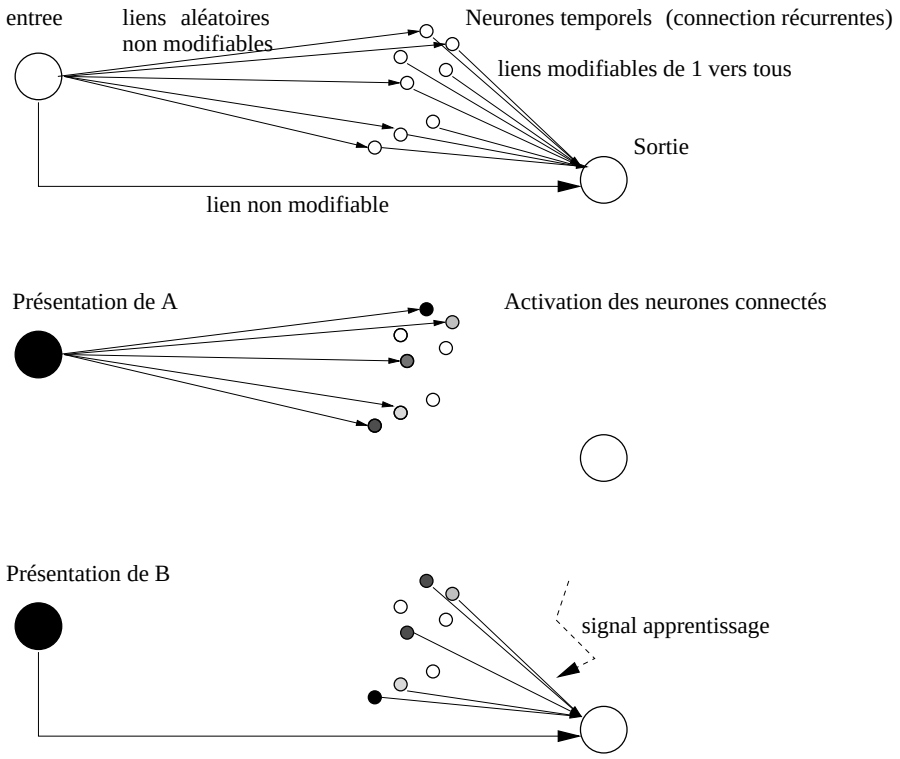


Fig. 5-1 – Illustration d'un réseau suffisant pour l'encodage d'un intervalle AB. L'apprentissage repose sur une règle associative et surtout sur le jeu des connexions aléatoires entre les groupes d'entrée et temporel, ainsi que les profils d'activité différents pour chaque neurone temporel.

Pour apprendre le rythme d'interaction (l'encodage de la temporalité), il suffit d'avoir un système neuronal qui soit capable d'apprendre l'intervalle de temps entre deux stimuli. Supposons deux stimuli distincts A et B émis à deux instants différents. Un réseau constitué de trois groupes de neurones différents sera capable d'apprendre la temporalité de l'activation A B (figure 5-1).

– Le premier groupe de ce réseau est un groupe d'entrée, dont le neurone réagit en présence du stimulus A ou B.

– Le deuxième groupe que nous appellerons groupe temporel est un groupe dont les neurones ont des propriétés temporelles.

– Le troisième groupe que nous appellerons groupe de sortie est relié au groupe temporel (connexion de un vers tous). Si ce groupe de sortie utilise le stimulus B comme un signal d'apprentissage, alors une simple loi d'apprentissage associant les activités des neurones du groupe temporel à l'activation du groupe de sortie permettra d'apprendre l'instant de présentation de B.

Des connexions entre le neurone d'entrée et le groupe temporel existent ce qui impliquent qu'une courte présentation du stimulus A engendrera l'activation d'une batterie de cellules du groupe temporel, cellules dont l'activité sera entretenue dans le temps en fonction de la force des connexions récurrentes: nous obtenons ainsi une mémoire à court terme du stimulus A.

Nous proposons un modèle de réseau de neurones capable d'apprendre un intervalle temporel basé sur une modélisation de l'hippocampe et du cervelet (Banquet et al., 1997; Andry et al., 2001; Moga et al., 2003). Ce système est un réseau de neurone composé de trois types de neurones ayant chacun des fonctionnalités particulières (Figure 5-2, 5-3).

– Des neurones d'entrée appelés GD (groupe d'entrée), leur fonction est essentiellement de dériver le signal d'entrée. Ces neurones reçoivent le flot d'information à apprendre (la séquence à apprendre).

– Des neurones temporels GT (groupe temporel). Ces neurones réagissent par bancs, et leurs activités correspondent à des réponses retardées selon des constantes de temps différentes.

– Des neurones de type prédictif, appelés GP (groupe de prédiction). Ces neurones reçoivent une information des neurones GD (le présent) et des neurones GT (le passé) et encodent l'instant de transition entre les deux événements.

Ce modèle repose sur le fait qu'un neurone de type GP peut apprendre, mais peut aussi prédire l'intervalle de temps écoulé entre l'activation de deux neurones de type GD en utilisant une batterie de cellules à réponse retardées (les neurones de types GT).

La sortie $D(t)$ d'un neurone GD s'active si la variation de son potentiel $P_D(t)$ entre deux instants dépasse un seuil ϵ_E (l'activité de GD dépend de la dérivée de son potentiel).

Une activation du neurone GD détermine une mise à zéro de chaque neurone GT grâce aux liaisons entre GD et GT. L'activation des neurones GD induit une modification des poids entre GP et GT à $t+1$ (Figure 5-1). Les neurones GT sont regroupés par lignes en batteries (des bancs de neurones). Chacune de ces lignes est reliée à un neurone GD dont l'activité déclenche les cellules de toute la ligne. Après chaque mise à zéro, par un neurone de type GD, les neurones de type GT ont une activité donnée par:

$$Act_{j,l}^{GT}(t) = \frac{1}{m_j} \cdot \exp - \frac{(t-m_j)^2}{2 \cdot \sigma_j} \tag{5.1}$$

où l correspond à l'indice de la cellule activée sur la ligne, j correspond à la ligne de la cellule activée, m_j est une constante de temps, σ_j son écart type et t le temps en milliseconde.

Les activités des cellules d'une ligne, se répartissent ainsi au cours du temps et représentent une trace sur plusieurs secondes de l'activité GD passée (le temps exact dépendant du nombre de cellules de la ligne et des constantes σ_j choisies).

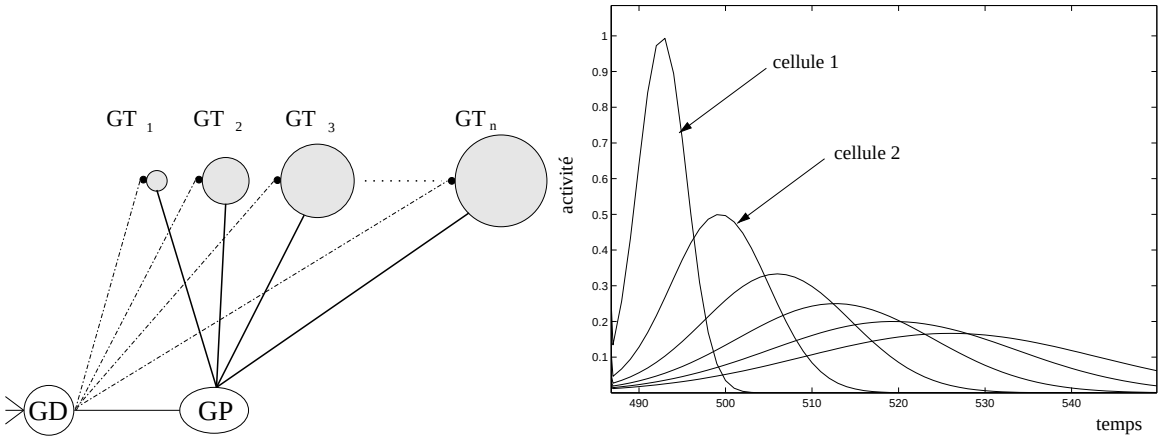


Fig. 5-2 – A Gauche : Détail des connexions entre un neurone GD, une batterie de neurones GT et un neurone GP. A droite : courbes d'activité de 6 cellules d'une ligne du groupe temporel. La ligne est activée à t=485 s, et chaque cellule décharge dans le temps, permettant de conserver une trace à court terme de l'événement (ici sur une minute environ)

Un neurone GP reçoit les connexions croisées d'une ligne de neurones GT et d'un neurone GD. L'information véhiculée par la ligne de neurones GT correspond au temps écoulé depuis l'apparition du dernier événement, tandis que l'information en provenance du neurone GD, correspondant à l'instant d'apparition du nouvel événement.

Le neurone GP somme les entrées relatives à ces deux types d'informations selon l'équation suivante:

$$Pot_{i,j}^{GP} = \sum_{l} W_{gp(i,j)}^{gt(j,l)} \cdot Act_{j,l}^{GT} + W_{gp(i,j)}^{gd(i)} \cdot Act_{j}^{GD} \tag{5.2}$$

$Act_{j,l}^{GT}$ est l'activité de la $l$ème cellule de la $j$ème ligne du groupe Temporel GT. $W_{gp(i,j)}^{gt(j,l)}$ est la valeur des poids des connexions entre les neurones temporels GT et la cellule du groupe GP. Act_{j}^{GD} est l'activité du $j$ème neurone GD. $W_{gp(i,j)}^{gd(i)}$ est la valeur des poids de la connexion entre GD et la cellule du groupe prédictif GP.

L'activation de GP est déterminée par la détection de la valeur maximale de son potentiel (impulsion émise au moment où la dérivée du potentiel repasse par zéro):

$$Act_{i,j}^{GP} = f_{GP}\left(Pot_{i,j}^{GP}\right) \tag{5.3}$$

$$avec\ f_{GP}\left(x\left(t\right)\right) = \begin{cases} 1 & si\ \frac{dx(t)}{dt} < 0\ et\ \frac{dx(t-1)}{dt} > 0 \\ 0 & sinon \end{cases} \tag{5.4}$$

Enfin, l'apprentissage de l'instant de transition est réalisé dans les poids des connexions reliant les neurones GT au neurone GP. Il s'agit d'un apprentissage en un coup. Il consiste à affecter aux poids des connexions l'activité normalisée des cellules GT seulement si l'une des cellules a été activée par le neurone GD.

$$W_{GP(i,j)}^{GT(j,l)} = \begin{cases} \frac{Act_{j,l}^{GT}}{\sum_{j,l}\left(Act_{j,l}^{GT}\right)^2} & si\ Act_{j}^{GT} \neq 0 \\ inchange & sinon \end{cases} \tag{5.5}$$

Ainsi, le rappel de la transition apprise (c'est à dire l'activation du neurone GP) ne pourra avoir lieu que si la somme des activités des cellules GT est égale à celle rencontrée lors de l'apprentissage (Figure 5-1).

5.3.3 Résultats

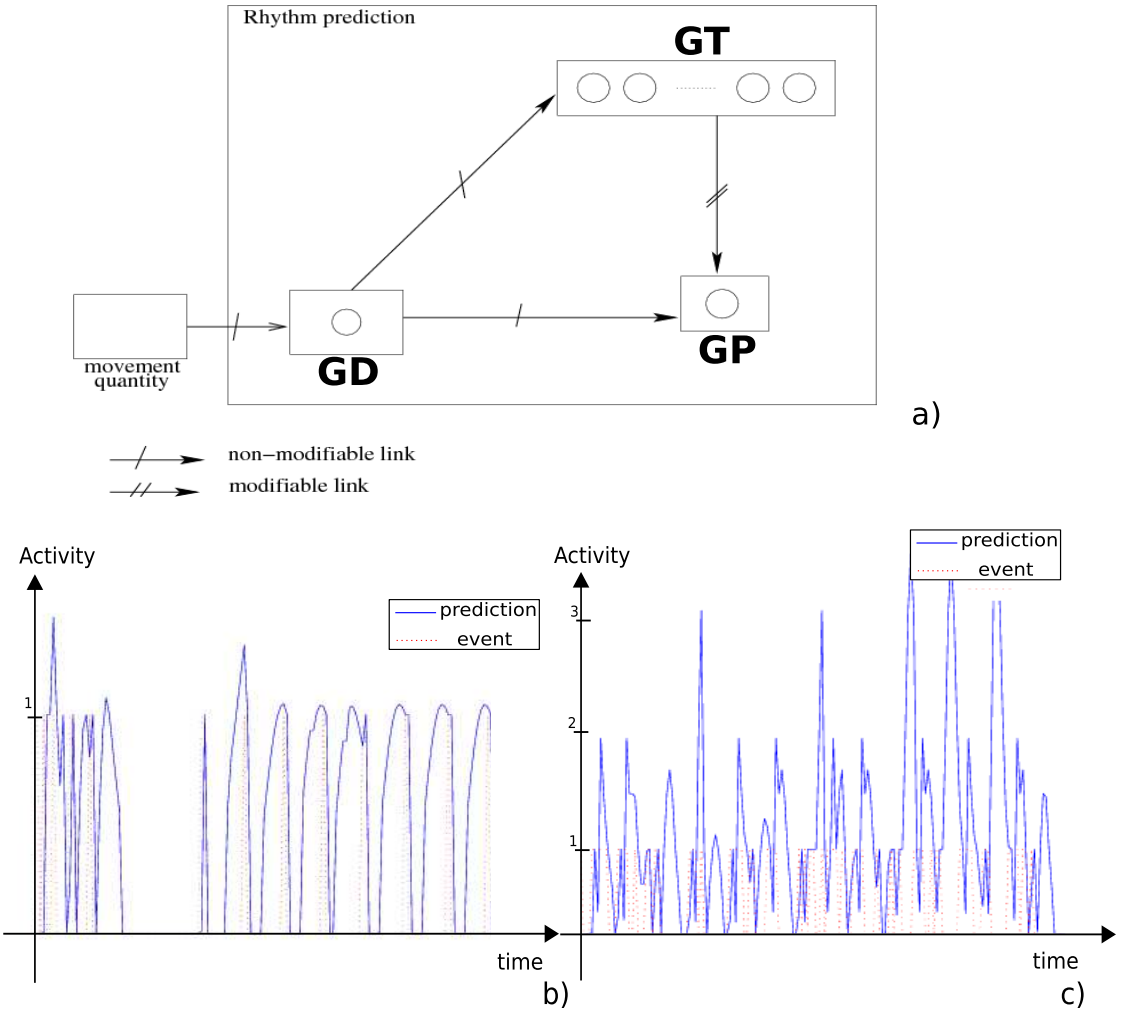


Fig. 5-3 – a) Modèle de prédiction du timing entre le sujet et le robot (prédiction du rythme de l'interaction). b) Activité des neurones GP et GD correspondant respectivement à la prédiction de timing et à l'événement réel. La prédiction et l'événement réel se superposent montrant ainsi une bonne prédiction dans le cas où le robot produit des expressions faciales et l'expérimentateur l'imite. c) Activité du neurone GP et GD dans le cas où le robot produit des expressions faciales et des mouvements incohérents sont réalisés devant le dispositif robotique. Dans ce cas, la prédiction de l'événement n'est pas possible.

Pour analyser le rythme de l'interaction, nous avons mis en place une expérience excessivement simple où le protocole expérimental est le suivant:

– Le robot produit des expressions faciales primaires (joie, colère, surprise, tristesse). Il exprime pendant 2 secondes une expression et passe par un visage neutre pendant 1 secondes

entre deux expressions faciales.

– L'expérimentateur se place devant le dispositif robotique à une distance d'interaction naturelle (une distance de bras).

– La seule consigne transmise à l'expérimentateur est d'imiter la tête expressive

Pour tester notre paradigme expérimental, nous avons réalisé l'expérience avec 10 personnes naïves. Ces personnes naïves sont des étudiants ou des personnels admistratifs n'ayant jamais interagit avec notre dispositif.

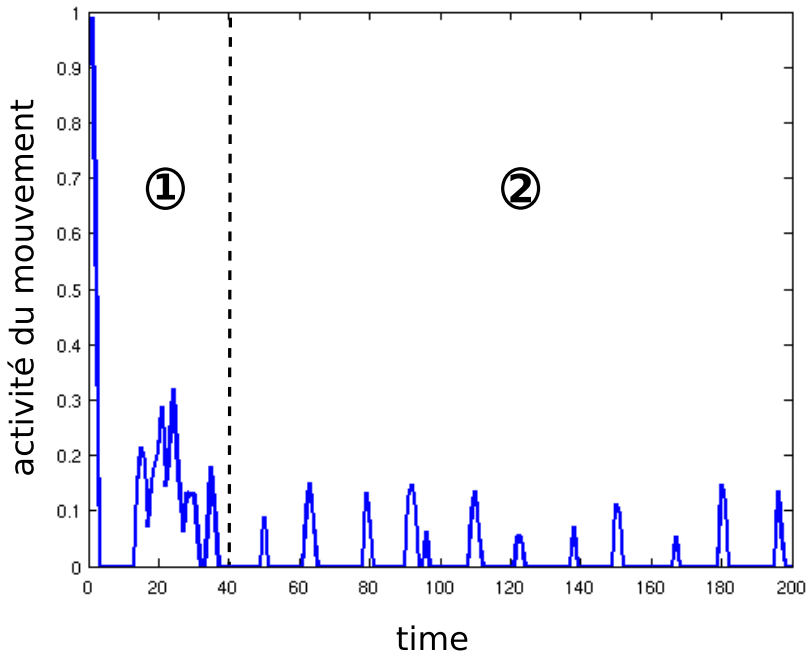


Fig. 5-4 – Mouvements produit par le partenaire social durant un jeu d'imitation. Le protocole expérimental est le suivant: l'humain se place devant le robot (le robot voit uniquement l'expression du partenaire) et doit imiter le robot pendant quelques secondes. L'unique consigne étant de reproduire les expressions faciales du robot. Nous observons que les pics de mouvements faciaux sont espacés très régulièrement dans le temps.

La figure 5-4 montre que les individus sans aucune consigne explicite de rythmicité sont capables de produire des pics de mouvements régulier lors de la phase d'imitation des expressions faciales du robot. Le signal mesuré est le mouvement brut produit par l'expérimentateur humain lorsqu'il effectue les expressions faciales. Nous pouvons observer sur cette figure deux phases: la première phase avant l'itération 40, est une phase d'adaptation de l'expérimentateur vis à vis du set-up expérimentale, le mouvement est irrégulier. La deuxième phase aprés l'itération 40, montre une stabilité de l'interaction une fois que l'expérimentateur s'est totalement adapté au robot. Ces pics correspondent aux différentes expressions que l'individu est entrain d'afficher lors de l'interaction avec le robot. Les mouvements faciaux de l'humain engendrent du mouvement à fréquence constante que le robot sera capable d'utiliser pour prédire une rythmicité entre lui et l'humain.

La figure 5-5 et la figure 5-6 montrent l'émergence de cette rythmicité qui s'installe entre les deux agents ainsi que la prédiction du rythme par le robot. Nous pouvons clairement observer

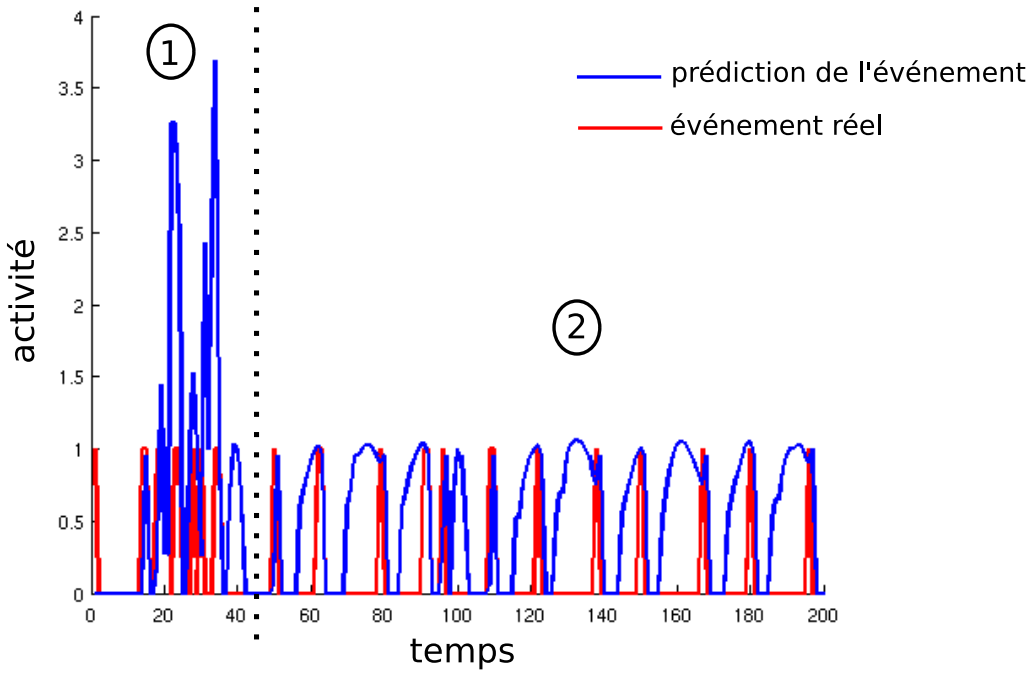


Fig. 5-5 – Emergence d'un rythme entre le robot et le partenaire social. Le protocole expérimental est le suivant: l'humain imite le robot pendant quelques secondes avec pour simple consigne d'imiter les expressions faciales du robot. La phase 1 montre le temps d'adaptation du partenaire social vis à vis du robot, la phase 2 montre qu'au bout de quelques secondes le partenaire est capable de reproduire les expressions faciales et malgré lui avec une certaine rythmicité.

que le robot est capable de prédire le rythme d'interaction avec le partenaire humain. Deux phase distinctes sont à noter, la première correspond explicitement à la phase d'adaptation du partenaire humain au dispositif robotique (phase qui dure environ 50 itérations) induisant une mauvaise prédiction de la rythmicité et la deuxième phase correspond à une bonne prédiction de la rythmicité de l'humain par le robot (l'humain s'est adapté à la tête expressive). Par conséquent, l'expérimentateur produit les expressions faciales suivant la fréquence imposée par le robot. Les prédictions et les événements réels se superposent montrant une corrélation temporelle entre la prédiction et l'événement.

Pour les 10 partenaires naïfs, il y a bien une émergence de la rythmicité, montrant ainsi une synchronie entre les mouvements faciaux du robot et de l'expérimentateur. La tête robotique prédit correctement le rythme de l'interaction 74% du temps.

Cependant, le robot a parfois plus de difficultés, l'expérimentateur ayant les résultats les moins performants tend vers 68% du temps dans le rythme d'interaction avec le robot. Néanmoins, l'expérimentateur naïf le plus performant tend vers de 80% du temps dans le rythme avec le robot. Généralement, le robot semble avoir des difficultés à prédire la rythmicité dans deux cas:

- l'expérimentateur n'est pas suffisament expressif. Dans ce cas, la détection du mouvement n'est pas suffisante pour déclencher une activité du neurone GD.
- l'expérimentateur produit des mouvements faciaux non homogènes, pouvant déclencher plusieurs pics de mouvements pour une expression.

Ce signal non explicite émergeant de l'interaction (rythme) est suffisant pour permettre au robot

de détecter si une personne interagit avec lui. Nous pouvons noter que pour les individus experts du système les résultats tendent vers 100%. La figure 5-6 illustre l'entrée visuelle (détection de

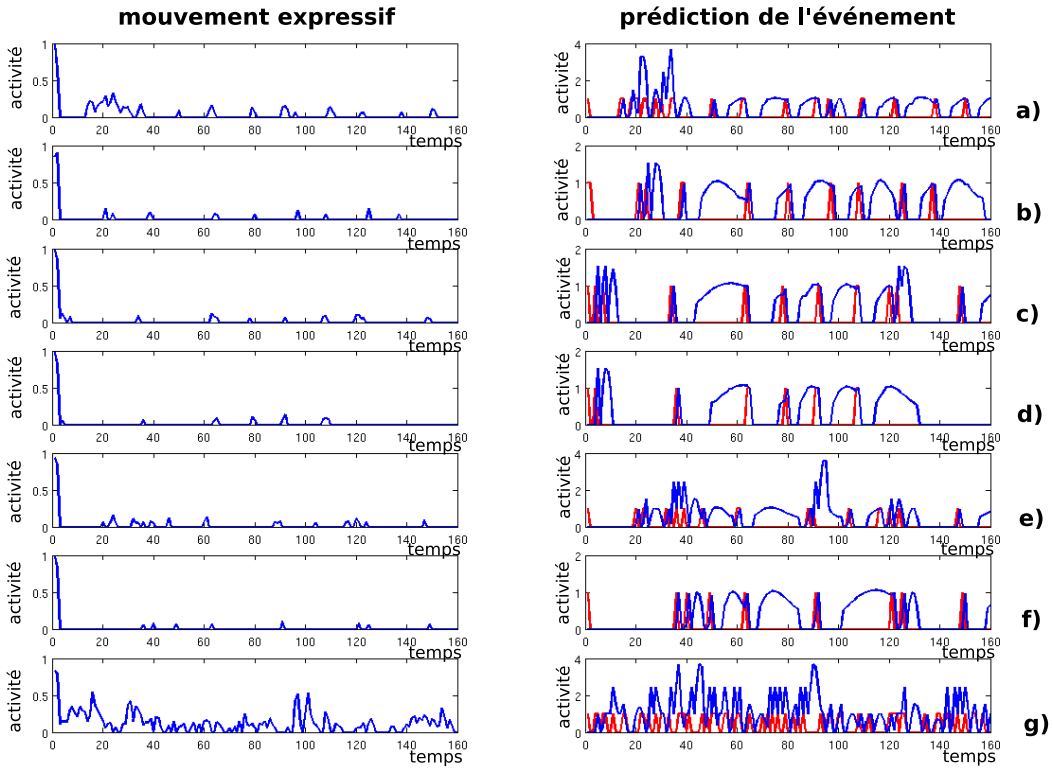


Fig. 5-6 – Illustration des pics de mouvement expressif de quelques expérimentateurs et du signal prédit par la tête robotique. La colonne de gauche correspond aux mouvements perçus par le robot (le partenaire imitant la tête de robot) et la colonne de droite correspondant au signal prédit par le robot. Les sept lignes correspondent aux différents partenaires. La ligne a) b) c) d) correspond à un événement correctement prédit par le robot. Le signal réel ainsi que le signal prédit sont en phase. e) f) montrent des événements beaucoup moins bien prédits, il y a un déphasage entre le signal réel et la prédiction. g) montre une complète désynchronie, le signal est imprédictible. Exemple parfait d'une "non" interaction, une ou plusieurs personnes qui passent aléatoirement devant le dispositif.

mouvement) ainsi que le signal prédit par le robot. Différents individus ont pu être testés, cette figure montre sur une séquence d'interaction la capacité du modèle à prédire correctement la rythmicité entre la tête expressive et l'expérimentateur imitant la tête. Cependant, la figure montre également que le modèle a des difficultés à prédire des personnes qui sont peu expressives (peu de mouvements faciaux: figure 5-6e-f) ou des personnes n'interagissant pas avec la tête robotique (figure 5-6g).

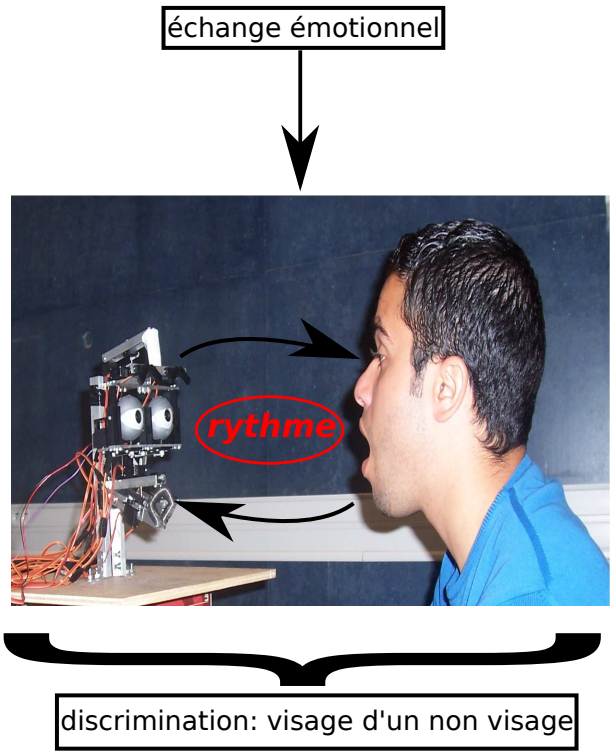


Fig. 5-7 – Emergence de la discrimination un visage d'un non visage grâce à l'interaction émotionnelle. La rythmicité entre les deux agents nous offre un signal de supervision implicite.

5.4 Discrimination d'un visage d'un non visage comme une conséquence de l'interaction émotionnelle

5.4.1 Contexte

Nous gardons le paradigme expérimental qu'au chapitre 4. Les distances d'interaction étant relativement courtes (60cm), le champ de vision du robot comprend exclusivement le visage de l'expérimentateur et le fond de la scène. L'objectif ici, est de montrer que la détection de visage peut être apprise de façon autonome par le biais de l'interaction. De l'interaction émotionnelle émerge une rythmicité entre la tête expressive et l'expérimentateur, pouvant être utilisée comme signal de supervision implicite (Figure 5-7). En fonction de ce signal, le robot pourra déterminer la présence ou l'absence de l'expérimentateur. L'idée interessante de cette étude est la capacité du robot à apprendre à reconnaître les expressions faciales sans détection de visage et au contraire ce jeux d'imitation (apprentissage) permet de discriminer un visage d'un non visage. En d'autres termes, la détection de visage est une compétence que le robot peut acquérir de façon autonome seulement si nous sommes dans un cadre interactif en face à face. Notons simplement que dans le modèle que nous défendons ici, la reconnaissance des expressions faciales semble moins "complexe" en termes de structures à déployer que la dicrimination de visage. Ceci ne semblait pas être évident au début de nos développements (comme tout le monde, nous étions plutot parti sur l'hypothèse classique: la détection précéde la reconnaissance).

Nos travaux soulignent l'importance de l'interaction émotionnelle comme étant un élément

important pour la construction d'autres capacités. Cette interaction induit une rythmicité entre les 2 agents, permettant l'émergence d'un signal. D'autres travaux (Andry et al., 2001) utilisent le rythme comme signal de renforcement. Les travaux de certains psychologues (Murray and Trevarthen, 1985; Muir and Nadel, 1998; Devouche and Gratier, 2001) soulignent l'importance de la temporalité des échanges au cours de l'interaction. Par exemple, les jeunes enfants sont extrêmement sensibles à la cohérence temporelle des échanges. Une forte rupture de l'interaction sociale, entraine chez un bébé des sentiments négatifs désagréables qui peuvent être traduits par une agitation ou même des pleurs. Par contre, un rythme d'interaction constant entre un bébé et sa maman produit des sentiments positifs qui peuvent se traduire par des sourires chez le bébé.

Ces études de psychologie laissent apparaitre l'importance du rythme de l'interaction dans l'apprentissage par imitation. Ce rythme nous servira de signal de renforcement pour l'apprentissage. Concrètement cela peut se traduire simplement par:

- Un renforcement positif est synonyme d'une interaction constante autrement dit chaque système produit au bon moment une action cohérente, qui prolonge l'interaction et entraine une réponse de l'autre.

- Un renforcement négatif peut lui être traduit par une rupture stricte, un changement de rythme de l'interaction.

5.4.2 Architecture de contrôle neuronale

Lorsqu'un sujet imite la tête robotique, des mouvements faciaux sont réalisés par son visage. Dans une interaction en face à face, seuls des mouvements du visage sont réalisés, les pics de mouvements du sujet sont corrélés avec les expressions faciales que le robot réalise successivement. Ces pics ont une fréquence constante qui dépend de la fréquence de changement d'expressions faciales de la tête robotique. Dans notre expérience, le robot change d'expression faciale toutes les 3 secondes.

Comme nous l'avons décrit dans la section précédente (Prédiction du rythme et son apprentissage), le rythme de l'interaction peut être déduit de la prédiction du temps entre 2 pics du signal visuel. Le signal visuel étant simplement la détection de mouvement (figure 5.4) qui est réalisé par l'expérimentateur lorsqu'il imite les expressions du robot.

La mesure de la prédiction de l'erreur est maintenant extrêmement facile à fabriquer. L'erreur à l'instant t est notée E_t, elle correspond simplement à la différence d'activité entre le signal prédit (GP_t) et le mouvement effectivement réalisé (signal d'entrée: GD_t):

$$E_t = GP_t - GD_t \tag{5.6}$$

Cependant, pour améliorer la robustesse du signal, nous allons considérer une fenêtre temporelle correspondant à une période de quelques itérations notée N. Ce signal d'erreur à l'instant t est noté ξ_t:

$$\xi_t = \frac{\sum_{t-N}^{t} E_t}{N} \tag{5.7}$$

ξ_t est le signal d'erreur et par conséquent il permet la détection de la rythmicité. Le signal utilisé étant non spécifique (pics de mouvement), il sera un excellent moyen de renforcer l'absence ou

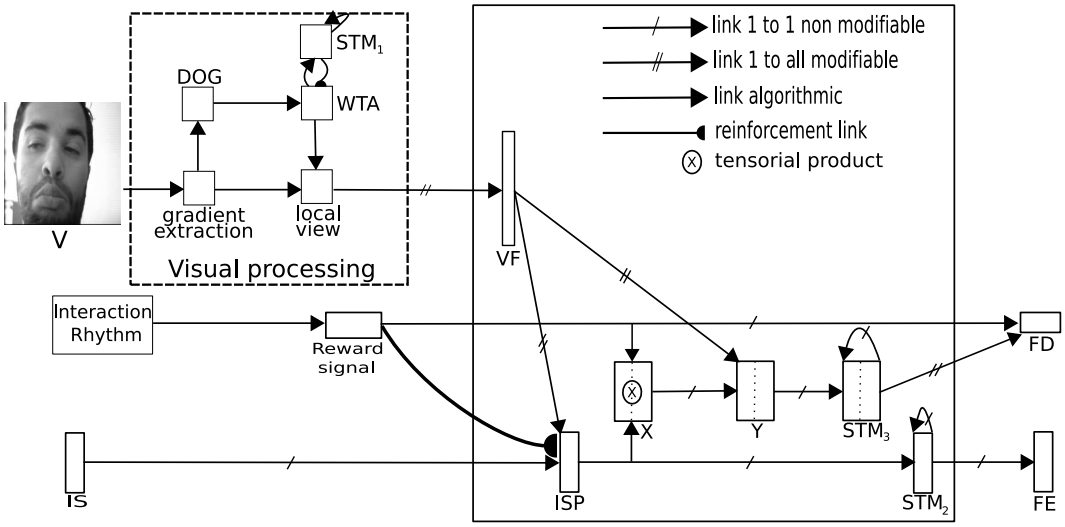


Fig. 5-8 – L'architecture globale pour reconnaître l'expressivité d'un partenaire, imiter et discriminer le visage d'un non visage. Le processus visuel permet d'extraire séquentiellement les vues locales. Le groupe VF (reconnaissance de la vue locale) apprend les vues locales. Chacun des groupes de neurones suivant IS, ISP, STM_2 et FE contiennent 5 neurones correspondant aux 4 expressions faciales et au visage neutre. Un produit tensoriel est réalisé entre ISP (prédiction de l'état interne correspondant à une sorte d'état émotionnel) et un signal de récompense sélectionnant le neurone devant apprendre. Y apprend l'association entre une vue locale et une expression faciale, un neurone spécifique est activé suivant que le lien de récompense venant de l'interaction est présent ou non. Le groupe FD (détection du visage) apprend lui l'association entre le produit tensoriel et le signal de récompense. Son activité correspond quand à elle à la reconnaissance ou au contraire à l'absence d'un visage dans la scène. Le signal de récompense permet également de moduler la vitesse d'apprentissage des expressions faciales en fonction de la situation interaction/non interaction.

la présence d'un partenaire social. Autrement dit, trois cas s'imposent à nous pour la détection de la rythmicité:

- une interaction avec l'expérimentateur où l'humain imite le robot. L'erreur ξ_t sera minimale, une rythmicité est prédite par le robot (Figure 5-3b)
- personne n'interagit avec le robot. L'erreur ξ_t est élevée, le robot est dans l'incapacité de prédire la moindre rythmicité (figure 5-3c)
- Aucun mouvement détecté par le robot. l'erreur ξ_t est egale à zéro, le robot ne peut pas prédire s'il n'a pas de signal d'entrée.

Les détails du réseau de neurone sont présentés dans la section précédente et certains aspects ont été présentés par d'autres auteurs (Andry et al., 2001; Banquet et al., 1997). L'interaction émotionnelle semble fournir des signaux suffisamment riches et intéressants pour pouvoir apprendre des tâches plus complexes. La rythmicité qui émerge de cette interaction fournie un signal de renforcement robuste pour apprendre à reconnaître un partenaire sociale. Dans notre set-up expérimental, les individus interagissent avec le dispositif à des distances courtes (situation de face à face), le robot est par conséquent capable d'apprendre à discriminer un visage d'un non visage. Néanmoins, si les distances d'interactions étaient plus grandes, on pourrait très bien apprendre au robot à discriminer une personne d'une non personne.

Un produit tensoriel est calculé entre le signal provenant du groupe de sortie du sous-réseau

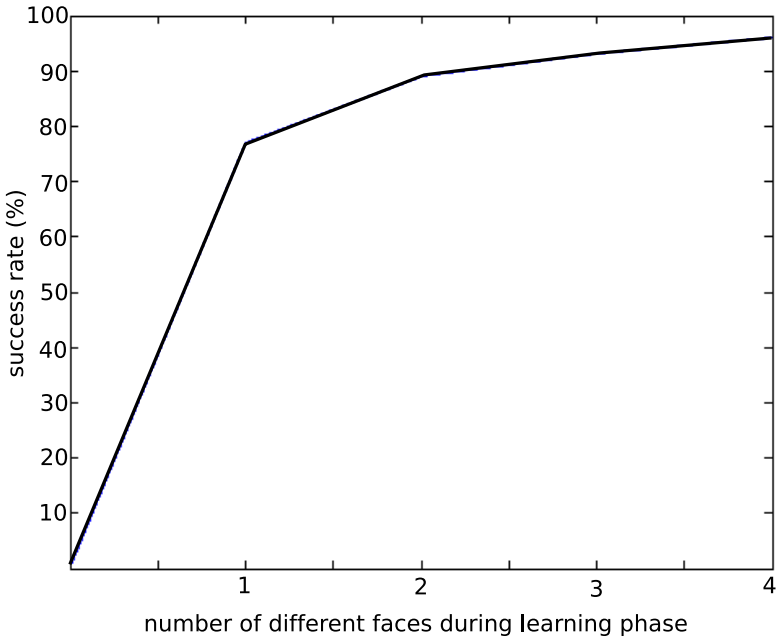


Fig. 5-9 – Généralisation de la capacité à discriminer un visage d'un non visage. Ces résultats montrent le taux de succès de la discrimination d'un visage d'un non visage en fonction du nombre de personnes interagissant avec le dispositif robotique pendant la phase d'apprentissage. Les résultats montrent qu'en apprenant seulement simplement sur 4 individus, l'architecture est capable de généraliser sur la totalité de la base (21 autres personnes soit 3360 images).

capable de prédire le rythme d'interaction et ISP (5 neurones) correspond à l'état émotionnel. Ce groupe de sortie caractérisant l'interaction, comporte seulement deux neurones: le premier est actif lorsque le rythme d'interaction est correctement prédit alors que le second est actif en cas d'échec indiquant une interaction incohérente. Ce produit tensoriel noté X forme une matrice de neurones constituée de 10 neurones (5 lignes et 2 colonnes). La première colonne va correspondre à un état émotionnel si un individu est en interaction avec le robot et la deuxième colonne correspond à un état émotionnel si personne n'interagit avec le robot (un objet pouvant procurer du plaisir au robot).

Un mécanisme de conditionnement très simple utilisant la règle du LMS (Widrow and Hoff, 1960) est utilisé pour associer l'activité des neurones des vues locales reconnues VF (caractéristiques visuelles) avec l'activité du produit tensoriel X. Le groupe de neurones Y se charge d'apprendre ce conditionnement. Cette association se construit au fur et à mesure des interactions. Après l'apprentissage, les associations entre le groupe VF et le groupe Y sont suffisamment élevées pour ignorer l'activité de bas niveau (réflexe) venant de $X_{i,j}$. Par conséquent, les vues locales VF suffisent à activer de manière cohérente le groupe de neurone Y.

Tous les points de focalisations appris sur une même image sont accumulés à l'aide d'une mémoire à court terme STM_3 constituée de 10 neurones (5 lignes et 2 colonnes):

$$STM_{3,(i,j)}(t+1) = \frac{1}{N} Y_{(i,j)}(t+1) + STM_{3,(i,j)}(t) \tag{5.8}$$

(i, j) étant l'indice du neurone $(0 < i \leq 5$ and $0 < j \leq 2)$ et N étant le nombre de points de focalisation analysés.

Une fois l'apprentissage terminé, la matrice STM_3 tend à activer d'avantage la première colonne en présence d'un visage contrairement à la seconde colonne qui s'active en présence d'un non visage. Reste maintenant à effectuer un conditionnement très simple, pour apprendre l'association entre l'activité des neurones du groupe STM_3 et le signal de récompense fourni par le rythme de l'interaction. Le groupe de neurone FD (détection de visage: discrimination d'un visage d'un non visage) constitué de 2 neurones utilisant également la règle d'apprentissage du LMS (Widrow and Hoff, 1960) est utilisé pour apprendre cette tâche (l'un des deux neurones s'activant en présence d'un visage tandis que l'autre en son absence). Le groupe FD est capable d'apprendre en quelques minutes d'interaction (2 à 3 min) la discrimination d'un visage d'un non visage.

Enfin, le signal de récompense émergeant de l'interaction émotionnelle permet également de moduler la vitesse d'apprentissage des expressions faciales en fonction de la situation interaction/non interaction. Le robot apprend à catégoriser les expressions faciales seulement dans le cas où un individu interagit avec lui. Le robot est non seulement autonome dans sa capacité à développer la reconnaissance des expressions faciales mais il est également autonome dans sa capacité à apprendre. Autrement dit, le robot est capable d'apprendre des expressions faciales en fonction de la situation contextuelle.

5.4.3 Résultats

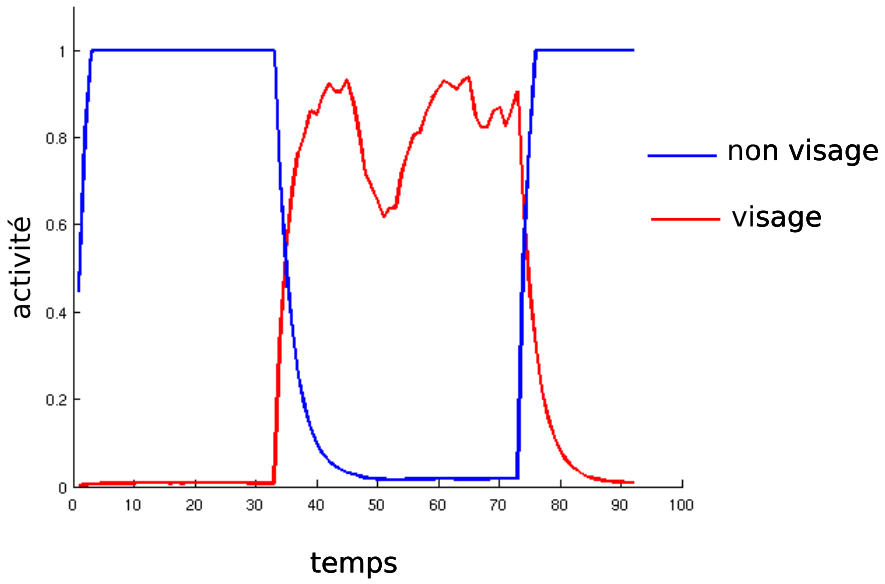


Fig. 5-10 – Activité neuronale de la détection visuelle de visage lors d'une phase d'interaction. Tout d'abord, aucun individu n'interagit avec la tête expressive, puis un expérimentateur se met en face du robot: le neurone codant pour la classe visage est le plus activé. Enfin, l'individu se retire ce qui implique aucune détection de visage: le neurone codant pour la détection de la classe non visage est à son tour le plus activé.

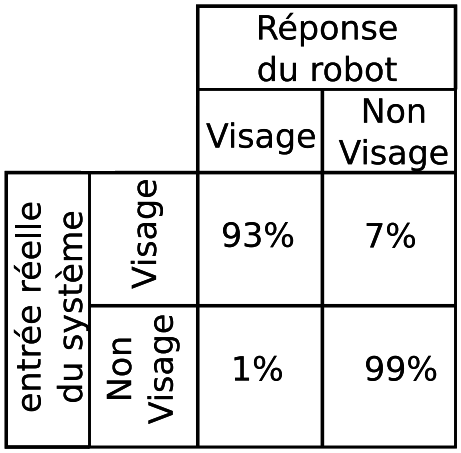


Fig. 5-11 – Taux de confusion montrant la capacité à discriminer un visage d'un non visage sur des images statiques. Les résultats sont effectués en ayant appris sur 4 individus et la généralisation est réalisée sur 21 autres personnes (3360 images).

Les premiers résultats liés à cette apprentissage en ligne et totalement autonome sont très positifs. Lors d'un apprentissage sur une personne, le taux de détection de visage sur cette même personne tend vers 100% mais on souligne aussi que dans 80%, le réseau est capable de détecter un visage d'une autre personne (21 individus testés). Les résultats montrent un modèle très robuste car l'apprentissage dure seulement 2 min (temps réel) avec une seule personne. Ces résultats montrent la robustesse et la capacité de généralisation de notre système visuel et justifie également le choix des filtres DOG (Différences de Gaussienne) pour la focalisation du robot sur des zones intéressantes de l'image (visage expressif du partenaire). Les résultats de la figure 5-9 mettent en évidence deux choses: La première est que plus le nombre d'individus interagissant avec le dispositif robotique croît, plus les résultats sont bons (ceci montre la capacité du système à accroitre ses connaissances). La deuxième est qu'un apprentissage sur un petit nombre d'individus (4 individus appris), permet une très bonne généralisation sur un grand nombre d'individus (21 dans notre cas) avec un taux de reconnaissance de 95%.

Nous avons également mis évidence dans ce chapitre que les interactions émotionnelles peuvent structurer l'apprentissage d'autres tâches : la discrimination d'un visage d'un non visage. Ce que nous essayons de mettre en avant dans ce chapitre est la nécessité d'étudier les interactions entre différents agents pour comprendre des mécanismes cognitifs plus complexes. La dynamique entre 2 agents est un processus important, qui nous offre certains signaux utiles à la construction du "soi" et à la connaissance du monde.

5.5 Discussion et Conclusion

Le modèle que nous proposons dans ce chapitre possède la capacité à distinguer un visage d'un non visage sans supervision explicite. L'idée de base est de prendre en compte la dynamique de l'interaction plutôt que d'analyser les deux agents séparément. Durant une interaction entre l'homme et la machine, un rythme s'installe montrant d'une part que l'humain imite de manière constante les expressions faciales que le robot affiche et d'autre part la volonté de l'humain à transmettre une connaissance au robot. L'idée sous-jacente est de profiter de ce jeu d'imitation pour que le robot détecte la rythmicité entre lui et le partenaire social. La prédiction du rythme

est une manière de fabriquer un signal de supervision permettant de contrôler le comportement du robot. Dans le modèle que nous proposons, ce signal permet la discrimination entre un partenaire interagissant avec le robot ou non. Plus précisément, le robot possède maintenant la capacité à discriminer un visage d'un non visage. Cependant, notre modèle reste tout à fait plausible si nous voulions discriminer une personne d'une non personne. Les limites sont fixées par le set-up expérimental que nous mettons en place. Nous avons choisi d'interagir avec le robot à des distances courtes essayant de mimer les interactions émotionnelles du nourrisson ce qui sous entend que le robot a dans son champ de vision exclusivement un visage. Si nous imaginons maintenant des interactions à des distances plus grandes, le modèle aurait la capacité à prédire la présence d'un humain.

De nombreux auteurs comme Viola et Jones (Viola and Jones, 2004), Littlewort (Littlewort et al., 2004), Rowley (Rowley et al., 1998) et d'autres ont développé des modèles pour la détection de visage même si des efforts doivent encore être effectués pour améliorer la robustesse. Cependant, des connaissances a priori sur la structure des visages sont introduites dans ces architectures (yeux, bouche, nez) pour améliorer les performances. Contrairement aux différentes méthodes existantes, nous essayons de comprendre quels sont les mécanismes minimaux à introduire pour détecter un visage. Autrement dit, quels sont les réseaux à mettre en oeuvre pour rendre le système autonome, ceci impliquant une réflexion sur le paradigme expérimental que nous devons utiliser. De plus, nous cherchons à mettre en évidence quelles sont les informations essentielles pour la détection de visage, les caractéristiques faciales doivent émerger de l'apprentissage.

Nous montrons également au travers de ces expériences que la détection de visage semble plus complexe à obtenir en termes de briques élémentaires à introduire que la reconnaissance des expressions faciales. En effet, la construction d'un signal de supervision est nécessaire pour rendre le système autonome. Les interactions émotionnelles peuvent être vues comme un moyen de structurer l'apprentissage d'autres tâches. Dans cette étude, nous avons montré que l'interaction émotionnelle peut amorcer la détection de visage. Cependant, notre approche peut être généralisée à d'autres apprentissage d'une plus grande complexité puisque nous avons montré dans (Gaussier et al., 1998; Andry et al., 2001; Andry et al., 2002) qu'une simple architecture sensori-motrice est suffisant pour déclencher un bas niveau d'imitation.

En Conclusion, ce travail suggère qu'un système parent/bébé est un système autopoiétique (Mataruna and Varela, 1980) dans lequel les interactions émotionnelles et l'empathie sont des éléments importants pour maintenir l'interaction et permettre l'apprentissage de capacités toujours plus complexes.

Chapitre 6

Intensité expressive

6.1 Introduction

L'être humain est capable de produire une multitude d'expressions faciales autres que les six émotions primaires d'Ekman. Il peut exprimer des expressions mélangées avec une palette variée d'intensités expressives. En effet, une simple expression peut exprimer plusieurs états selon son niveau d'intensité. Si nous prenons l'exemple de la peur, le sujet humain peut exprimer une peur légère, une forte angoisse ou à l'extrême une peur panique. La reconnaissance d'une même expression peut avoir plusieurs significations selon le niveau d'intensité reconnu. Par conséquent, si l'on désire progresser dans la compréhension des interactions sociales et réaliser des systèmes capable d'assurer un mimétisme fin de nos expressions (i.e j'ouvre plus ou moins la bouche, plus ou moins vite et le robot mime fidèlement l'état de mon visage) alors l'intensité de l'expression est un paramètre très important à prendre en compte.

Le système FACS de P. Ekman introduit le concept d'échelle d'activation suivant les différentes unités d'action (UA). Les UA peuvent avoir plusieurs intensités, 5 graduations allant de A à E. A notre connaissance, cet axe de recherche n'a pas beaucoup été développé. Les travaux de Fasel (Fasel and Luettin, 2000) propose de discriminer, les 5 niveaux d'intensité associés aux UA. La méthode est classique: une différence entre l'image courante et une image de référence (visage neutre) est réalisée, suivie d'une réduction de la dimension par une analyse en composante principale (ACP) ou une analyse en composante indépendante (ACI). Le rôle de ce genre transformation est de réduire la complexité du problème. Enfin, un classifieur "K plus proches voisins" est utilisé.

D'autres auteurs comme (Ghanem et al., 2008; Ghamen and Caplier, 2008) réduisent les niveaux d'intensité à 3 niveaux d'activation: faible intensité, forte intensité et intensité moyenne. Leur modèle ne se base pas sur les UA mais sur l'analyse de distance faciales (ouverture et étirement de la bouche, ouverture des yeux, distance des yeux aux sourcils). L'intuition de ces méthodes réside dans la variation des distances lorsqu'on produit des expressions faciales. Pour améliorer la modélisation, la théorie de l'évidence est introduite pour permettre une classification plus fine.

Enfin, il existe d'autres modèles se basant sur la déformation d'un visage expressif (déplacements de points et estimation de vecteurs vitesses) vis à vis d'un visage neutre. Ces techniques sont par conséquent un moyen de définir la notion d'intensité. (jier Lien and Cohn, 1998; Lee and Xu, 2003) définissent une dimension d'intensité continue allant de l'intensité nulle vers une intensité maximale.

Dans ce chapitre, nous désirons vérifié si notre architecture peut être adaptative pour réaliser ce type de reconnaissance. Contrairement aux modèles existants, nous cherchons à faire émerger cette capacité au travers de l'architecture neuronale. Dans un premier temps, nous montrons comment la tête expressive sera capable de reproduire des expressions secondaires correspondant à des expressions un peu plus sophistiquées (mélanges d'expressions primaires). Dans un second temps, nous montrons comment la notion d'intensité peut émerger d'une architecture sensori-motrice. Enfin, nous montrons comment cette architecture peut intégrer la notion d'expression faciale intensive.

6.2 La reconnaissance de primitives motrices

6.2.1 La notion de primitives motrices

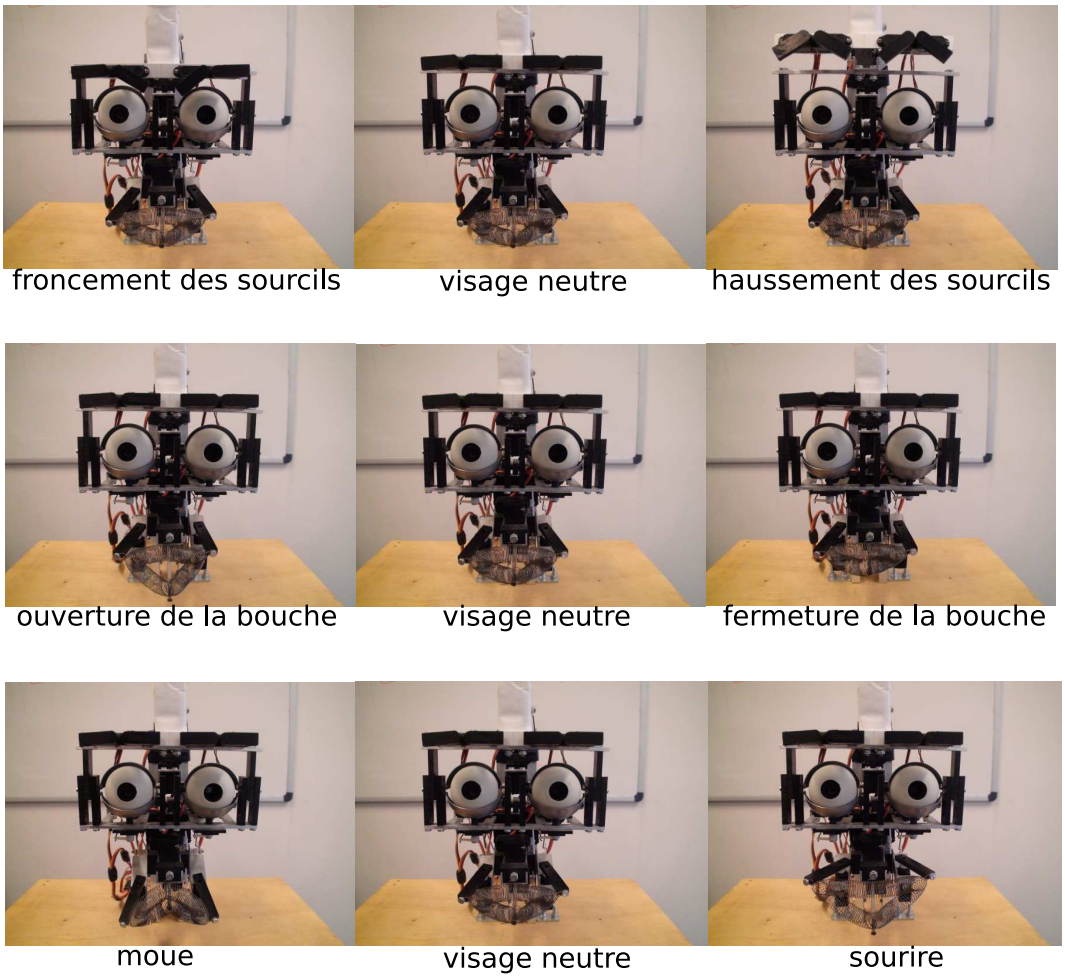


Fig. 6-1 – Différentes primitives motrices que le robot est capable de produire.

Le chapitre 4 avait pour but de développer un système capable de reconnaitre cinq expressions faciales : la joie, la tristesse, la colère, la surprise et le visage neutre. Néanmoins, ce système a comme principale limitation la production d'un ensemble restreint d'expressions, ne

pouvant pas exprimer des mimiques faciales plus sophistiquées. Afin d'obtenir un robot capable de produire n'importe quelle expression faciale, plutôt qu'un ensemble fini d'expressions faciales, notre solution est d'apprendre des groupements musculaires du visage se contractant ensemble. Ceux sont ces groupes musculaires que nous appellerons primitives motrices qui seront apprises séparément par le robot. La tête de robot étant composée de treize servos moteurs agissant chacun comme un muscle du visage. Ceux-ci sont regroupés par groupements musculaires ou groupe de servos moteurs fonctionnant ensemble. Trois primitives seront exploitées:

- le haussement et le froncement des sourcils (P1): cinq servos moteurs contrôlent ce groupement musculaire. Cette primitive donne un air de surprise ou de colère au robot du moins pour la partie haute de la tête robotique.
- l'ouverture (du milieu) de la bouche (P2): un seul servo moteur contrôle cette primitive.
- la position des coins de la bouche (P3) est controlée par 4 servos moteurs. Ces servos moteurs sont capables d'exprimer un sourire ou une moue.

Les 3 primitives motrices ont été choisies arbitrairement (figure 6-1), elles semblent être suffisantes pour exprimer une palette expressive variée. Nous avons fait le choix d'utiliser trois neurones codant pour trois niveaux d'intensité différents par primitives motrices. Pour les sourcils correspondant à la primitive 1, trois neurones sont utilisés pour coder trois positions motrices: haussement, froncement et sourcils neutre. Pour la bouche (primitive 2), trois positions motrices sont également utilisées pour l'ouverture, la fermeture et la bouche neutre. Pour la primitive 3 correspondant à un sourire, une moue ou une bouche neutre, trois neurones sont également utilisés pour coder ces positions motrices.

6.2.2 Premier modèle neuronal pour la reconnaissance et la production de primitives expressives analogiques

Les primitives motrices sont apprises de la même manière que les expressions émotionnelles. Le jeux d'imitation entre les 2 agents fournit le cadre expérimental de notre approche. Pour tester notre modèle, nous proposons de développer une architecture de contrôle neurale ainsi que d'adopter le protocole expérimental suivant: Dans une première phase de l'interaction, le robot produit des primitives motrices de manière aléatoires (haussement de sourcil, ouverture de bouche, sourire etc) durant 2 secondes et nous demandons à l'expérimentateur de l'imiter. Entre chaque mouvement moteur, le robot repasse par le visage neutre pour éviter les mauvaises interprétations. Une procédure équivalente est utilisée en psychologie expérimentale pour éliminer le biais du set-up expérimental. Après cette première phase durant entre 3 et 4 minutes selon la "patience" du sujet, le générateur de primitive motrice est stoppé. On constate alors la capacité du robot à imiter les mimiques faciales de l'expérimentateur. Le robot n'est plus limité à un ensemble réduit d'expressions faciales mais il peut maintenant produire des expressions plus évolués comme un sourire surpris ou même une mixture entre la joie et la colère.

Dans la perspective d'un apprentissage autonome, nous évitons les mécanismes ad-hoc comme le cadrage de visage exactement de la même manière que pour la reconnaissance des expressions faciales (cf chapitre 4). Le système visuel est basé sur une exploration des points de focalisations qui sont le résultat d'une convolution entre une filtre de différence de gaussiennes (DOG) et l'image du gradient (obtenue par un filtre de Canny-Derriche), permettant de focaliser sur les zones de courbures (sourcils, coins de bouche, etc). Ces points de focalisation subissent une transformation log-polaire augmentant la robustesse des vues locales. L'apprentissage est réalisée par le SAW (Self Adaptatif Winner) où le fonctionnement de ce modèle de réseau de neurones est

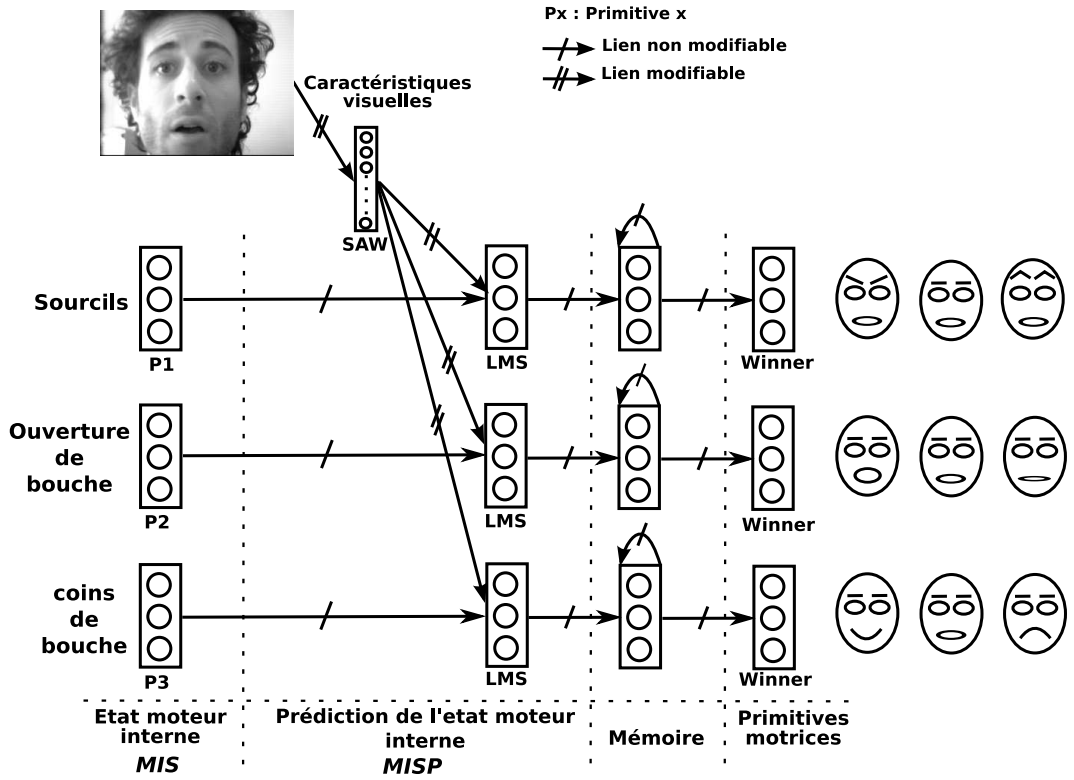


Fig. 6-2 – Modèle d'apprentissage des primitives motrices. Pour chaque primitive, trois neurones codent pour trois intensités. Primitive motrice P1: haussement des sourcils, froncement des sourcils et sourcils neutre. Primitive motrice P2: ouverture de la bouche, fermeture de la bouche et bouche neutre. Primitive motrice P3: sourire, moue, bouche neutre. Ce modèle montre qu'une architecture purement sensori-motrice est capable de réaliser l'association entre la sensation visuelle du robot et son état moteur interne. Après 3 à 4 minutes d'apprentissage, l'activation de certaines sensations visuelles permet d'activer les états moteurs internes correspondants.

décrit en détails dans le chapitre 4. Le SAW apprend les caractéristiques expressives. Il permet à la fois un apprentissage en un coup et un moyennage des prototypes dans le temps. La règle d'apprentissage permet d'apprendre les choses nouvelles très rapidement (toutes les caractéristiques visuelles encore jamais vues peuvent être apprises en un coup) et permet d'adapter les prototypes dans le temps (moyennage des caractéristiques visuelles qui sont proches par exemple un haussement de sourcil). De nombreux distracteurs peuvent être pris sur le fond de la scène ou bien sur des parties non expressives du partenaire comme les cheveux. La distance d'interaction homme-robot restant courte, les points pris sur le fond de la scène sont peu nombreux. Ces distracteurs sont appris par le groupe de neurones SAW, impliquant que des neurones s'activent en présence de certaines caractéristiques non expressives. Néanmoins, seules les caractéristiques expressives du visage sont renforcées par le modèle (Figure 6-2) grâce à l'association entre l'état interne moteur et les caractéristiques visuelles apprises par le SAW. $MISP$ correspond à la prédiction de l'état interne moteur qui associe l'activité du SAW avec l'état interne moteur courant MIS du robot. L'association se fait par simple mécanisme de conditionnement utilisant la règle des moindres carrés (LMS) légèrement modifiée. La variation des poids w_{ij} du groupe

$MISP$ suit la règle suivante:

$$\Delta w_{ij} = \epsilon.SAW_i.(MIS_j - MISP_j).\sum_j MIS_j \tag{6.1}$$

avec le terme $\sum_j MIS_j$ servant de modulation. En effet, si aucune activité est présente sur le groupe des primitives motrices alors la prédiction n'est pas considérée. Ce terme sert à l'apprentissage d'une primitive motrice à la fois. Nous avons également choisi d'analyser seulement 10 points de focalisations pour permettre au système d'être réactif durant l'interaction avec le partenaire humain. Le système est capable d'analyser 3 à 5 images par secondes. Une mémoire à court terme STM est introduite pour mémoriser tous les points de focalisation appartenant à la même image:

$$STM_i(t+1) = \frac{1}{N}.MISP_i(t+1) + \frac{N-1}{N}STM_i(t) \tag{6.2}$$

i est l'indice des neurones codant pour l'intensité d'une primitive motrice, par exemple $MISP_i$ correspond au i^{th} neurone pour la primitive 1 ($0 < i \leq 3$). La STM est utile pour avoir une analyse d'une image complète car elle accumule les vues locales explorées. Une seconde mémoire à court terme peut également être utilisée comme mémoire glissante sur les images consécutivement analysées. Ces mémoires rendent le contrôle moteur plus stable pour l'interaction. Pour finir la prise de décision est réalisée par un mécanisme de Winner-take-all pour les trois primitives motrices.

6.2.3 Résultats

Après l'apprentissage, les bassins d'attraction sont créés grâce aux associations entre SAW et $MISP$ pour passer l'activité réflexe de bas niveau venant de l'état interne moteur MIS. Dans ce cas, les primitives motrices résultent d'une intégration temporelle de la prédiction de l'état interne moteur associé aux différentes caractéristiques visuelles analysées par le système. Chaque primitive motrice est corrélée avec un ensemble spécifique de points de focalisations. Ces points d'intérêt correspondent aux zones locales importantes pour la reconnaissance des primitives motrices. Par exemple, certaines zones autour des sourcils caractérisent un haussement de sourcil, certaines autour de la bouche caractérisent une ouverture de la bouche. Les vues locales sont prises dans les mêmes zones indépendamment des individus comme pour la reconnaissance des expressions faciales où les points de focalisation étaient pris dans les même zones pour la majorité des expérimentateurs.

La figure 6-3 montre l'activité neuronale des signaux capturée durant la phase de reproduction de la tête expressive. Cette figure montre que le robot est capable de reproduire correctement les intensités de 3 primitives motrices (ouverture bouche, sourcils, coins de la bouche) en ayant réalisé un apprentissage en ligne et autonome. Le temps d'apprentissage est relativement court, 4 minutes d'apprentissage ont suffit pour que le robot soit capable de reproduire les intensités motrices du partenaire humain et ceci pour les trois groupements moteurs. Cette figure montre également que dans cette séquence de reproduction la tête robotique réagit correctement à toutes les intensités motrices et que la mémoire à court terme lisse l'activité neuronale (figure 6-3 première ligne). Cette mémoire a deux conséquences, la première est que le robot aura un temps de réaction légèrement plus long mais se stabilisera plus facilement sur une intensité motrice (il ne passera pas d'une intensité à l'autre). Ceci expliquant la difficulté du robot à reproduire les mimiques faciales lors d'un changement d'intensité motrice par le partenaire humain. Autrement dit, les phases de transitions introduisent des artéfacts faisant échouer la reconnaissance par

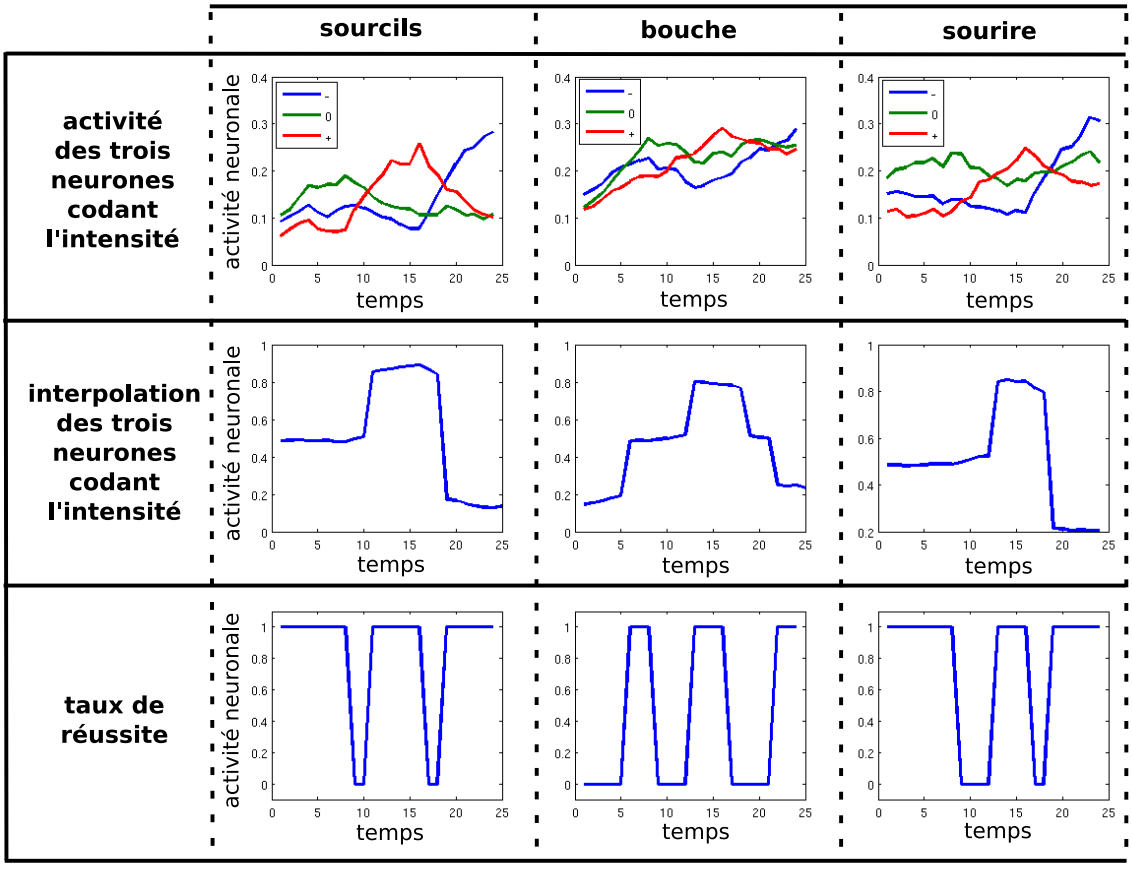


Fig. 6-3 – Activité neuronale des trois neurones codant pour les intensités motrices. Une colonne montre l'activité neuronale de différents signaux pour les différentes primitives motrices apprises. Les primitives motrices analysées sont: les sourcils (haussement, froncement, neutre), la bouche (ouverture, fermeture, neutre), les coins de la bouche (sourire, moue, neutre). Les lignes représentent les différents signaux capturés: la première ligne représente l'activité analogique des 3 neurones codant l'intensité motrice, la deuxième ligne représente l'interpolation neuronale des 3 neurones codant l'intensité motrice et la troisième ligne représente le taux de réussite à savoir si le robot imite correctement l'humain (si l'activité est égale à 1 alors le robot reproduit correctement les intensités motrices du partenaire humain et au contraire quand cette activité est égale à 0 la tête de robot se échoue dans la reproduction). Notons simplement que le robot échoue dans l'imitation lors des changements d'intensité motrice par exemple lorsque l'humain passe des sourcils froncés au haussement des sourcils.

exemple lorsque l'humain passe des sourcils froncés au haussement des sourcils ou de la bouche ouverte à la bouche fermée. De plus la figure 6-3 montre l'interpolation neuronale réalisée suivant l'équation:

$$f(x) = \frac{\sum_i pos_i.x_i}{\sum_i x_i} \tag{6.3}$$

x_i étant l'activité neuronale du i-ème neurone codant pour l'intensité motrice i, pos_i étant la position du neurone $pos_i \in [0,1]$, $f(x) \in [0,1]$ étant le résultat de l'interpolation. Nous montrons à travers cette figure que l'activité de l'interpolation est très nettement inférieur à 0.5 quand le neurone codant pour l'intensité faible est gagnant, autour de 0.5 quand le neurone codant pour l'intensité moyenne est gagnant et supérieur à 0.5 quand le neurone codant pour

l'intensité forte est gagnant. La figure 6-4 montre les taux de confusion pour la reconnaissance de

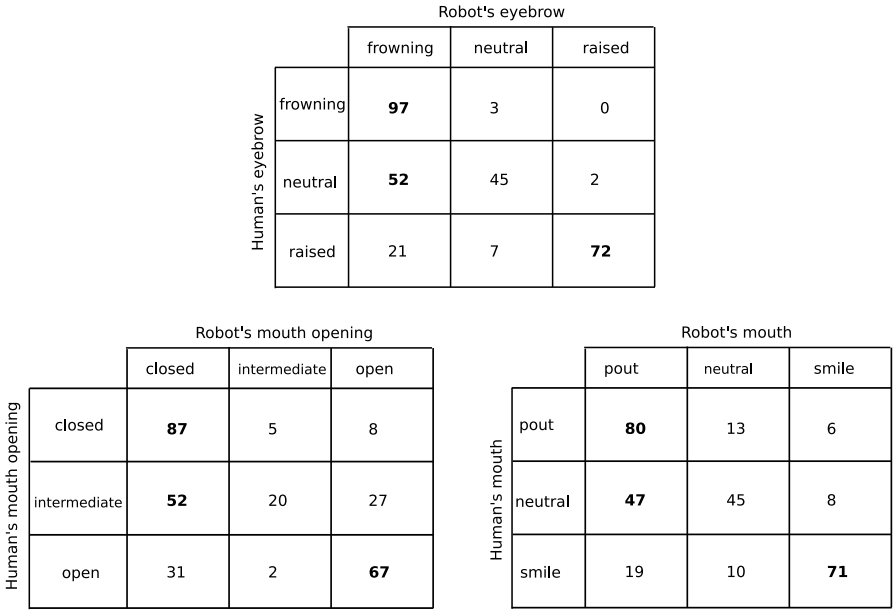


Fig. 6-4 – Tableaux montrant le taux de confusion pour les 3 primitives motrices apprises. La base de donné est constituée de 5040 images des différentes primitives motrices et de leur intensité. 14 personnes constituent la base de donné, ces expérimentateurs ont préalablement imité la tête de robot pour annoter les images en fonction des mimiques faciales du robot. Ces différents tableaux illustrent les taux de confusion pour la reconnaissance de l'intensité expressive des différentes primitives motrices. Le système a appris 14 individus durant la phase d'apprentissage (4 minutes par personnes) et par la suite la tête de robot a dû imiter les mimiques faciales de ces 14 individus.

l'intensité expressive des différents groupements moteurs. Cette figure montre qu'un temps très court d'apprentissage (4 minutes pour apprendre 3 groupements musculaires contenant chacun 3 intensités), le dispositif robotique est capable de reproduire les mimiques faciales de 14 individus appris durant la phase d'apprentissage. Les taux de réussite sont nettement supérieurs à 70% pour les intensités extrêmes (haussement et froncement des sourcils, ouverture et fermeture de la bouche, sourire ou moue de la bouche) par ailleurs les résultats sont moins performants pour les positions neutres. En moyenne, le taux de reconnaissance des intensités expressives des différents groupements moteurs est de 65%. Le robot expressif semble avoir plus de faciliter à imiter les positions des sourcils que les mimiques de la bouche. Par ailleurs, une conséquence de notre modèle est que ces taux de réussite n'illustrent pas la capacité du système à combiner des groupements musculaires (primitives motrices). Le robot est maintenant capable de produire des mimiques faciales très étonnantes. Le robot peut exprimer pour le haut du visage un étonnement en haussant les sourcils et pour la bouche afficher un sourire ou encore un haussement de sourcils et une moue. Ceci laissant supposer que le robot est capable d'afficher des expressions faciales secondaires (expressions faciales mélangées). Cependant, le robot est également capable d'exprimer des expressions faciales primaires comme la joie (sourcils neutre et sourire) ou la surprise (haussement des sourcils et bouche ouverte). La combinaison des primitives motrices nous laisse la possibilité d'envisager que le système ne soit plus figé à un petit ensemble d'expression. La figure 6-5 montre les tables de confusion sur le test en généralisation à savoir si la capacité du robot à imiter des sujets avec lesquels il n'a pas encore interagi. Etant donnée que notre base d'apprentissage ne contient pas suffisamment de personnes (14 personnes), nous avons utilisé une

Robot's eyebrow

Human's eyebrow	frowning	neutral	raised
frowning	**50**	32	17
neutral	**35**	32	30
raised	21	20	**52**

Robot's mouth opening

Human's mouth opening	closed	intermediate	open
closed	**56**	29	13
intermediate	32	**35**	29
open	23	29	**42**

Robot's mouth

Robot's mouth	pout	neutral	smile
pout	31	**40**	23
neutral	**37**	36	22
smile	26	27	**41**

Fig. 6-5 – Test de généralisation du système obtenue durant une interaction naturel avec le robot: Après que 14 personnes ont interagit avec la tête de robot durant la phase d'apprentissage (les partenaires doivent imiter le robot), les rôles sont inversés et le robot imite des individus avec lesquelles il n'a jamais interagit durant l'apprentissage. Ces tableaux montrent le taux de succès et les taux de confusions pour chaque primitive motrice. Ces statiques sont obtenues grâce à la méthode de cross validation (on retire à chaque fois une personne de la base d'apprentissage pour pouvoir la tester en généralisation).

cross-validation pour avoir ces premiers résultats (un individu est enlevé de la base d'apprentissage pour ensuite être testé en généralisation et ceci pour tous les individus appartenant à la base de donnée). Les résultats montrent une capacité de généralisation mais les résultats sont moyens (42% de réussite sur des individus non appris). Cependant le haussement et le froncement des sourcils ainsi que l'ouverture de la bouche sont quantitativement mieux reconnus par le système. Ces expériences laissent sous entendre que la généralisation sur des expérimentateurs inconnus pour le robot est faisable. Néanmoins, le nombre de personne interagissant avec le dispositif robotique doit augmenter pour obtenir des résultats meilleurs. La base de donnée n'est pas encore assez fournie, elle n'est pas suffisamment diversifiée pour permettre une généralisation plus précise. Nous devons également prendre en compte l'expressivité de l'expérimentateur, la qualité de l'imitation de l'expérimentateur est une condition nécessaire pour obtenir des résultats de qualité. Grâce à l'interaction avec l'environnement social, l'apprentissage est autonome, mais en contre partie des difficultés peuvent apparaître puisque les individus ne sont pas des experts du système. Enfin, comme pour la reconnaissance des expressions faciales, la dynamique de l'interaction entre l'homme et le robot induit un apprentissage contenant des faux exemples car le temps de réaction de l'humain vis à vis des intensités motrices du robot, n'est pas instantané.

6.2.4 Amélioration par l'utilisation des champs de neurones dynamiques

Jusqu'à présent, nous avons montré qu'il était possible de contrôler le niveau intensité de plusieurs groupements musculaires en utilisant 3 neurones chacun codant une intensité. Le niveau d'intensité n'était qu'une simple interpolation de ces neurones. Nous allons proposer un

mécanisme plus robuste qu'une simple fonction d'interpolation: un champ de neurones (neural field) pour obtenir plus de finesse vis à vis du contrôle moteur de la tête expressive.

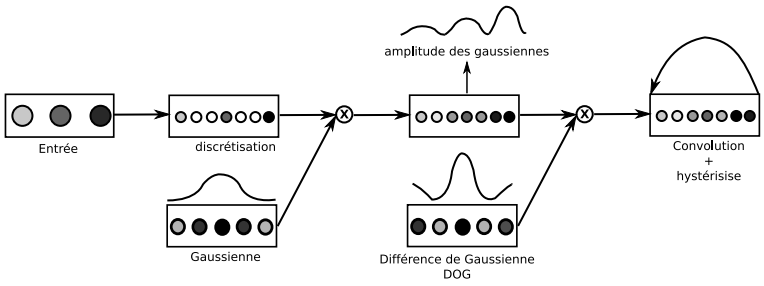


Fig. 6-6 – Modèle du Neural Field (champ de neurones). L'équation du neural field peut être facilement découpée en une succession de groupe de neurone. Les champs de neurone ont des propriétés intéressantes pour un contrôle moteur stable. Les champs de neurone sont utilisés pour obtenir un signal plus stable dans le temps (propriété temporelle) ainsi qu'une capacité de bifurcation.

Le modèle reste identique au précédent excepté pour le contrôle moteur de la tête expressive (Figure 6-7). L'équation de $f(x)$ (équation 6.3) est remplacé par l'équation dynamique du champ de neurones (Amari, 1977) (Figure 6-6).

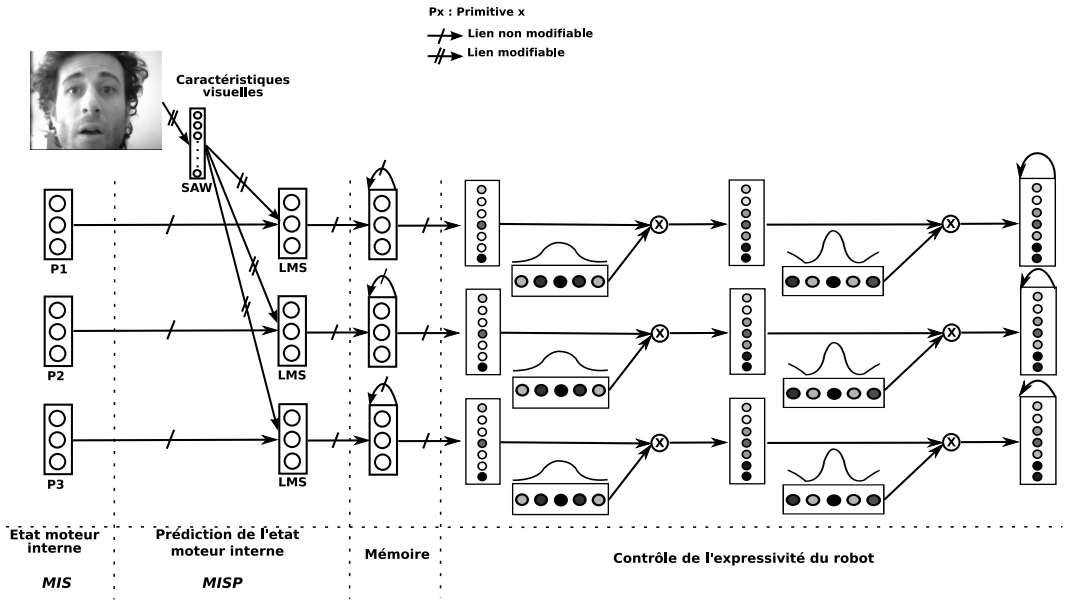


Fig. 6-7 – Reconnaissance des intensités motrices à l'aide des champs de neurones. Pour chaque primitives, trois neurones codent pour trois intensités. Primitive motrice P1: haussement des sourcils, froncement des sourcils et sourcils neutre. Primitive motrice P2: ouverture de la bouche, fermeture de la bouche et bouche neutre. Primitive motrice P3: sourire, moue, bouche neutre. Ce modèle montre qu'une architecture purement sensori-motrice est capable de réaliser l'association entre la sensation visuelle du robot et son état moteur interne. Après 3 à 4 minutes d'apprentissage, l'activation de certaines sensations visuelles permet d'activer les états moteurs internes correspondant. L'ajout du champ de neurone induit un contrôle plus fin de l'intensité motrice des différentes primitives. Les 3 neurones sont étalés sur plusieurs neurones ce qui permet d'avoir une palette d'intensités plus variées.

Le champ de neurones vient de la théorie des systèmes dynamiques. Dans la théorie des

systèmes dynamiques, le robot expressif est un système dynamique dont un ensemble de variables détermine un état X du robot (par exemple l'intensité moteur). Cet ensemble de variables contrôlent les mimiques faciales du robot lors d'une interaction dans un environnement naturel. L'équation d'un tel système est de la forme suivante:

$$\frac{dX}{dt} = f(X) = \sum_i w_i.f_i(X) \tag{6.4}$$

L'utilisation d'une telle équation combine un ensemble de comportements qui peuvent être élémentaires et qui influencent le comportement globale du robot pour la réalisation de différentes tâches complexes (les poids w_i pondèrent les comportements élémentaires f_i). Dans notre cas, les comportements élémentaires sont les différents niveaux d'intensités motrices provenant de plusieurs chaînes de traitements. Les poids w_i ne sont pas constants, ils peuvent évoluer dans le temps ce qui permet des changements de dynamique (bifurcations du système).

Les champs neuronaux permettent la conception de tels systèmes parfois très complexes, ils ont été introduit par Amari en 1977 (Amari, 1977). Ces champs ont très souvent été utilisés pour le contrôle moteur (Schöner et al., 1995; Bicho and Schoner, 1997). Le champ de neurones est constitué de neurones dont l'abscisse x code une intensité motrice, pouvant être assimilée à une direction. L'équation de l'évolution temporelle des potentiels des neurones suit la loi suivante:

$$\tau\frac{du(x,t)}{dt} = -u(x,t) + h + s(x,t) + \int w(x-v)f(u(v,t))dv \tag{6.5}$$

$u(x,t)$ correspond au potentiel courant du neurone d'abscisse x, f(u(x,t)) correspond à l'activité de ce neurone, s(x,t) correspond à la stimulation reçue par le champ de neurones, h correspond au potentiel de repos et τ est une constante de temps. La fonction w symbolise l'interaction avec les neurones voisins, les interactions sont de types chapeaux mexicains.

Les propriétés du champ de neurones restent très intéressantes pour le contrôle moteur par exemple le contrôle de l'intensité motrice:

- des consignes contradictoires peuvent cohabiter sur un même champ de neurone
- la dynamique du champ permet de fusionner ou de scinder des consignes proches permettant l'élimination des régimes oscillants
- des effets mémoires apparaissent procurant une stabilité su système malgré d'éventuelles absences temporelles

L'ajout du champ de neurone induit un contrôle plus précis de l'intensité motrice comparativement aux 3 neurones codant pour les 3 niveaux d'intensités. Les 3 neurones sont étalés sur une population de neurones (31 neurones codant chacun pour une intensité). Dans cette population, tous sont inactifs sauf les neurones extrêmes et le neurone central codant chacun pour les intensités extrêmes et intermédiaire. Une convolution avec une gaussienne est réalisée sur cette population permettant d'avoir un voisinage actif et un étalement des activités neuronales. Dans un second temps, une différence de gaussienne est réalisée, le but étant de mettre en compétition les neurones voisins. Enfin, une hystérésis temporelle est appliquée permettant ainsi une plus grande stabilité dans le temps (mémoire).

Afin de mettre en évidence différents paliers d'intensités motrices ainsi que la comparaison des intensités motrice produites par l'homme et imitées par le dispositif robotique, nous avons choisi comme primitive l'ouverture de la bouche pour montrer l'effet de l'utilisation d'un champ neuronal. L'ouverture de la bouche semble être une primitive motrice pouvant être mesurée

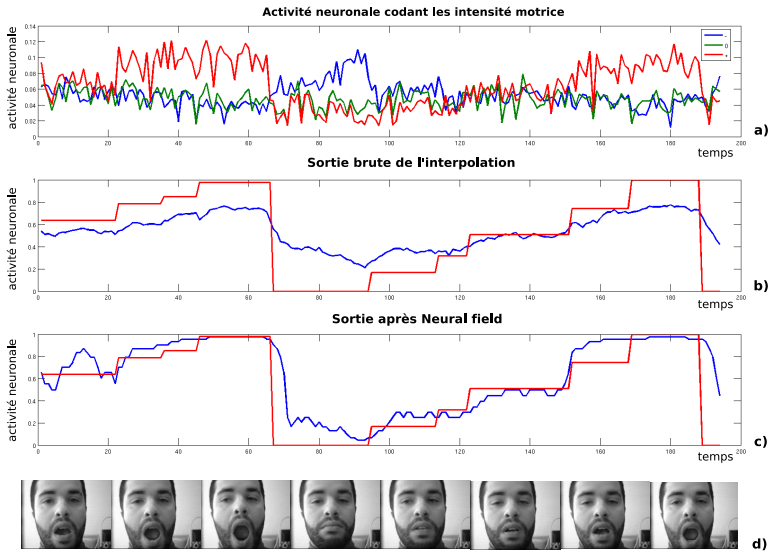


Fig. 6-8 – Intensité motrice de l'ouverture et la fermeture de la bouche durant la phase de reproduction de la tête expressive. Ces courbes montrent l'activité analogique de plusieurs neurones qui sont impliqués dans l'ouverture et la fermeture de la bouche. a) montre l'activité neuronale des 3 neurones codant chacun pour une intensité motrice (ouverture, fermeture et position intermédiaire de la bouche). b) montre l'interpolation des 3 neurones codant pour des ouvertures différentes de la bouche. c) montre l'effet du champ de neurones sur la reconnaissance des différentes ouvertures de la bouche. Pour les courbes b) et c): la courbe rouge correspond à l'ouverture réelle de la bouche de l'expérimentateur et la courbe bleu représentant l'intensité reconnue par la tête expressive. d) montre la séquence d'image analysée.

facilement. La figure 6-8 montre l'intensité motrice de l'ouverture de la bouche durant la phase de reproduction de la tête expressive (ouverture que nous avons mesuré manuellement sur chaque image pour obtenir une vérité terrain). Les différents signaux montrent la capacité du modèle à reconnaître les différentes intensités motrices de la bouche (courbes équivalentes pour les autres groupements moteurs). La figure 6-8 b) et c) montre l'apport du Neural Field. Sans ce dernier, la dynamique neuronale est écrassée et les différents paliers que l'humain effectue lors de l'ouverture ne sont pas présents. Au contraire en présence de celui-ci, nous montrons que la courbe "théorique" (mesuré par une personne extérieure) est bien corrélée avec la courbe de reconnaissance de l'ouverture de la bouche. L'expérimentateur produit des paliers lorsqu'il ouvre sa bouche, qui sont reproduits par la tête expressive. Ces courbes montrent très clairement que les neurones qui codaient initialement pour les 3 intensités motrices impliquant la reproduction de 3 intensités motrices par le système, ont été remplacés par un contrôle moteur plus précis où émerge des paliers intensifs grâce au champ de neurone. La figure 6-9 montre la reproduction de l'intensité motrice du haussement et froncement des sourcils. De la même manière que pour l'ouverture de la bouche, le robot est capable d'imiter plus finement ce groupement musculaire. De plus, nous montrons au travers de la figure 6-10 l'activité temporelle du champ de neurone. La séquence que l'humain produit est la suivante: ouverture, fermeture et ouverture de la bouche. Nous remarquons que des bulles d'activités se forment sur les neurones codant pour une bouche grande ouverte, plus la bouche se ferme moins ces bulles sont importantes, laissant se former une bulle d'activité sur les neurones codant une bouche fermée.

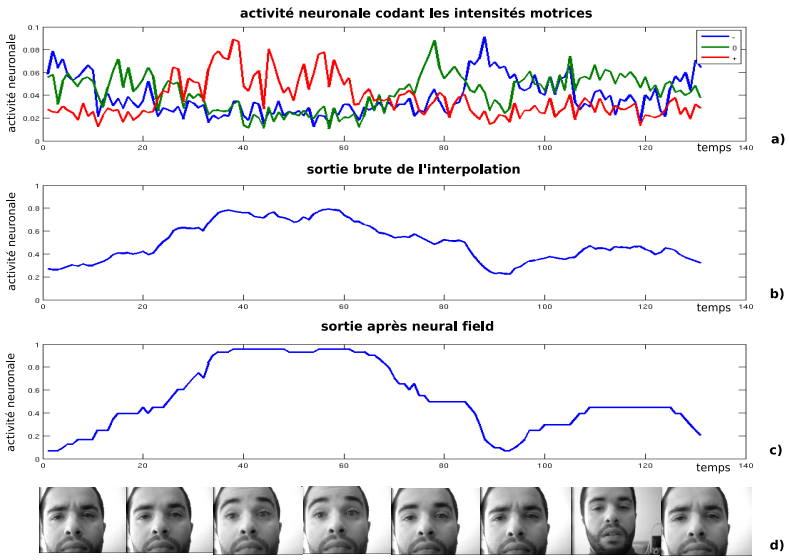


Fig. 6-9 – Intensité motrice du haussement et froncement des sourcils durant la phase de reproduction de la tête expressive. Ces courbes montrent l'activité analogique de plusieurs neurones qui sont impliqués dans le contrôle des sourcils. a) montre l'activité neuronale des 3 neurones codant chacun pour une intensité motrice (haussement, froncement, position intermédiaire des sourcils). b) montre l'interpolation des 3 neurones codant les 3 positions des sourcils. c) montre l'effet du champ de neurones sur la reconnaissance des intensités motrices des sourcils. d) montre la séquence d'image analysée.

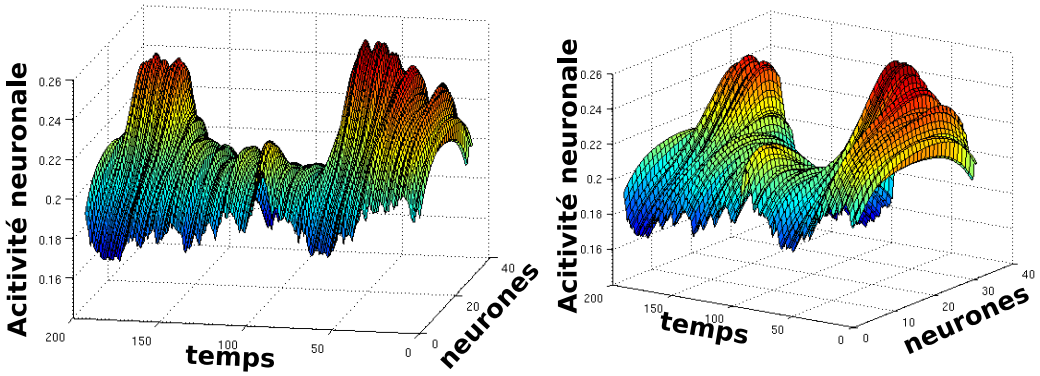


Fig. 6-10 – Activité temporelle du champ de neurones, visualisation sous 2 angles différents. Ces différentes courbes 3D montrent l'activité temporelle du champ de neurone lors d'une séquence de mimique faciale que l'humain réalise de manière lente et continue. La séquence est la suivante ouverture, fermeture et ouverture de la bouche.

6.2.5 Conclusion

Nous montrons dans cette section que le paradigme utilisé pour la reconnaissance des expressions faciales peut être appliqué pour la reconnaissance des intensités de plusieurs groupements musculaire (sourcils, ouverture de la bouche, coins de la bouche). Une tête expressive dôtée d'une architecture sensori-motrice de bas niveau peut apprendre des mouvements musculaires via un jeux d'imitation avec un expérimentateur (environnement social). L'humain imite les mouvements musculaires du robot durant 4 minutes. Une fois l'apprentissage stabilisé, les rôles sont

inversés, permettant à la tête expressive d'imiter les mimiques faciales de l'expérimentateur. Les

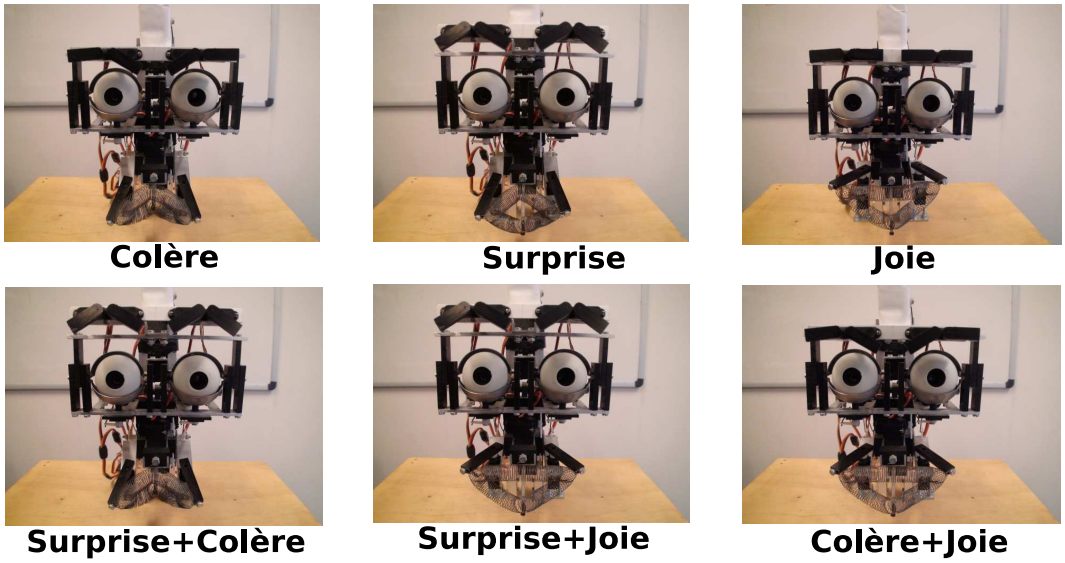


Fig. 6-11 – Expressions faciales primaires et secondaires que la tête expressive est capable de reproduire.

résultats mettent en évidence que le même modèle permet l'apprentissage (en ligne et autonome) de la reconnaissance des groupements musculaires et des expressions faciales émotionnelles. La propriété émergente de notre système réside dans la capacité du robot à produire des expressions faciales comme la joie, la surprise ou la colère mais il peut produire des expressions faciales secondaires (mélange d'expressions primaires: figure 6-11). Par exemple, si l'humain fronce les sourcils et sourit alors la tête expressive mimera l'expression du visage de l'humain. Cependant, la question du sens sémantique de ce qu'elle exprime, n'est pas résolu. Autrement dit, les commandes motrices ne sont pas liées aux états internes. Malgré les contraintes dues à un apprentissage en ligne, nous avons montré que des capacités de généralisation apparaissent pour des personnes n'ayant jamais interagi avec le dispositif. Cependant les résultats pourraient être meilleurs si le nombre d'individus appris était plus conséquent (notre base de donnée ne contient que 14 personnes). De plus, le nettoyage de la base de donnée améliorerait les résultats comme pour la reconnaissance des expressions faciales. En analysant plus finement la base de donnée, nous constatons que certains individus ont des difficultés à imiter certaines primitives motrices, par exemple certains expérimentateurs ne reproduisent pas correctement les différentes positions de la bouche (ouverte, fermée et position intermédiaire), la position intermédiaire est souvent confondue avec les deux autres. Enfin, différentes manières d'ouvrir la bouche et de sourire (en montrant les dents ou non) constituent beaucoup de variabilités pour permettre au système de reconnaitre les primitives motrices en n'ayant appris que sur un petit ensemble d'individus. En particulier, "ouvrir en grand la bouche" chez un sujet peut correspondre à "bouche mi-ouverte" chez un autre.

Nous avons egalement montré qu'il était faisable d'introduire une notion d'intensité motrice plus précise qu'une simple interpolation neuronale. Notre solution a été l'introduction des champs de neurones car ils possèdent les propriétés idéales pour ce type de contrôle moteur (suppression de régime oscillant, mémoire temporelle, etc). Les expériences ont montré qu'en partant d'une architecture ayant 3 neurones par primitives motrices (chaque neurone codant une intensité) donc une architecture capable de reproduire 3 intensités motrices, nous étions capables d'introduire une notion d'intensité neuronale par l'intermédiaire du champ neuronal. La tête de robot est

maintenant capable de reproduire les mimiques faciales de l'expérimentateur interagissant avec elle en ayant une notion d'intensité plus précise que 3 niveaux intensifs.

6.3 L'ajout de la reconnaissance des intensités expressives

Dans cette section, nous proposons deux modèles incorporant la notion d'intensité pour la reconnaissance des expressions faciales. En utilisant le même paradigme expérimental, il est possible de voir apparaître la notion d'intensité expressive pour une expression positive (la joie) et une expression ayant une connotation négative (la colère). Avoir une notion d'intensité expressive permet de transmettre des signaux plus précis au robot. Dans une perspective développementale, il est important de faire la distinction entre une joie intense et un léger sourire. Les signaux transmis au robot sont importants pour amorcer des comportements sociaux plus complexes.

6.3.1 La reconnaissance des expressions faciales influençant ou non la reconnaissance des primitives motrices

Deux modèles ont été développés pour la reconnaissance de l'intensité expressive et de la reconnaissance des primitives motrices. La figure 6-12 nous montre la première architecture développée. La reconnaissance des expressions faciales a le même statut qu'une primitive motrice (ouverture de la bouche ou encore haussement des sourcils). Cette architecture est capable d'imiter les différents groupements musculaires mais elle est également capable d'avoir une notion d'intensité expressive. Nous entendons par intensité expressive, la capacité à obtenir une valeur analogique. Nous avons fait le choix que cette valeur $\in [0, 1]$ (0 codant pour l'expression de colère et 1 codant pour la joie). Ce choix est fait pour des raisons d'affichage et pour manipuler qu'une grandeur pour les émotions. Pour simplifier, on prend une émotion positive (la joie) et une émotion négative (la colère) qui seront mises sur le même axe. Il s'agit d'un a priorie utilisé pour simplifier le réseau de neurones. Nous ne rentrerons pas dans une discussion sur modèle discret contre modèle dimensionnel. On aurait pu laisser les deux dimensions séparées mais l'affichage aurait été plus délicat.

Cependant, le contrôle des mimiques faciales est obtenu comme la conséquence de la combinaison des primitives motrices. De ce premier point de vue, le contrôle moteur ne serait pas influencé par les expressions émotionnelles. Ce qui laisse supposer que l'expressivité du robot n'est qu'une simple résonnance motrice et non émotionnelle. Tandis que l'autre modèle développé (Figure 6-13) montre l'influence de la reconnaissance des expressions émotionnelles sur la reconnaissance des groupements musculaires du visage. Ce modèle est bien entendue capable de reproduire des mouvements musculaires (primitives motrices) mais il est également capable de mimer les expressions émotionnelles. Ce modèle met en évidence l'influence de la reconnaissance des expressions émotionnelles sur le contrôle des mimiques faciales. Une première voie est capable d'apprendre les primitives motrices indépendamment les unes des autres et les expressions émotionnelles comme pour le modèle de la figure 6-12. Cependant une autre voie (flèche bleu) apprend l'association entre les primitives motrices et la prédiction d'un état interne (reconnaissance de l'expression émotionnelle). Par conséquent, aussi bien des caractéristiques visuelles de bas niveau que la prédiction d'un état interne sont capables de déclencher des configurations motrices particulières. Le principal avantage du modèle de la figure 6-13 contrairement au premier modèle est une robustesse à certaines lésions. Imaginons des lésions dans les réseaux de neurones chargés de la reconnaissance des primitives motrices, l'architecture globale fonctionnera toujours car la prédiction de l'état interne (expressions émotionnelles) influencera le contrôle

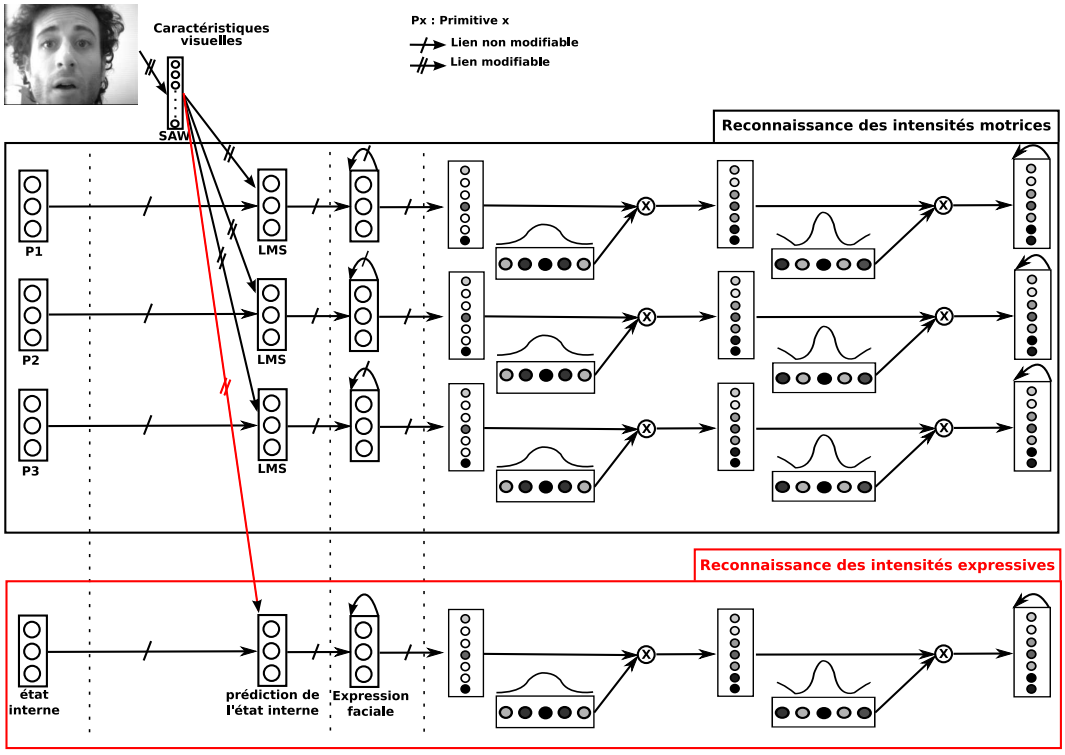


Fig. 6-12 – Modéle de la reconnaissance des intensités motrices et des intensités expressives. Le réseau de neurone est toujours capable d'apprendre les différentes primitives (P1, P2, P3) et est capable d'apprendre également 2 expressions faciales l'une positive et l'autre négative (la joie et la colère) ainsi que le visage neutre. Dans ce modèle la reconnaissance des expressions faciales n'influencent pas la reconnaissance des primitives motrices.

des mimiques faciales. Ce qui laisse suggèrer que des aires corticales distinctes seraient chargées des différentes mimiques faciales, par exemple une sous aire par mimique. La reconnaissance des expressions faciales émotionnelles serait induite par un ensemble d'aires corticales tandis que la reconnaissance des groupements musculaires serait controlée par un autre ensemble d'aire corticale.

6.3.2 Résultats

De la même manière que pour la reconnaissance des primitives motrices, le robot est capable d'imiter plusieurs niveaux d'intensités expressives. Nous obtenons d'une part une notion d'intensité expressive émotionnelle purement interne par l'intermédiaire de l'état interne prédit (catégories émotionnelles dicrètes) mais également purement motrice. Autrement dit, le système est capable de générer des signaux internes comme la conséquence d'une perception de l'expression faciale venant de son environnement (expérimentateur interagissant avec le robot) et d'autre part d'exprimer cette expression (contrôle des mimiques faciales). La figure 6-14 et la figure 6-15 montrent ces deux aspects sur une même séquence d'interaction avec un partenaire humain. La tête expressive a appris les primitives motrices ainsi que les expressions émotionnelles durant 4 minutes, ces figures montrent la phase de reproduction (l'humain exprime des mimiques faciales et le robot l'imite). La figure 6-14 illustre le modèle de la figure 6-12 (la reconnaissance de l'expressivité faciale n'influence pas la reconnaissance des primitives motrices). Plusieurs signaux sont

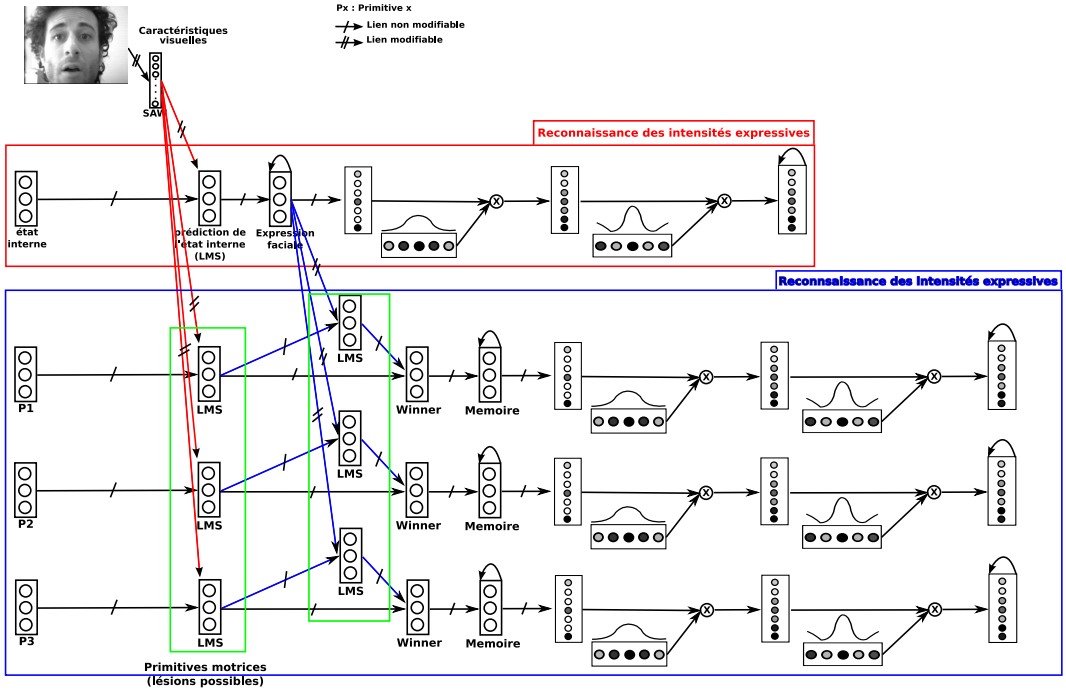


Fig. 6-13 – Modèle de la reconnaissance des intensités motrices et des intensités expressives. Le réseau de neurone est toujours capable d'apprendre les différentes primitives (P1, P2, P3) et est capable d'apprendre également 2 expressions faciales l'une positive et l'autre négative (la joie et la colère) ainsi que le visage neutre. Dans ce modèle la reconnaissance des expressions faciales influencent la reconnaissance des primitives motrices.

analysés correspondant chacun à la sortie du "champ de neurones" pour la reconnaissance des expressions faciales et des primitives motrices (figure 6-14-b-c-d-e) cependant il est intéressant de mettre en évidence plusieurs phases dans cette séquence de reproduction. La première phase montre l'humain exprimant une expression faciale de surprise, on observe une activité moyenne pour l'intensité expressive (correspond ni à de la joie ni à de la colère), une activité élevée pour le haussement des sourcils ainsi que l'ouverture de la bouche correspondant aux caractéristiques de l'expression de surprise et une activité moyenne pour les coins de la bouche (l'humain n'exprime ni un sourire ni une moue). La deuxième phase montre l'expérimentateur exprimant un visage de joie. On observe une activité élevée sur le versant expression émotionnelle ainsi que sur les coins de la bouche (sourire), une activité inférieure à la moyenne pour l'ouverture de la bouche et supérieur à la moyenne pour le haussement des sourcils. Ces signaux correspondent bien aux caractéristiques de l'expression de joie (sourire, haussement des sourcils et bouche peu ouverte). La troisième phase montre l'humain exprimant une expression secondaire (mélange de 2 expressions faciales primaires) de la joie et de la colère. On observe une activité basse pour l'expression émotionnelle ainsi qu'une activité basse pour le haussement des sourcils (équivaut à un froncement des sourcils), une activité moyenne pour l'ouverture de la bouche et une activité élevée pour les coins de la bouche (reconnaissance du sourire). Cette phase met en évidence, la capacité du robot à reproduire des expressions complexes. La quatrième phase montre le partenaire humain exprimant une expression de joie intense, on remarque que les caractéristiques motrices utiles à l'expression de joie sont plus hautes que pour la deuxième phase. Le sourire est à une activité légèrement supérieure à la quatrième phase tandis que le haussement des sourcils est nettement supérieur. Cette phase met en évidence la capacité du système à créer des paliers

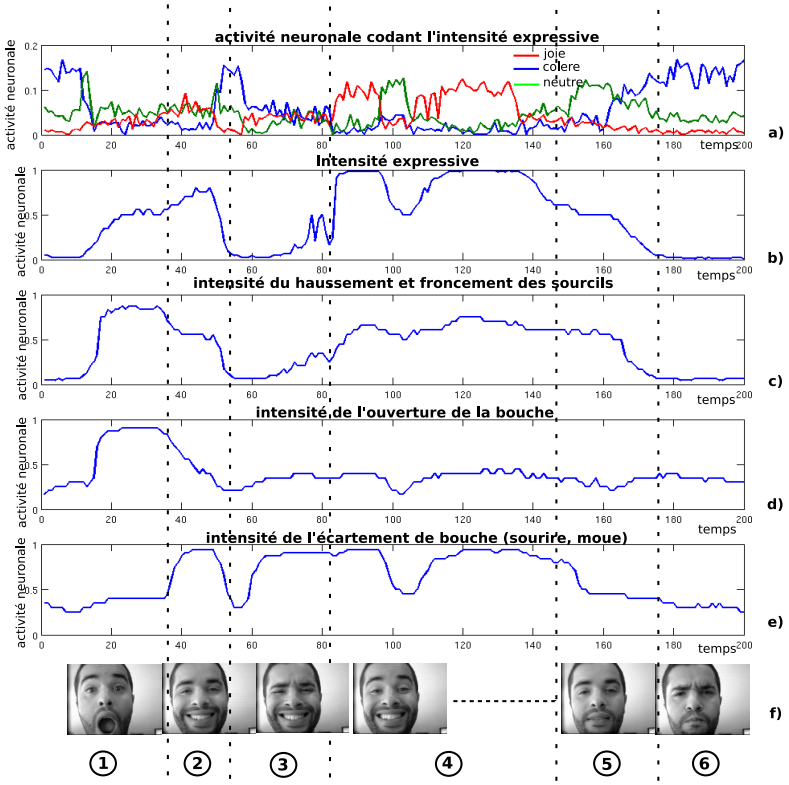


Fig. 6-14 – Résultats lors d'une séquence d'interaction en ligne avec la tête de robot expressive (le robot imite l'humain). Ces courbes montrent les différents signaux utiles à la reconnaissance des expressions émotionnelles, cependant ces dernières n'influencent pas la reconnaissance des intensités motrices. a) montre l'activité neuronale des 3 neurones codant pour les 2 expressions faciales (joie et colère) et le visage neutre. b) montre l'effet du neural field sur la reconnaissance des intensités expressives émotionnelles. c) montre l'effet du neural field sur la reconnaissance des intensités motrices pour le haussement et froncement des sourcils. d) montre l'effet du neural field sur la reconnaissance des intensités motrices pour l'ouverture de la bouche. e) montre l'effet du neural field sur la reconnaissance des intensités motrices pour les coins de la bouche (sourire et moue). d) montre la séquence d'image analysée.

intensifs aussi bien sur le plan expressif pur (contrôle moteur) que sur le plan "émotionnel" (prédiction des états internes). La cinquième phase met en évidence la capacité du système à exprimer un visage neutre, on observe que toutes les activités sont proches de la valeur moyenne aussi bien pour l'intensité expressive que pour les primitives motrices. La tête expressive reproduit un visage neutre mais ne "ressent" ni de la joie ni de la colère. Cela souligne que le robot est complètement neutre vis à vis de son environnement aussi bien du côté de l'expressivité que du côté "émotionnel". Enfin la sixième phase montre un robot capable de reproduire l'expression de la colère et de la "ressentir". L'intensité expressive tend vers 0 correspondant bien à l'expression de colère, toutes les caractéristiques motrices sont présentes pour la reproduction des mimiques: froncement des sourcils, bouche légèrement fermée et légère moue. Ces différentes phases montrent la capacité du système à pouvoir reproduire (mimer) des expressions faciales primaires et secondaires ainsi que la capacité à pouvoir "ressentir" ces expressions émotionnelles. Nous entendons par "ressentir", le simple fait de pouvoir prédire des états internes grâce à la perception des expressions faciales (interaction avec l'environnement). Cependant, la moindre perturbation neuronale de ce type de réseau (par exemple, lésion neuronale) détruirais certaines

capacités. Par exemple, des lésions au niveau des primitives motrices engendreraient une incapacité à reproduire des expressions faciales alors que le robot serait toujours capable de "ressentir" les émotions associées.

Le modèle décrit figure 6-13 est robuste à certaines ablations, la figure 6-15 montre les activités des neurones utiles à la reconnaissance de l'expression émotionnelle ainsi qu'à la reconnaissance des différentes primitives motrices (sans lésion). L'influence de l'expression émotionnelle lisse davantage la reconnaissance des primitives motrices. Ces résultats illustrent la capacité du robot à reproduire et "ressentir" les expressions émotionnelles aussi bien que le premier modèle. Pour

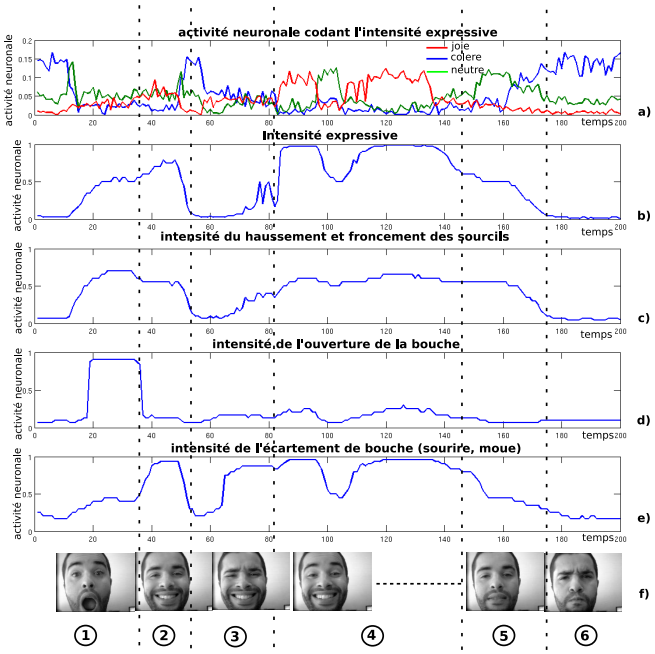


Fig. 6-15 – Résultats lors d'une séquence d'interaction en ligne avec la tête de robot expressive (le robot imite l'humain). Ces courbes montrent les différents signaux utiles à la reconnaissance des expressions émotionnelles ainsi que leurs influences sur la reconnaissance des primitives motrices. a) montre l'activité neuronale des 3 neurones codant pour les 2 expressions faciales (joie et colère) et le visage neutre. b) montre l'effet du neural field sur la reconnaissance des intensités expressives émotionnelles. c) montre l'effet du neural field sur la reconnaissance des intensités motrices pour le haussement et froncement des sourcils. d) montre l'effet du neural field sur la reconnaissance des intensités motrices pour l'ouverture de la bouche. e) montre l'effet du neural field sur la reconnaissance des intensités motrices pour les coins de la bouche (sourire et moue). d) montre la séquence d'image analysée.

tester la robustesse du modèle, nous avons introduit des lésions dans le réseau de neurones. Ces lésions peuvent correspondre à l'ablation de structures neuronales importantes au fonctionnement du système global. La figure 6-16 témoigne de la capacité du système à pouvoir exprimer des expressions émotionnelles même après l'ablation des groupes de neurones dédiés à la reconnaissance des primitives motrices. Les aspects intéressants de ces résultats sont la compétence du robot à pouvoir exprimer des expressions émotionnelles grâce à l'association apprise (flèche bleu: figure 6-13)entre les primitives motrices et l'expression émotionnelle (prédiction de l'état interne). Le robot est toujours capable d'exprimer soit une expression positive soit une expression négative mais est dans l'incapacité de mimer d'autres expressions faciales. Durant la premiére phase par exemple, l'expérimentateur exprime de la surprise en ouvrant la bouche, le robot ne mime pas l'expression de son visage car l'ouverture de la bouche n'est pas une caractéristique des

expressions émotionnelles apprises. Les ablations au niveau de la reconnaissance des primitives motrices n'empêche pas la reproduction des expressions de joie et de colère mais ne permet pas la reproduction de certaines caractéristiques comme la simple ouverture de bouche.

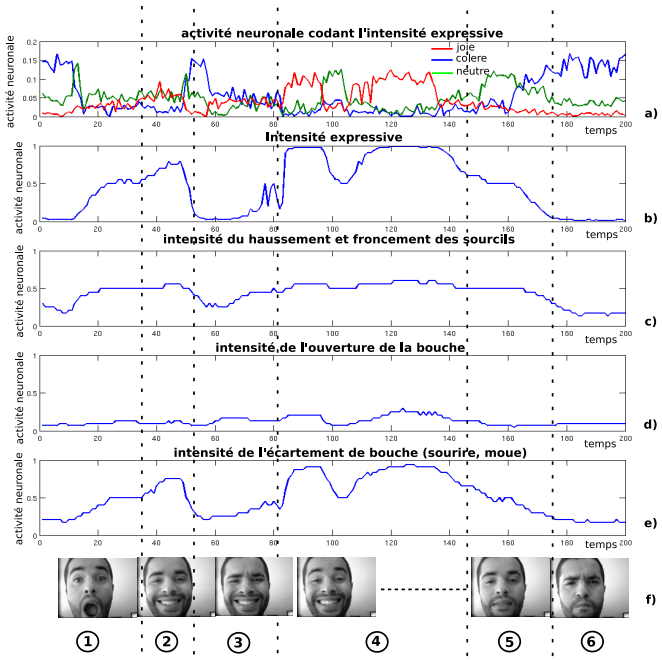


Fig. 6-16 – Résultats lors d'une séquence d'interaction en ligne avec la tête de robot expressive (le robot imite l'humain). Ces courbes montrent les différents signaux utiles à la reconnaissance des expressions émotionnelles ainsi que leurs influences après lésion des groupes de neurones apprenant les différentes primitives motrices. a) montre l'activité neuronale des 3 neurones codant pour les 2 expressions faciales (joie et colère) et le visage neutre. b) montre l'effet du neural field sur la reconnaissance des intensités expressives émotionnelles. c) montre l'effet du neural field sur la reconnaissance des intensités motrices pour le haussement et froncement des sourcils. d) montre l'effet du neural field sur la reconnaissance des intensités motrices pour l'ouverture de la bouche. e) montre l'effet du neural field sur la reconnaissance des intensités motrices pour les coins de la bouche (sourire et moue). d) montre la séquence d'image analysée.

6.3.3 Conclusion

Les différents modèles développés (figure 6-12 et figure 6-13) montrent la capacité du robot à pouvoir exprimer des expressions émotionnelles (joie, colère) ainsi que certaines primitives motrices comme l'ouverture de la bouche ou le haussement des sourcils. Une notion d'intensité est introduite dans l'architecture par l'intermédiaire des champs neuronaux. De plus, le robot est capable de prédire des états internes (état "émotionnel") en ayant plusieurs niveaux d'intensité. Aussi bien sur le plan expressif et que sur le plan du ressenti des états internes, la notion d'intensité émerge du champ de neurones. Le modèle de la figure 6-13 (avec la réentrance de la reconnaissance de l'état émotionnel sur la reconnaissance des primitives expressives) semble plus plausible car il paraît clair que la reconnaissance de l'expression émotionnelle peut influencer sa propre expressivité. Mais également lors de l'introduction de certaines lésions neuronales (lésions des réseaux de neurones déclenchant la reconnaissance des primitives motrices) où le robot possède toujours la capacité à imiter les expressions faciales de joie et de colère, cependant

il perd la capacité à réaliser des mimiques faciales simples comme l'ouverture de la bouche où le haussement des sourcils.

6.4 Conclusion

Nous avons montré qu'une architecture sensori-motrice est capable d'apprendre à reproduire des mouvements musculaires grâce à jeux d'imitation. La tête expressive possède la capacité de reproduire chaque groupement musculaire les uns indépendamment des autres mais possède également une propriété très intéressante qui est la capacité à reproduire des expressions secondaires comme un sourire mélangé avec de la colère.

D'autre part, la notion d'intensité est introduite par l'intermédiaire de mécanisme neuronale. Les champs de neurones ont des propriétés intéressantes permettant l'émergence d'une notion d'intensité musculaire. Par exemple, la tête de robot est capable d'ouvrir plus ou moins la bouche suivant le visage du partenaire humain.

Ces propriétés ont été introduites pour la reconnaissance d'expressions faciales, la tête expressive est maintenant capable de "ressentir" le niveau expressif de l'expérimentateur. Cette capacité est importante pour améliorer l'interaction avec le robot, le robot doit faire la distinction entre un léger sourire et un sourire intense car le sens n'est pas le même lors d'un interaction avec l'environnement. Cette notion d'intensité n'est pas simplement une amélioration algorithmique mais elle peut nous faire poser des questions sur les modèles émotionnelles (discret, dimensionnelle ...).

Troisième partie

Interaction triadique: Expérimentateur - Objet - "Bébé robot"

Chapitre 7

Référencement social dans le cadre d'une manipulation d'objet

7.1 Introduction

Dans ce travail, nous nous intéressons à comprendre comment des interactions émotionnelles avec un partenaire social peuvent amorcer un comportement plus complexe tel que le référencement social. L'idée que nous défondons, est que le référencement social aussi bien que la reconnaissance d'expressions faciales peut émerger d'une architecture sensori-motrice. Sans connaissance de ce qu'est l'autre, autrement dit sans intersubjectivité, le robot devra être capable d'apprendre et de développer des capacités sociales. Nous soutenons l'idée que le référencement social ainsi que d'autres comportements peuvent être amorcés par une simple cascade d'architectures sensori-motrices qui à la base ne sont pas dédiées aux interactions sociales.

En particulier, le référencement social est la capacité à récolter des informations au travers des interactions émotionnelles. Le référencement social est un processus développemental incorporant la capacité à reconnaitre, comprendre, modifier son comportement en fonction de l'état émotionnel du partenaire sociale. Cela permet à un enfant ou à un robot(Breazeal et al., 2005; Thomaz et al., 2005) de chercher de l'information venant d'un autre individu et d'utiliser cette information pour guider son comportement vis à vis d'un objet ou d'un événement qui est ambigüe(Klinnert et al., 1983a). De manière générale, les situations qui sont ambigües pour un nouveau né, sont les situations qui sont toutes nouvelles pour lui par exemple un objet avec lequel le bébé n'a encore jamais interagi ou encore une situation environnementale nouvelle comme une falaise.

L'ensemble des informations qui proviennent des interactions émotionnelles semblent être une manière rapide et efficace de déclencher l'apprentissage. De plus, les émotions sont de notre point de vue un excellent moyen de communication. Elles permettent par la même occasion de désambiguiser certaines situations. Ceci est encore plus évident dans des étapes précoces du développement cognitif humain et des primates (Russell et al., 1997). L'interaction entre les deux partenaires (mère et bébé) est une véritable symbiose: l'enfant communique son attente dans certaines situations nouvelles et la mère est toujours présente pour l'aider ce qui peut être une source de grande simplification dans le monde de la robotique.

Dans cette forme de cognition sociale, un bon (ou un mauvais) objet (ou évènement) est

identifié ou signalé grâce à un message émotionnel sans label verbal évitant ainsi de devoir comprendre des expressions telles que "ne touche pas à cette objet. Il est dangereux". Nous savons que la valeur émotionnelle peut être fournie par un certain nombre de modalités comme les expressions faciales, la voix, les gestes etc. Nous avons choisi d'utiliser les expressions faciales car non seulement elles sont un excellent moyen pour communiquer des informations importantes dans des situations ambigües mais aussi elles peuvent être apprises rapidement et de façon autonome comme on l'a vu dans le chapitre 4. D'autre part, les expressions faciales primaires sont universelles, elles sont reconnues et portent le même sens dans toutes les cultures. L'idée que nous allons développer est que le référencement social peut émerger à travers une interaction basée sur de simples mécanismes associatifs (perception-action).

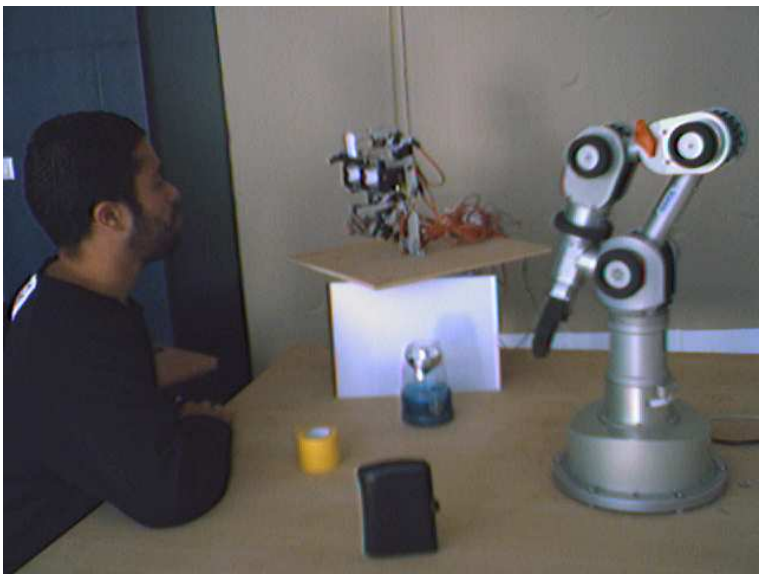


Fig. 7-1 – Set-up expérimental pour le référencement social. Nous utilisons une tête robotique capable de reconnaitre et d'imiter les expressions faciales et un bras robotique capable de prendre des objets à valence positive ainsi que de s'éloigner des objets dits négatifs.

Le rôle de l'interaction est souvent sous-estimé en robotique pour l'apprentissage de comportements comme par exemple l'apprentissage par démonstration avec des phases d'apprentissage figées et/ou l'apprentissage hors ligne à partir d'une base de donnée (Calinon et al., 2007; Billard et al., 2008). De notre point de vue, nous essayons de mettre en évidence que le développement autonome d'un robot ne peut être envisagé qu'au travers des interactions avec son environnement (Figure 7-1). Sans cette capacité à interagir avec l'environnement physique et social, le robot ne peut finalement pas agir et le développement de capacités cognitives semble difficile. Particulièrement, la notion d'agent autonome est valable si et seulement si l'agent, son environnement et leurs interactions sont pris en considération. L'ensemble de ces interactions pourra faire émerger la notion d'agentivité (se savoir auteur de ses propres actions).

Cette étude peut donc être à la fois importante pour le domaine de la robotique autonome et très intéressante pour la psychologie du développement. Les modèles computationnels que nous développons, peuvent être considérés comme une nouvelle manière de comprendre ce qui est observer chez les jeunes enfants. Dans notre paradigme expérimental, le robot est un bébé qui

n'a apriori aucune information sur le monde mais qui à force d'interagir avec son environnement extérieur apprend des "choses" d'une complexité toujours croissante. Au début, le "robot bébé" ne connaît rien sur le monde, il possède que des comportements réflexes.

Nos travaux exploitent l'ambiguïté de la perception, c'est à dire dans le cas présent la non capacité à faire la différence entre son propre corps et celui de l'autre ou entre ses propres actions et celles de l'autre. L'ambiguité de la perception couplée avec un système homéostatique est suffisant pour déclencher premièrement la reconnaissance d'expression faciale et par la suite l'apprentissage de valeurs émotionnelles associées à des objets qui sont arbitraires situés dans l'environnement du robot.

7.2 Set-up expérimental

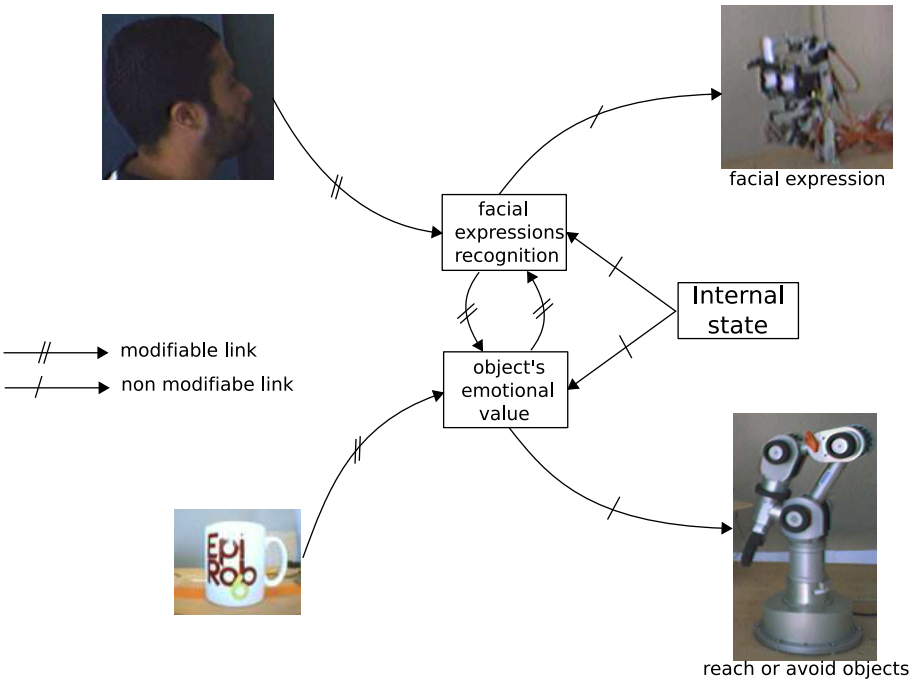


Fig. 7-2 – Modèle simplifié du référencement social. Mise en évidence des interactions bidirectionnelles des structures indispensables à cette capacité. Le référencement social n'est possible qu'au travers l'interaction émotionnelle avec un partenaire humain. On considère les 2 agents (Homme - Robot) comme un seul système interagissant en boucle fermée avec l'environnement.

Dans notre expérience de référencement social (figure 7-2 et figure 7-3), nous avons le set-up suivant: une tête de robot pouvant imiter et reconnaître les expressions faciales, un bras Katana pouvant interagir avec différents objets présents dans son espace péri-corporel et une autre caméra regardant justement l'espace péri-corporel du robot. Une conséquence de ce set-up expérimental est que notre pseudo-robot humanoide (tête et bras) a des capacités d'interactions avec son environnement aussi bien social que physique et par la même occasion peut manipuler des objets qui lui procureront un certain plaisir social. Dans l'architecture développée, le robot apprend à attraper les objets qui sont "positifs" et au contraire à éviter les objets qui sont "négatifs". Un seul objet est présent dans l'espace péri-corporel du robot. Nous mettons en

évidence dans cette expérience que l'interaction émotionnelle via l'expressivité faciale, permet de communiquer un état émotionnel à un robot afin d'influencer son comportement. Les expressions faciales émotionnelles sont un excellent média de communication non-verbal entre l'humain et le robot.

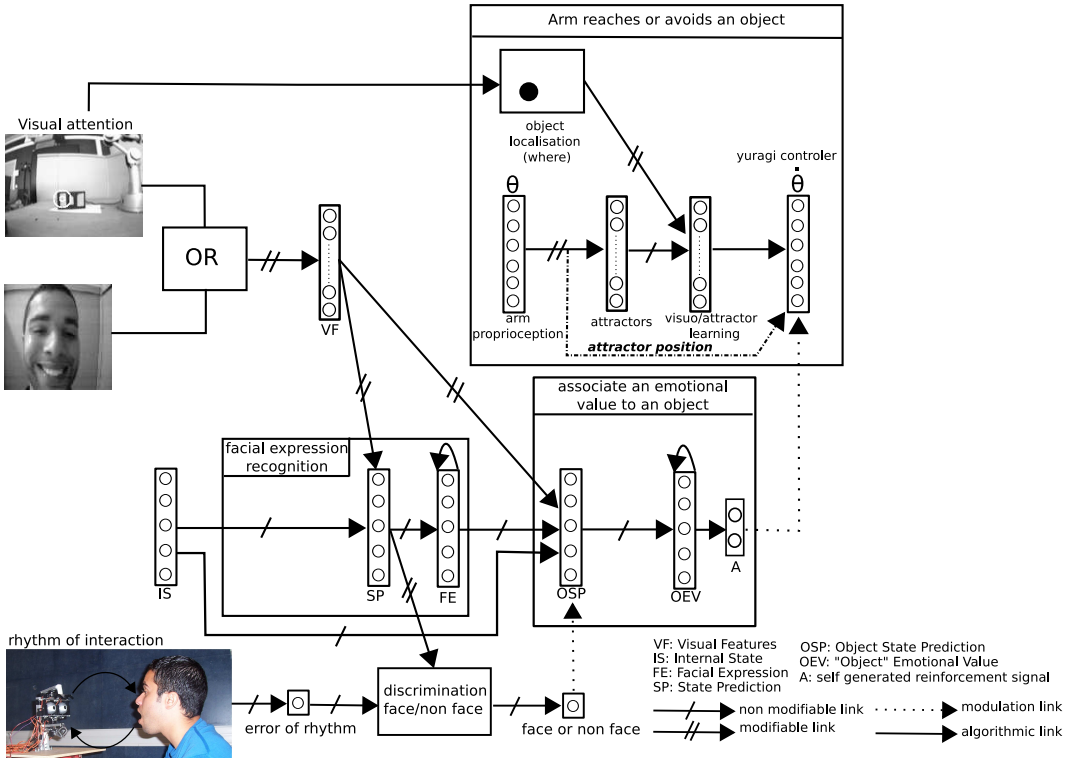


Fig. 7-3 – Modèle de référencement social. Le référencement social émerge des interactions sensori-motrices entre la reconnaissance des expressions faciale, l'attribution de valeur à des objets et l'apprentissage visuo-moteur pour le contrôle du bras robotique. Notre modèle architectural est capable d'apprendre et de reconnaitre les expressions faciales ainsi que de discriminer si un partenaire interagit avec lui (détection d'un visage). Le robot focalise sur les objets grâce à un mécanisme pré-attentionnel très rudimentaire utilisant des caractéristiques issus des filtres de Gabor et de la couleur. Après l'apprentissage visuo-moteur, le bras robotique peut dans son espace de travail attraper ou s'éloigner de certains objets qui lui procurent du plaisir ou réciproquement du déplaisir et ceci grâce à un signal de renforcement auto-généré nommé A (valeur émotionnelle venant de la reconnaissance d'expression faciale). A est construit comme la conséquence directe de la reconnaissance des expressions faciales (A_1 est le neurone correspondant à l'expression faciale de joie, A_2 est le neurone correspondant à l'expression faciale de colère). "OR" signifiant simplement qu'une seule image est analysée à la fois.

Le modèle proposé (figure 7-3) souligne l'aspect constructiviste de notre approche car il s'attache à étudier les mécanismes et processus permettant la construction du référencement social à partir d'éléments déjà intégrés comme la reconnaissance des expressions faciales ou la détection de visage. Ce modèle se base sur des compétences déjà acquises par le robot.

Après l'apprentissage visuo-moteur du bras, plusieurs positions peuvent être atteintes par celui-ci (Andry et al., 2001). Une position visuelle correspond à une ou plusieurs configurations motrices du bras (une configuration motrice correspond à un vecteur proprioceptif), nous les appellerons des attracteurs. Ces attracteurs entrainent le bras dans un bassin d'attraction

centré sur la position de l'objet que le robot souhaite attraper. Ce contrôle est réalisé grâce à un système dynamique dont le but est de lisser la trajectoire (Fukuyori et al., 2008). Ce système dynamique utilise aussi un signal de renforcement donnant beaucoup ou peu d'importance à certains attracteurs. Par exemple, une récompense peut être attribuée si le bras suit la bonne direction. Dans le cas contraire, l'influence de l'attracteur est réduite. Le renforcement est ici de nature émotionnelle (e.g l'expression de joie est un signal positif tandis que l'expression de colère est considérer comme un signal négatif).

Un scénario possible est le suivant: le robot est dans un état émotionnel neutre tandis que le partenaire humain exprime de la joie en présence d'un objet particulier. Le comportement naturel du robot sera de changer son propre état émotionnel en passant de l'état neutre à l'état de joie et d'associer une valeur positive à l'objet qui était neutre à l'origine. Dans le cas contraire, si l'humain exprime une expression négative (expression de colère), la valeur associée à l'objet sera négative. Par la suite, le bras robotisé cherchera à attraper ou à éviter l'objet selon la valeur émotionnelle qui lui a été associée. En d'autres termes, la valeur émotionnelle associée à l'objet est le signal de renforcement que le bras utilise pour réaliser son contrôle moteur (exactement comme l'expression faciale du partenaire humain ou son propre état interne). Dans ce scénario, nous essayons de mettre en évidence la dimension émotionnelle. L'émotion joue un rôle non seulement pour essayer de réguler nos propres signaux internes mais également pour adaptater le comportement à l'environnement réel.

7.3 Reconnaissance de l'état émotionnel du partenaire

Cette section montre comment la reconnaissance des expressions faciales et la détection de visage ont été intégrées dans le but d'associer une valeur à un objet donnée. Nous montrons comment l'expressivité est utilisée dans le cas où un partenaire social interagit avec l'ensemble du dispositif robotique. Il est nécessaire dans ce scénario expérimental de détecter la présence de l'expérimentateur car la tête expressive produit constamment des expressions faciales même si aucun humain n'interagit avec elle. Dans ce cas, la reconnaissance de l'expression faciale par le robot doit être prise en compte seulement lors de la présence de l'expérimentateur. Cette solution évite l'envoie de messages émotionnels en l'absence d'humain.

Cette architecture (figure 7-4) permet au robot de reconnaitre des caractéristiques visuelles de l'expérimentateur et d'apprendre les caractéristiques expressives du sujet humain qui sont corrélées avec ses propres états internes. En parallèle, une autre structure neuronale apprend à détecter la présence d'un humain dans le champ visuel. Ce modèle tend à reconnaître les caractéristiques expressives d'un sujet humain tout en discriminant les caractéristiques non expressives relatives à l'absence de l'expérimentateur.

La figure 7-5 montre une séquence entre l'environnement et la tête expressive. Deux aspects sont mis en évidence: le premier est la détection du partenaire humain et la seconde est l'utilisation du signal émotionnel seulement dans le cas où la présence du partenaire est détectée. Cette figure illustre les deux cas pouvant être rencontrés:

– L'expérimentateur communique expressivement avec le robot. Sa présence est détectée par l'intermédiaire du réseau de neurones apprenant la détection de visage. Dans ce cas, l'expression reconnue pourra être utilisée pour communiquer de l'information sur l'environnement extérieur.

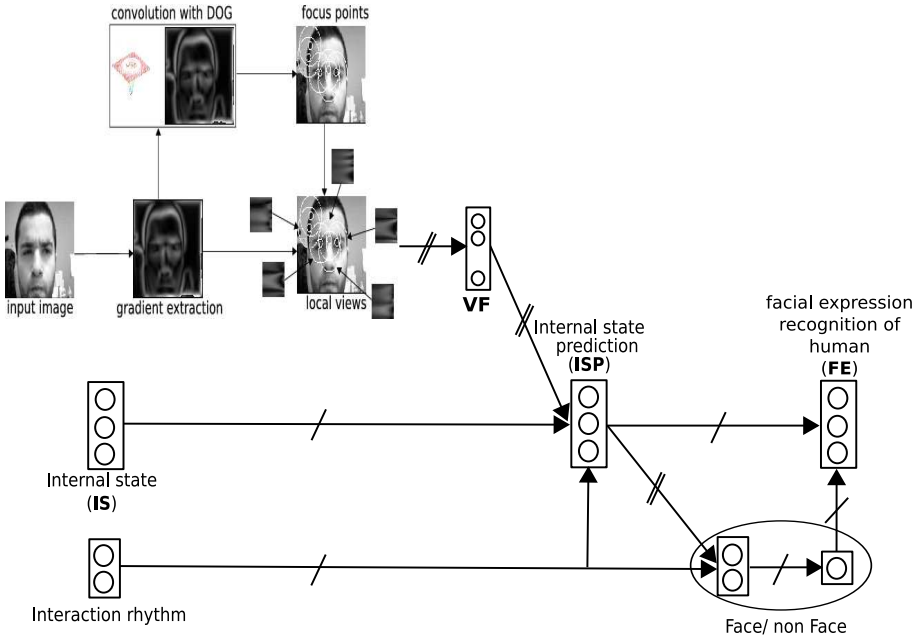


Fig. 7-4 – Modèle intégrant la reconnaissance des expressions faciales et la détection de visage.

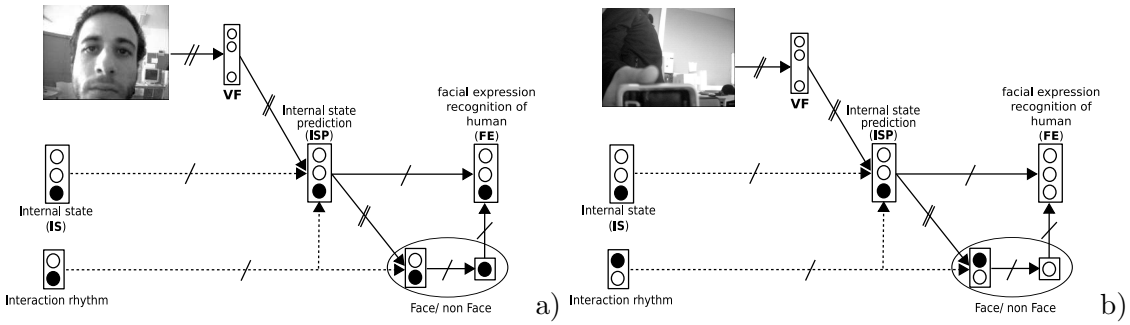


Fig. 7-5 – Modèle montrant une séquence interactive. a) montre un expérimentateur produisant une expression faciale. Sa présence est détectée et par conséquent l'expression pourra être utilisée pour fournir des informations sur l'environnement. b) montre la capacité du modèle à ne pas détecter de présence humaine impliquant l'inhibition de la reconnaissance de l'expression faciale (aucune activité pour le groupe FE).

– Aucune présence physique n'est détectée induisant. L'état interne prédit n'est pas propagé: aucune activité dans le groupe de reconnaissance FE (facial expressions). Dans ce cas, l'information expressive qui aurait pu être détectée malgré la non présence d'un visage, n'est pas utilisée.

A ce stage du développement, la tête de robot est capable de reconnaître et comprendre les expressions faciales du sujet humain. Le robot peut maintenant associer des signaux émotionnels à l'environnement comme une conséquence de l'interaction. L'expression émotionnelle est vue comme un moyen de communication pouvant transmettre des informations sur l'environnement

(Figure 7-6).

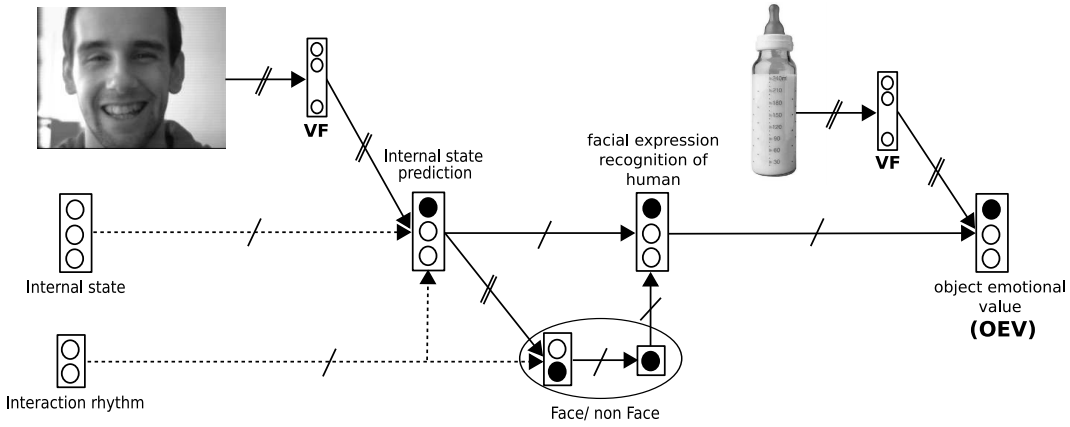


Fig. 7-6 – Modèle simplifié montrant la transmission de l'expression émotionnelle du partenaire humain en direction d'un objet se situant dans le champ d'action du robot.

7.4 Association d'une valeur émotionnelle à un objet

7.4.1 Vue d'ensemble

Après que le partenaire humain a imité la tête de robot durant 2 min environ, la tête robotique est capable de reconnaître les expressions de l'humain ainsi que de les reproduire. Une fois que cet apprentissage est réalisé, l'humain peut interagir avec la tête de robot pour associer une valeur émotionnelle à un objet (valeur positive ou négative).

Le réseau de neurones traite de la même manière les signaux qui viennent de l'état interne du robot et les signaux qui lui sont corrélés (par exemple les expressions faciales). Nous rappelons qu'un état interne peut déclencher une expression faciale de même que la reconnaissance d'une expression faciale. En cas de conflit, les poids venant de l'état interne contrôle en priorité l'expressivité du robot. L'état interne est toujours prioritaire. Le robot affiche donc une expression en accord avec l'état interne plutôt que celle issue de la reconnaissance de l'expression de son partenaire. Cette hypothèse est relativement forte mais elle nous évite l'utilisation de structures complexes permettant un contrôle volontaire de l'expression faciale. En l'absence de l'état interne, la reconnaissance de l'expression faciale induit un état qui peut être associé à l'objet présent dans le champ de vision du robot (simple chaine de conditionnement: la figure 7-6 montre une vue d'ensemble de cette cascade associative "simplifiée").

L'architecture détaillée est illustrée dans la figure 7-3. Un conditionnement classique est utilisé pour permettre l'association entre l'expression faciale de l'humain et certaines zones saillantes de l'image correspondant à l'objet.

7.4.2 Mécanisme de vision pré-attentive

Le mécanisme de vision pré-attentive attire le regard du robot sur certaines zones de l'environnement. Ce processus possède la capacité à chercher de l'information sur les zones intéressantes de l'environnement, en se focalisant particulièrement sur un objet attirant l'attention du robot.

Le processus attentionnel utilisé dans ce modèle est très simple puisque le robot focalise à la fois sur des zones colorées et sur les zones texturées. Les différentes primitives sont sommées sur une carte de saillance où la zone d'intêret est trouvée grâce à un mécanisme de compétition WTA (Figure 7-7). Cependant, des modèles plus sophistiqués peuvent être utilisés comme par

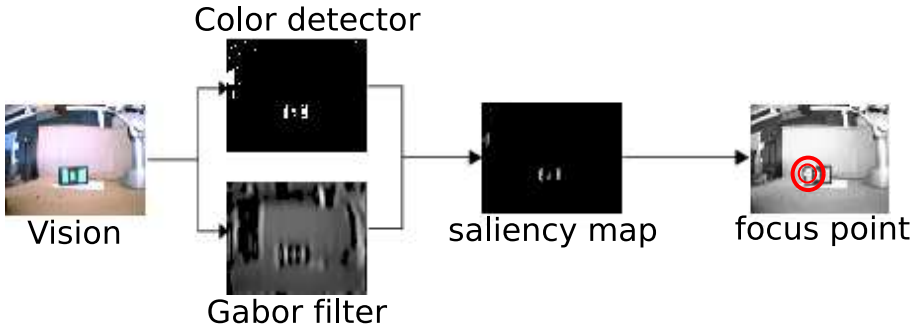


Fig. 7-7 – Vision pré-attentive. Le robot peut focaliser sur certains objets dans l'espace de travail. Une carte de saillance est activée pour focaliser sur les zones intéressantes de l'image. Les primitives visuelles calculées indépendamment (couleur, caractéristiques issus de filtres de Gabor) sont fusionnées par la suite pour trouver la zone que le robot doit analyser.

exemple (Gallet et al., 1998; Itti and Koch, 2001; Chevallier and Tarroux, 2008). Néanmoins, ces modèles sont souvent un peu plus couteux en temps de calcul mais ont l'avantage de mimer les saccades oculaires de l'humain. Ici, on se contente d'un mécanisme pré-attentif simple qui suffit à focaliser le champ de vision du robot sur un objet de l'environnement.

Une fois cette zone d'intêret sélectionnée grâce au mécanisme de vision pré-attentive, des points de focalisation sont extraits exclusivement sur cette zone pour ensuite être associées avec la valeur émotionnelle transmise par l'expressivité faciale du partenaire (figure 7-8).

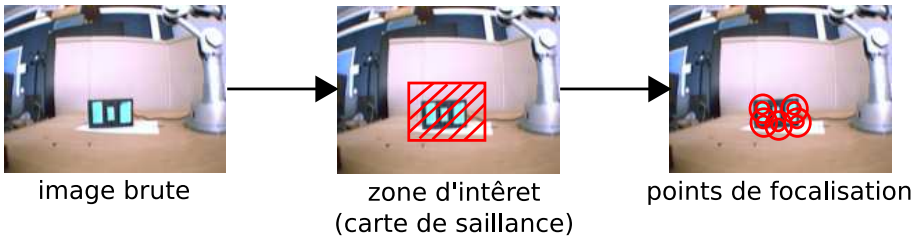


Fig. 7-8 – Chaîne de traitement visuelle. Le robot focalise sur une zone d'intêret dans l'environnement (zone saillante: couleurs et textures). Une fois cette zone sélectionnées, des points de focalisations sont extraits sur cette zonne grâce aux Différences de Gaussiennes (DOG).

7.4.3 Stabilité du système de vision pré-attentive

Nous discutons ici de la stabilité visuelle du processus de vision pré-attentive. Le but de cette section est d'obtenir une carte de saillance avec les caractéristiques décrites dans la section précédente (couleur, contraste, ...) et qui soit robuste au bruit de l'image. L'environnement naturel ainsi que la qualité moyenne du matériel utilisé (caméra firewire bas de gamme) apportent des perturbations non négligeables sur le champ visuel du robot. En d'autres termes, nous cherchons à augmenter la stabilité de la focalisation du robot vis à vis du bruit dans l'image.

De nombreux modèles existent notamment (Rougier and Vitay, 2006; Rougier, 2009) qui se rapprochent de nos travaux. Ainsi, lorsqu'on présente plusieurs images similaires au dispositif robotique ayant pour seule différence des pertubations causées par la qualité du capteur, nous souhaiterions que la zone pré-attentive trouvée reste stable dans l'espace (x,y). Les propriétés du champ de neurones (Amari, 1977) sont idéales pour ce type de problème. Ce modèle posséde des propriétés de stabilité dans l'espace et dans le temps ainsi qu'une capacité à bifurquer.

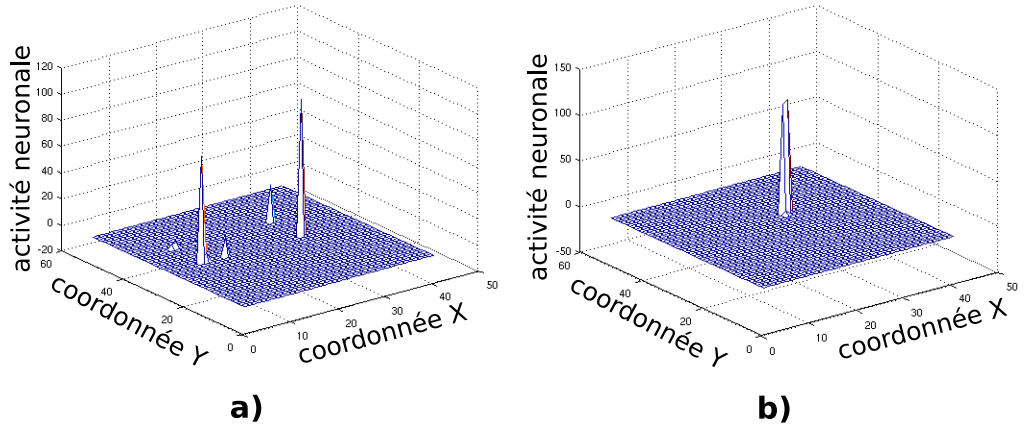


Fig. 7-9 – Influence du champ de neurones sur la stabilité du centre de la zone pré-attentive. a) montre le résultat du mécanisme pré-attentive après avoir construit une carte de saillance (couleur, contour, gabor) b) montre la stabilité obtenue de la zone de pré-attentive par l'intermédiaire du champ de neurones. L'activité montre l'histogramme de focalisation (centre de la zone pré-attentive) en fonction des coordonnées (x,y): plus l'activité est grande, plus cette coordonnée (x,y) est choisie.

Fig. 7-10 – Les cercles rouges montrent le centre de la zone pré-attentive sélectionnée. a) montre le traitement sans le champ de neurones où le premier point trouvé change d'une image à l'autre. b) montre l'influence du champ de neurones sur la stabilité de la zone pré-attentive. Le système se stabilise sur une seule zone pré-attentive dont le centre est représenté par le cercle rouge.

Les figures 7-9 et 7-10 montrent les propriétés intrinsèques du champ de neurones vis à vis de la stabilité du système visuel. La figure 7-10 montre la localisation dans l'espace (x,y) de la première saccade oculaire avec et sans le champ de neurones. Cette figure montre l'influence du champ de neurones sur la stabilité de la première saccade oculaire (sélection de la zone pré-attentive). Le test est réalisé sur environ 200 images dont la qualité est dépendante des pertubations causées par l'environnement et le capteur utilisé. On observe sur la figure 7-9a) que plusieurs zones sont privilégiées montrant l'instabilité du mécanisme pré-attentif. L'ajout du champ de neurones montre l'amélioration apportée sur la stabilité du premier point de focalisation (centre de la

zone pré-attentive). Le centre de la zone pré-attentive est toujours localisé au même endroit (pic d'activité de l'histogramme: figure 7-9b). Les propriétés du champ de neurones (hystérésis temporelle, bifurcation) permettent bien de stabiliser le mécanisme pré-attentif.

7.4.4 Apprentissage d'objets via l'interaction émotionnelle

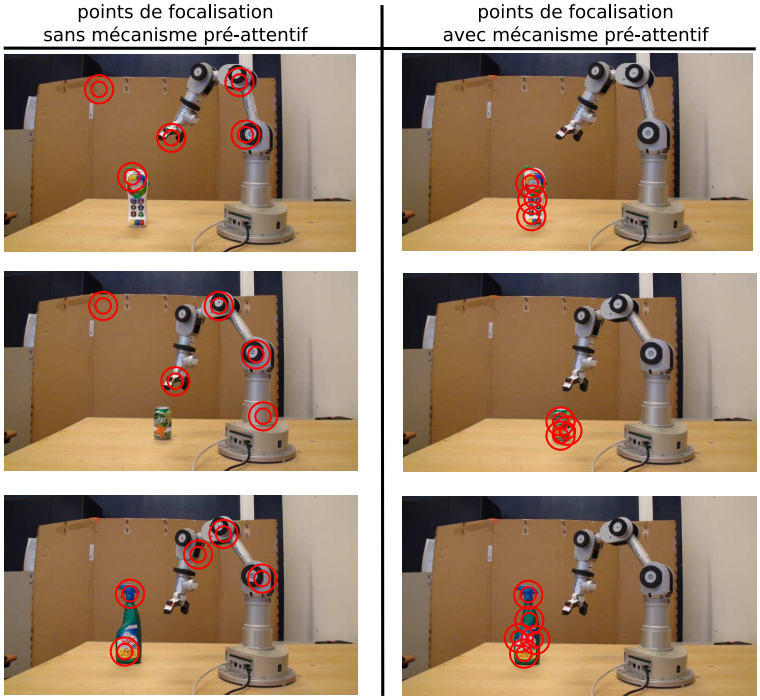


Fig. 7-11 – Processus visuel avec ou sans mécanisme de vision pré-attentive.

Une fois que l'attention visuelle du robot est focalisée sur une zone précise de l'image, le traitement visuel est le même que pour le chapitre 4. La figure 7-11 met en évidence l'importance du mécanisme de vision pré-attentive pour sélectionner ensuite les points de focalisation qui permettront d'associer à l'objet une valeur émotionnelle en fonction de l'expressivité du partenaire. Grâce à ce mécanisme, les points d'intêrets sont pris exclusivement dans la zone où se trouve l'objet. La figure 7-12 montre un exemple du traitement visuel sur un objet particulier.

Chaque vue locale est apprise par le groupe de neurones de VF de type SAW (Selectif Adaptatif Winner) qu'on a décrit en détails dans le chapitre 4. Les propriétés intéressantes de ce réseau sont une capacité d'apprentissage rapide (en un coup) et peut réaliser un moyennage des motifs.

OSP ("objet state prediction" correspond à la prédiction de l'état interne) associe l'activité de VF avec l'état interne du robot correspondant à l'expression faciale FE. FE est déclenchée par l'intermédiaire de l'expression faciale du partenaire humain. L'association est un simple mécanisme de conditionnement utilisant la règle d'apprentissage des moindres carrés dans laquelle est introduit un terme de modulation.

$$\Delta w_{ij} = \epsilon . VF_i . (FE_j - OSP_j) \sum_k FE_k \tag{7.1}$$

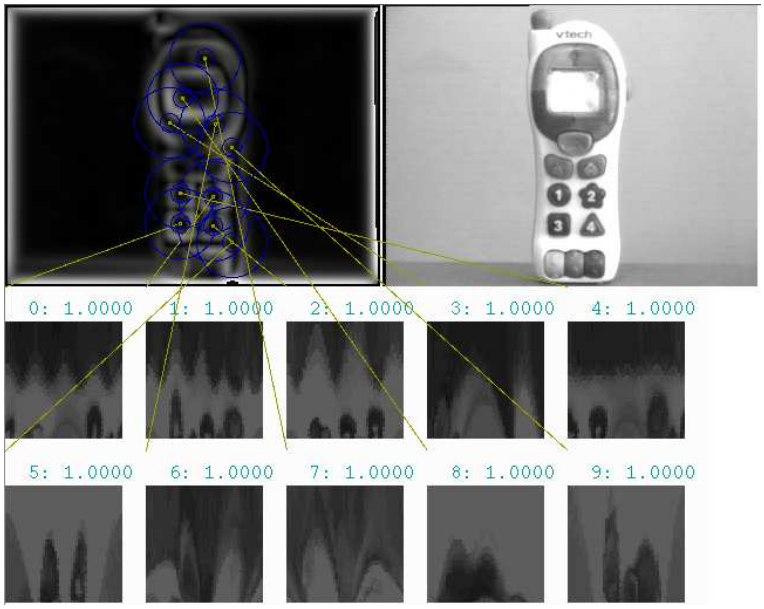


Fig. 7-12 – Traitement visuel. Le processus visuel utilisé est le même que pour la reconnaissance des expressions faciales. Les points de focalisations sont le résultat d'une convolution entre une Différence de Gaussiennes (DOG) et l'image, focalisant particulièrement sur les zones de courbures. Une transformation log-polar est utilisée pour obtenir plus de robustesse vis à vis de la vue locale.

w_{ij} correspond à un poid synaptique reliant l'entrée i et la sortie j.

Une mémoire à court terme (STM) est utilisée dans le but de sommer et de filtrer sur une courte période, la valeur émotionnelle OSP associée avec chaque vue locale explorée. OEV (objet emotional value) correspond ici à la valeur émotionnelle de l'objet, la plus grande activité OEV_i déclenche la valeur émotionnelle i^{th} $(0 < i \leq 5)$ comme une conséquence d'un Winner Take All.

Après l'apprentissage, les associations entre VF (reconnaissance des vues locales) et OSP (état émotionnelle) sont suffisamment élevées pour bypasser l'activité réflexes de bas niveau venant de l'état interne IS et FE. Dans ce cas, la valeur émotionnelle OEV est le résultat de l'intégration temporelle venant des états émotionnels associés aux différentes caractéristiques visuelles analysées par le système. Les caractéristiques visuelles ont une valeur émotionnelle si elles ont été associées lors de l'apprentissage avec l'expression faciale du robot.

7.4.5 Mise en pratique

Les résultats du processus globale (processus attentionnel et classification) sont montrés sur la figure 7-14. Nous avons placé sur la scène 10 objets les uns à la suite des autres (figure 7-13). Chaque objet est placé pendant 1 minute sur une table dans une zone accessible à notre robot robot (figure 7-11), permettant ainsi l'apprentissage de la valeur émotionnelle de l'objet via l'interaction émotionnelle. Les objets restent à des positions fixes durant la phase d'apprentissage. Une fois cette apprentissage terminé, les objets sont placés à différents endroits impliquant un changement de position des objets dans le repère image et une variation de distance vis à vis de la caméra (plus ou moins 50 cm). Ces résultats montrent que l'apprentissage est réalisable en un temps très court, 1 minute d'apprentissage par objet est suffisante (temps pendant lequel

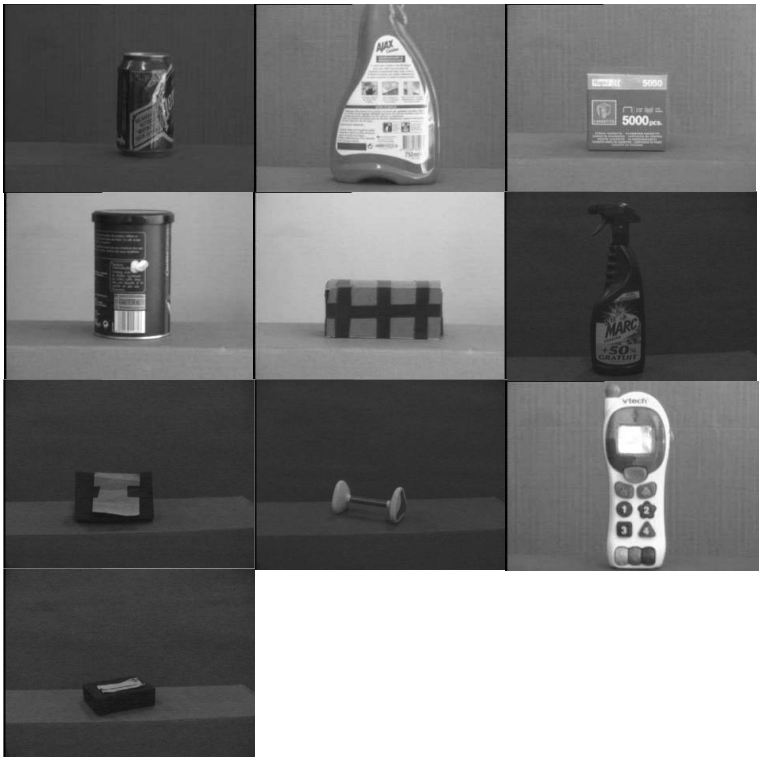


Fig. 7-13 – Base de donnée relative aux objets utilisée pendant l'expérience de référencement social.

		Réponse du robot	
		Positive	*Négative*
valeur émotionnelle donnée à l'objet	*Positive*	**87%**	13%
	Négative	2%	**98%**

Fig. 7-14 – Taux de reconnaissance et de confusion pour les différents objets. Nous avions à notre disposition 10 objets placés autour du robot (5 objets négatifs et 5 objets positifs). Chaque objet a été appris pendant 1 minute via l'interaction émotionnelle avec la tête expressive (l'objet resté à une position fixe). L'expérimentateur transmet un signal émotionnel positif lorsque l'objet peut procurer du plaisir au robot et au contraire un signal négatif si l'objet est dangereux pour le robot. Pour mesurer, le taux de reconnaissance, les objets sont ensuite replacés sur la scène à différents endroits (distances et positions différentes).

l'expérimentateur exprime une expression faciale). La reconnaissance de la valeur émotionnelle est de 87% pour les objets positifs et 98% pour les objets négatifs montrant ainsi que la classification est très bonne malgré l'introduction de changements importants (position, changement d'échelle et légère rotation) de l'objet dans la prise de vue.

A ce stade de développement, le robot est capable d'utiliser l'expression faciale du partenaire comme un signal de communication pour attribuer une valeur émotionnelle à des objets. Grâce à l'interaction avec un partenaire humain, le robot est capable de reconnaitre et d'utiliser les expressions de l'autre dans le but de désambiguïser certaines situations comme la présence d'un nouvel objet dans l'espace de travail. Sans connaissance explicite de ce qu'est l'autre et ce que veut l'autre, autrement dit sans théorie de l'esprit, le robot est capable de produire des comportements sociaux relativement intéressants.

7.5 Apprentissage visuo-moteur et contrôle de bras à plusieurs degrés de liberté

7.5.1 Introduction

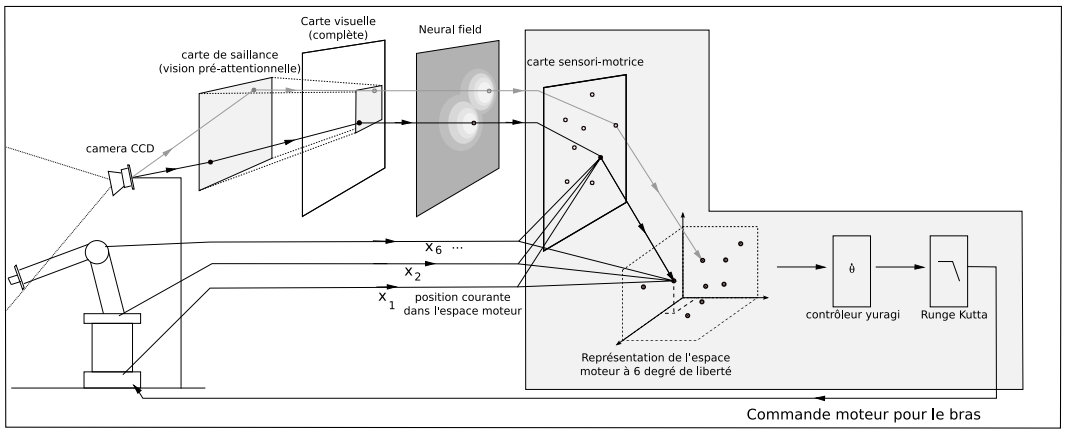


Fig. 7-15 – Modèle montrant l'architecture capable de contrôler le bras robotique à 6 degrés de liberté. La carte sensori-motrice peut apprendre à associer un stimulus visuel et des informations proprioceptives du bras de robot impliquant après apprentissage l'activation d'un ou plusieurs attracteurs dans l'espace moteur. Si la position courante du bras est différente des attracteurs sélectionnés alors une loi de commande est envoyée au bras robotisé pour atteindre la position cible. Cette loi de commande est régie par une équation dynamique (équation de langevin): le contrôleur Yuragi.

Pour l'apprentissage visuo-moteur correspondant à l'apprentissage entre l'extrémité du bras (la pince) et la proprioception de ce dernier, nous avons repris les travaux de (Andry et al., 2001). Le bras est alors capable de rejoindre plusieurs positions dans son espace de travail. Une position visuelle correspond à une ou plusieurs configurations motrices correspondant à des attracteurs dans l'espace moteur (Figure 7-15) lorsqu'une cible est présente dans le champ visuel du robot. Ces attracteurs entrainent le bras dans un bassin d'attraction centré sur la position visuelle de la cible. Le contrôle est réalisé par un système dynamique lissant la trajectoire (Fukuyori et al., 2008). Ce système dynamique utilise un signal de renforcement dans le but d'assigner plus ou moins d'importance aux attracteurs qui ont été sélectionnés. Une récompense positive est attribuée si le bras se dirige dans la direction souhaitée et inversement une récompense négative est donnée si le bras s'en éloigne. Cette récompense peut être de modéliser par une fonction d'énergie à minimiser, une fonction modélisant le comportement du robot ou une fonction dépendant de l'état émotionnel.

7.5.2 Contrôle moteur: Modèle Yuragi

Les travaux de (Andry et al., 2001) utilisent un champ de neurones pour contrôler le bras robotisé. Cependant, ce contrôle est réalisé dans l'espace visuel (x, y) avec un bras ayant 3 degrés de liberté $(\theta_1, \theta_2, \theta_3)$. C'est l'alignement $(x- > \theta_1)$ et $(y- > \theta_2, \theta_3)$ qui facilitent les choses. Le principal problème est le temps d'apprentissage très long en haute résolution de la carte visuo-motrice. Si cette carte n'est pas totalement construite ("trou" dans la carte) alors il y aura un manque de précision. Le passage à un bras à 6 degrés de liberté avec ce genre de modèle semble relativement difficile.

Par conséquent, nous avons choisi de développer un modèle dont le contrôle est réalisé entièrement dans l'espace moteur. Le modèle que nous allons décrire (Fukuyori et al., 2008), est régi par une équation de Langevin dans l'espace proprioceptif du bras:

$$\tau_x \dot{x} = A.f(x) + \epsilon \tag{7.2}$$

$$A = \begin{cases} d & \text{si } \frac{\Delta d}{\Delta t} < 0 \\ 0 & \text{sinon} \end{cases} \tag{7.3}$$

x est l'état (proprioception du bras) et $f(x)$ représente la dynamique du modèle de sélection des attracteurs, $\tau_x = 0.1$ est une constante de temps, ϵ représente le bruit, d correspond à la distance entre la cible et l'extrémité du bras et $\frac{\Delta d}{\Delta t}$ la dérivée de la distance à la cible. A correspond au signal de renforcement qui indique l'adéquation de l'état x vis à vis de l'environnement et contrôle le comportement du modèle de sélection des attracteurs. $A.f(x)$ devient dominant lorsque l'activité est large signifiant que $f(x)$ et A ont une certaine activité permettant au bras de se diriger vers la cible: l'état x se rapproche du but à atteindre de manière déterministe. D'autre part, le bruit ϵ devient dominant quand l'activité $A.f(x)$ est petite et par conséquent l'état x devient plus probabiliste.

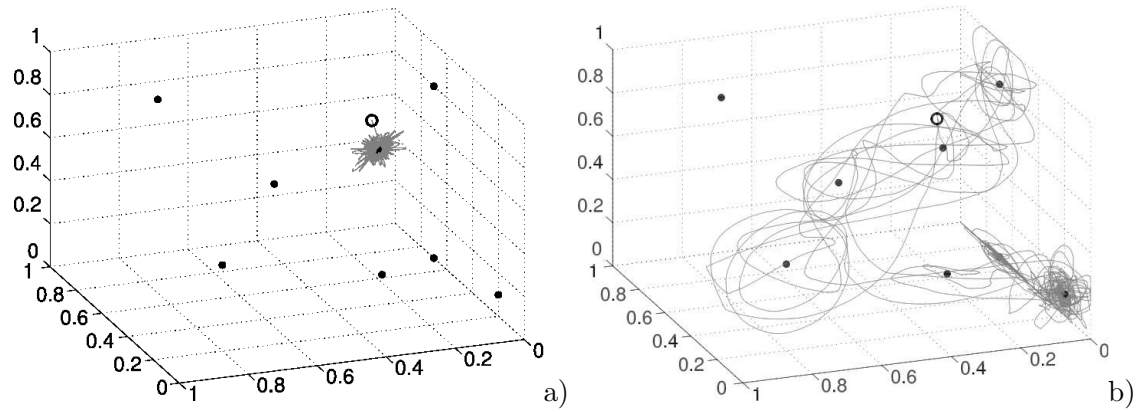


Fig. 7-16 – Illustration de l'équation de langevin pour une simulation d'un bras à trois degrés de liberté. "o" représente la position initiale du bras. Ce test est réalisé avec 8 attracteurs. a) montre une trajectoire convergeante vers un attracteur proche dont les paramètres suivant sont fixés: $\beta = 0.01$, $\epsilon = 0$, $A = 1$ et $\tau = 0.01$. b) montre la trajectoire dépendante de A et ϵ. Elle nous montre l'effet du bruit sur la commande en vitesse permettant une exploration de l'espace moteur.

$$f(x) = \sum_{i=1}^{n_a} N_i \frac{(X_i - x)}{||X_i - x||} \tag{7.4}$$

$$N_i = \frac{g_i(x)}{\sum_{j=1}^{n_a} g_j(x)} \tag{7.5}$$

$$g_i(x) = exp\{-\beta||X_i - x||^2\} \tag{7.6}$$

n_a est le nombre d'attracteurs sélectionnés. X_i (i=1, ... , n_a) est un vecteur représentant le centre du i-ème attracteur. La fonction N_i est une gaussienne normalisée dont le comportement est fonction de la position du bras. N_i a une activité élevée dans le cas où la position du bras est proche de l'attracteur i et inversement une activité faible lorsque ce dernier est loin de l'attracteur i. Le comportement du système est tel que le bras se rapproche des attracteurs les plus proches.
La figure 7-16 montre l'influence du contrôleur vis à vis des deux paramètres essentiels ϵ et A. La figure 7-16 a) montre la convergence du contrôleur lorsque le bruit n'a pas d'influence sur le comportement du bras. Au contraire, la figure 7-16 b) montre le fonctionnement du modèle si le bruit ϵ est actif. On peut voir d'après cette figure que des comportements de stabilité et de bifurcation sont émergents à ce système dynamique. Le bras est capable de converger si le bruit est moins présent que le renforcement A et au contraire une exploration est réalisée lorsque le bruit est plus fort que le renforcement. Cependant, à partir du moment où ϵ et A fluctuent alors il est possible de construire des "attracteurs virtuels" correspondant simplement à la capacité de se stabiliser entre différents attracteurs.

7.5.3 Résultat

Les résultats montrent la capacité du robot à atteindre des cibles dans l'espace péri-corporel du bras robotique. La figure 7-17 montre que le bras robotique est capable de se diriger vers un stimulus visuel. Il rejoint la cible en partant de positions d'origine différentes, toutes les trajectoires convergent vers la cible en ayant appris un nombre réduit d'états attracteurs (couples proprioception-vision). La figure 7-17a) montre la distance à la cible qui décroit au fur et à mesure que l'on se rapproche de l'objet ce qui montre une balistique curviligne du bras.

De plus, d'autres résultats montrent également la convergence du modèle même dans un cas où l'attracteur n'est pas existant (Figure 7-18). Le modèle crée un "attracteur virtuel" permettant de rejoindre des positions n'ayant encore jamais été apprises. Ce modèle possède une capacité de généralisation car il peut rejoindre des positions non explorés. Le temps d'apprentissage se trouve donc diminué, le robot n'aura pas besoin d'apprendre l'ensemble de son domaine péri-corporelle.

Les figures 7-17c), 7-18c) met en évidence une des propriétés émergentes du modèle à savoir l'obtention d'une balistique curviligne du bras robotique (bout de la pince). D'après ces figures, la précision pour atteindre un objet reste très bonne, le bras se stabilise à proximité de la cible à une distance inférieure à 5 pixels. Une telle précision est largement suffisante pour la prise d'un objet étant donnée la taille que peut prendre ce dernier dans l'image (supérieure à 5 pixels).

7.5.4 Conclusion

Tous ces résultats expérimentaux nous montrent que le contrôle du bras à 6 degrés de liberté peut être effectué grâce à un couplage entre une carte sensori-motrice (création d'attracteurs) et

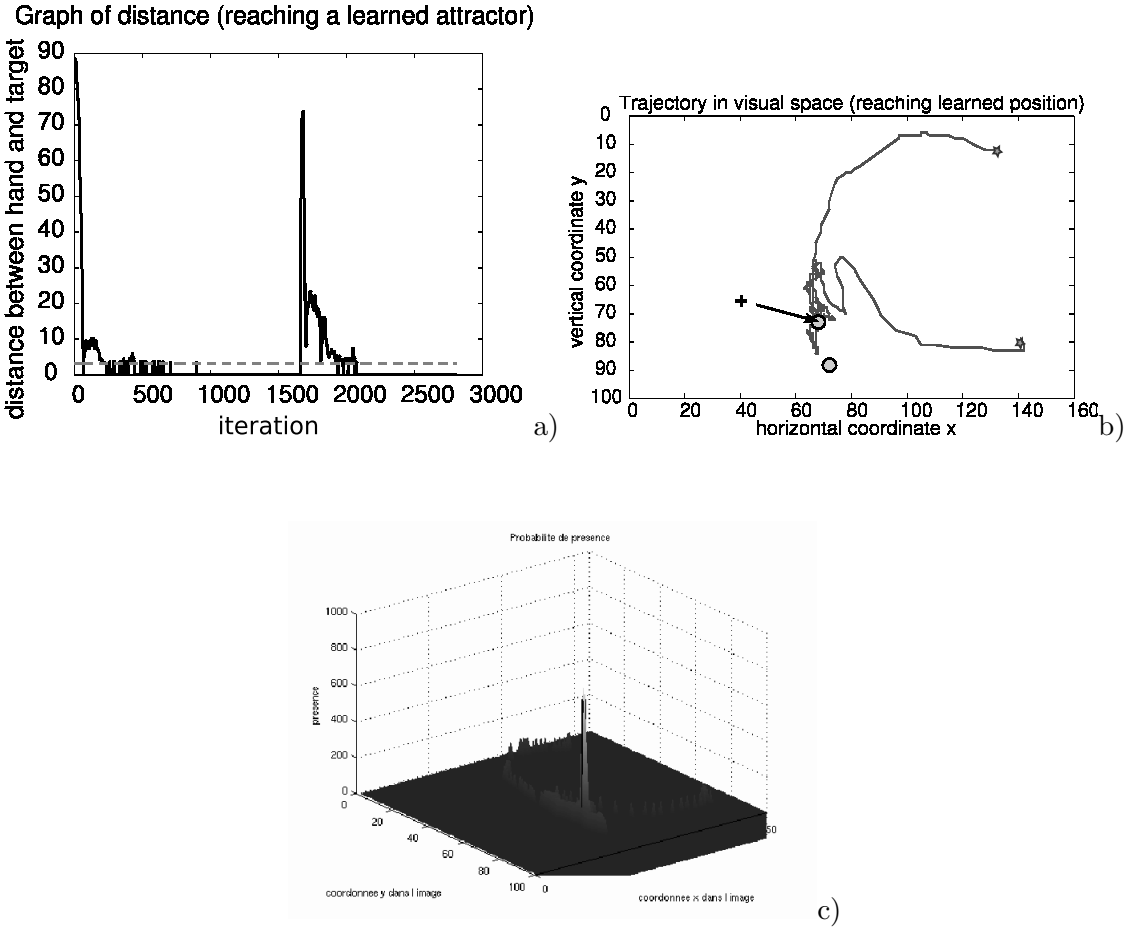


Fig. 7-17 – Convergence du bras robotique vers la cible. a) montre la distance à la cible et à 1500 itérations la cible a été déplacée. b) montre la trajectoire du bras robotique lorsque la cible à rejoindre est proche d'un attracteur. Les cercles gris correspondent aux différents attracteurs appris, la croix correspond à la cible à rejoindre et les étoiles correspondent aux différents points de départ. Deux attracteurs sont activés. c) montre l'histogramme de la trajectoire (endroit où se situe le bout de la pince). Le pic noir est la cible à rejoindre et les pics blancs correspondent à la présence du bout de la pince.

un système dynamique utilisant les attracteurs appris pour réaliser le contrôle moteur (Figure 7-15). Cette architecture possède plusieurs propriétés:

- La solution trouvée est stable: le bras converge vers la solution avec une bonne précision.
- Le bras peut fusionner des attracteurs pour créer une solution ("création d'attracteurs virtuels").
- La balistique du bras est souvent curviligne montrant une continuité en direction du stimulus visuelle à rejoindre.

7.6 L'interaction émotionnelle influence le comportement du bras robotique

Une fois les expressions faciales apprises, la reconnaissance de l'expression de l'expérimentateur peut influencer la prise d'objet par l'intermédiaire de la valeur émotionnelle assignée aux ob-

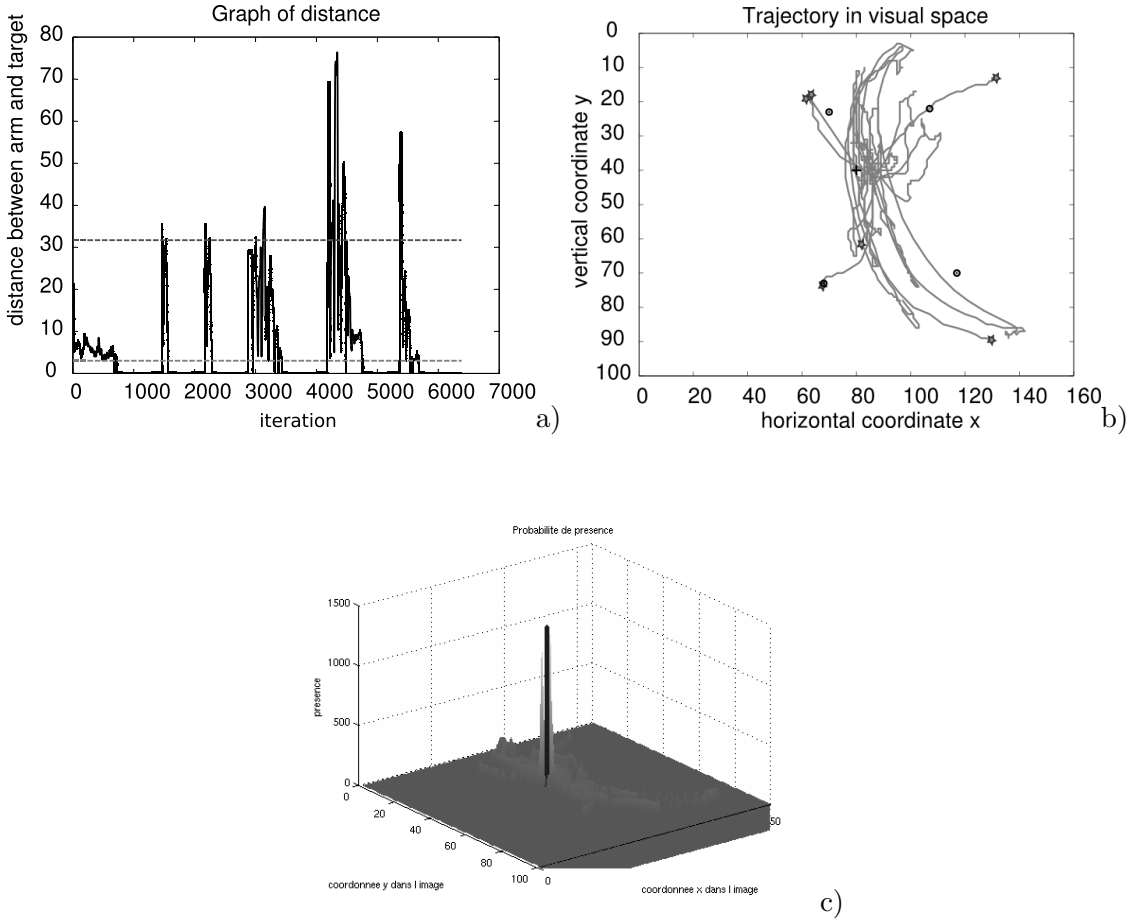


Fig. 7-18 – Trajectoire montrant la convergence du bras robotique. a) montre la distance à la cible. b) montre la trajectoire du bras robotique lorsque la cible à rejoindre est entre plusieurs attracteurs. Les cercles noirs correspondent aux différents attracteurs appris. La croix correspond à la cible à rejoindre et les étoiles correspondent aux différents points de départ. Quatre attracteurs sont activés. c) montre l'histogramme de la trajectoire (endroit où se situe le bout de la pince). Le pic noir est la cible à rejoindre et les pics blancs correspondent à la présence du bout de la pince.

jets. Par conséquent, les objets ainsi que l'expression faciale de l'humain peuvent fournir un signal de renforcement permettant l'adaptation du comportement du robot. Dans notre cadre expérimental, nous souhaitons par l'intermédiare de l'interaction émotionnelle réguler le comportement du bras robotique. A est maintenant une fonction dépendante de l'état émotionnel du robot. En d'autres termes, le bras sera capable de rejoindre ou d'éviter un objet en fonction de la valeur émotionnelle A:

$$A = \begin{cases} 1 & \text{si valeur émotionnelle positive} \\ -1 & \text{sinon} \end{cases} \tag{7.7}$$

Durant, l'expérience un seul objet est placé sur la scène. Néanmoins, un objet procurant du "plaisir" (objet positif) et/ou un sourire de l'expérimentateur incite le robot à attraper l'objet. L'évitement d'objet est la conséquence d'un objet dangereux et/ou d'une expression négative de l'expérimentateur.

La figure 7-19 montre les étapes importantes du modèle du référencement social. La Figure 7-

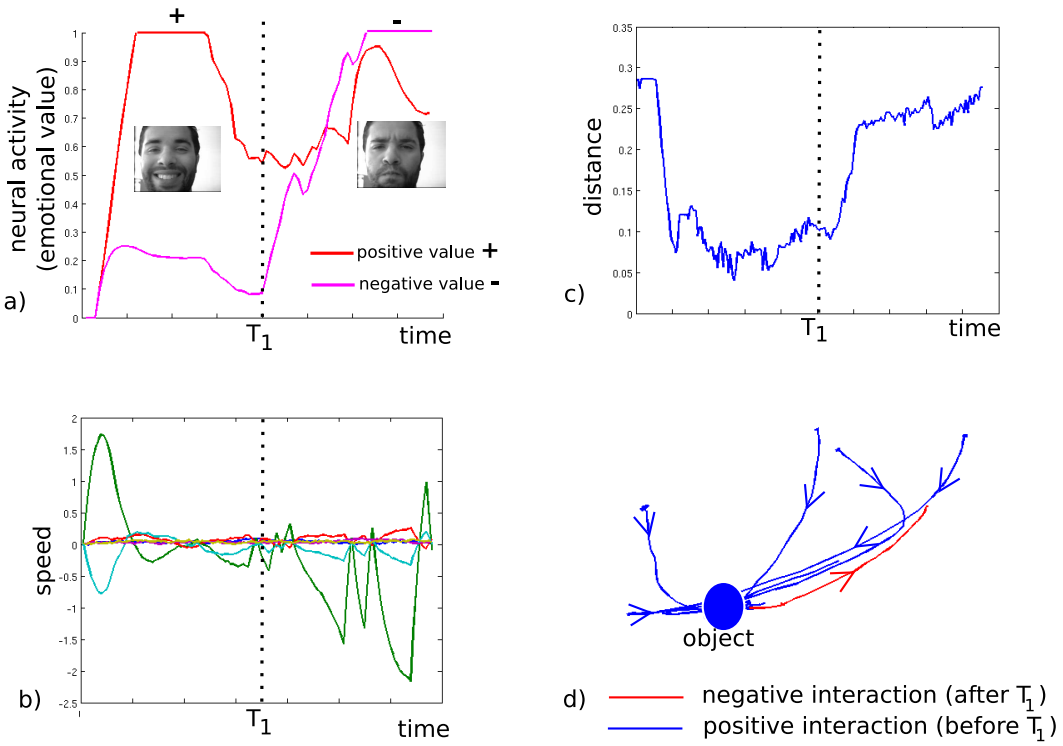


Fig. 7-19 – Ces courbes montrent: a) la valeur émotionnelle transmise à l'objet grâce à l'interaction avec le partenaire humain (Avant T_1). L'humain transmet une valeur positive. Après T_1, l'humain transmet une valeur négative. b)les vitesses de chaque moteurs du bras (6 degrés de liberté) c) La distance à l'objet d) les trajectoires du bras robotiques en partant de différents points de départ: le bras est capable d'attraper l'objet associé à une expression faciale positive et d'éviter l'objet associé à une expression négative.

19a montre la valeur émotionnelle de l'objet qui est associée avec l'expression du partenaire. Avant T_1, l'humain exprime de la joie en présence du nouvel objet, ce qui signifie pour le robot qu'il peut se rapprocher de cet objet, le robot associant l'objet à un signal positif (A_1 est activé). Nous pouvons voir (figure 7-19b, 7-19c) que plus la distance entre le pince et l'objet diminue plus la vitesse du bras diminue jusqu'à tendre vers 0 lorsque l'objet est atteint. Après T_1, le partenaire humain transmet une valeur négative, la valeur de l'objet est modifiée (A_2 est activé). Maintenant, le bras change son comportement et évite l'objet (l'objet apparaît comme "dangereux" pour le robot).

De plus, il nous semble important d'obtenir des résultats quantitatifs sur la capacité du bras à attraper les objets ayant une valeur émotionnelle positive. Un seul objet est placé dans l'espace péri-corporel du robot. Néanmoins, il peut être positionné à différents endroits (figure 7-20) montrant ainsi la capacité à attraper un objet indépendamment de sa position. La figure 7-21 montre que le bras est capable d'attraper les objets positif dans 82% des cas et de se diriger vers ces objets dans 92% des cas. L'analyse des résultats montre que le bras échoue seulement lorsque l'objet ne peut pas être atteint par le bras (zone de l'environnement où il n'existe pas d'attracteurs). Dans le cas où l'objet est négatif, le bras s'éloigne des différents objets 100% du temps. Ces résultats mettent en évidence la capacité du dispositif robotique à adapater son comportement en fonction de la valeur émotionnelle de l'objet transmise par l'expérimentateur

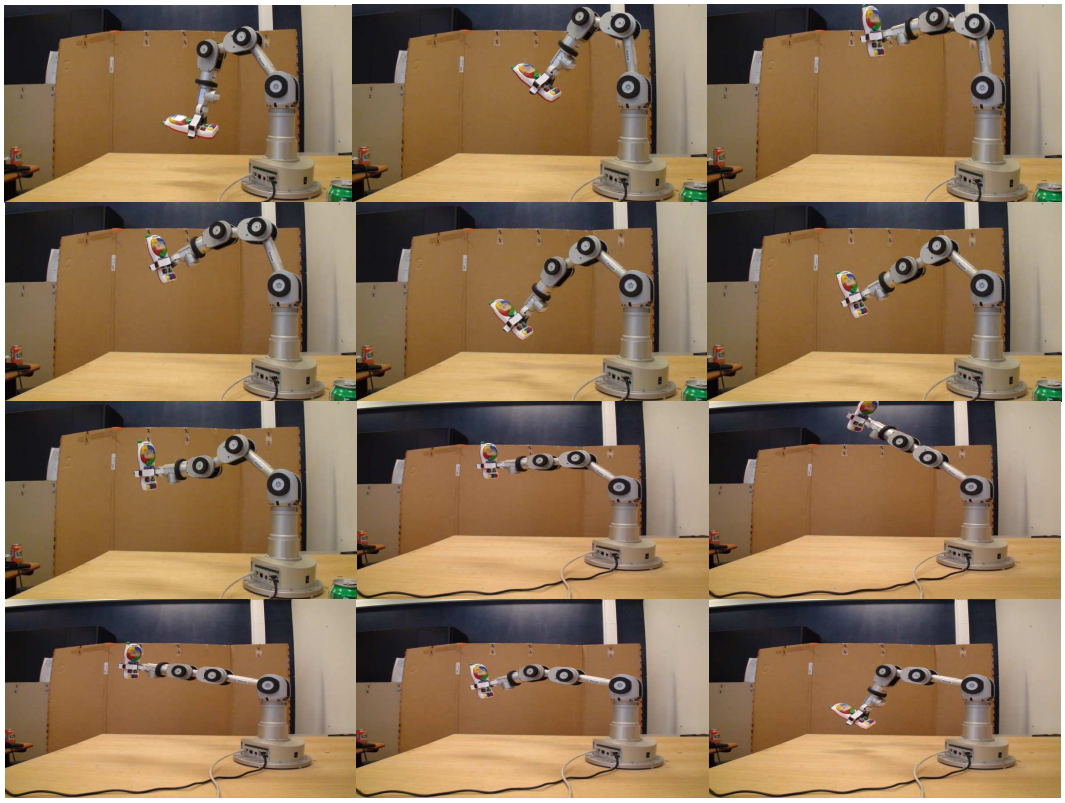


Fig. 7-20 – Un ensemble de positions que le bras robotique est capable de rejoindre.

humain.

	attrape	se dirige vers
objet 1	90%	90%
objet 2	80%	90%
objet 3	80%	100%
objet 4	80%	90%
objet 5	80%	90%
moyenne	82%	92%

Fig. 7-21 – Pourcentage de réussite lorsque le bras essaye d'attraper un objet dans son environnement. Les 5 objets positifs sont placés sur la scène les un après les autres. Chaque objet est placé à 10 endroits différents permettant ainsi d'obtenir des résultats quantitatifs sur la capacité du bras à attraper un objet positif.

A ce stade de développement, l'adaptation du robot à l'environnement est fonction de l'expressivité émotionnelle de l'expérimentateur. Selon la valeur émotionnelle de l'objet et de l'expression faciale de l'humain, le robot possède la capacité d'attraper ou d'éviter un objet. L'expression émotionnelle est vue comme un excellent moyen de communication permettant d'associer à un objet une valeur émotionnelle afin de réguler le comportement du robot.

7.7 Conclusion

A notre connaissance, ce système est le premier qui apprend de manière autonome un couplage entre l'émotion (reconnaissance d'expression faciale) et des capacités sensori-motrices. Nous pouvons ici entrevoir une mini séquence développementale. Sans modèle de théorie de l'esprit, le robot est capable d'apprendre des tâches de niveau toujours un peu plus complexe comme par exemple l'apprentissage d'expressions faciales, la détection de visage et l'attribution de valeurs émotionnelles à des objets. Une simple cascade d'architectures sensori-motrices qui au départ ne sont pas dédiées aux interactions sociales, possèdent des propriétés permettant de résoudre ces capacités cognitives. Pour améliorer le fonctionnement du système, il pourrait être nécessaire de moduler l'état émotionnel comme une fonction d'intensité de l'expression faciale et moduler le comportement vis à vis de l'objet selon l'intensité par exemple une intense expression de colère pourrait impliquer un vif retrait et une expression intense de joie pourrait impliquer une prise d'objet rapide.

D'autre part, ce modèle ne décrit pas tous les aspects du référencement social. Ici, la question de l'attention conjointe qui est une capacité importante du référencement sociale n'a pas été traitée. Nous essayerons dans le chapitre 9 de donner un début de réponse à ce problème. D'autres auteurs comme (Nagai et al., 2003) se sont intéressés à cette question et proposent un modèle très complémentaire du notre. D'autres auteurs (Breazeal et al., 2005; Thomaz et al., 2005; Jasso et al., 2008) se sont intéressés à des aspects de référencement social dans des expériences robotiques. Cependant, ces modèles possèdent de nombreux a priori ne permettant pas d'obtenir un modèle générique. Enfin, les robots utilisés dans ces expériences possèdent de nombreux degrés de liberté mais les comportements adaptatifs qu'ils proposent sont minimaux.

Nous pensons que notre approche peut fournir de nouvelles données intéressantes sur comment les humains peuvent développer les capacités de référencement social provenant des dynamiques sensori-motrices. A l'inverse des théories développementales courantes, qui avancent l'idée que le référencement social est un processus cognitif complexe d'une relation triadique, le travail développé dans ce chapitre montre 1) la primacie de l'émotion dans l'apprentissage, 2) des mécanismes de conditionnement simple par lesquels le signal émotionnel d'une autre personne peut être associé avec les états émotionnels internes du robot et 3) un simple jumelage d'états émotionnels avec les objets est capable de diriger le comportement du robot.

Ce travail met en évidence que la cognition sociale incluant le référencement social, pourrait avoir une dimension émotionnelle plus forte et un besoin moindre pour une cognition de haut niveau comme pouvait le penser certains auteurs(e.g. (Bard et al., 2005)). De nouvelles études de neuropsychologie du système miroir dans l'émotion (J. Bastiaansen, 2009), les bases neuronales de l'intersubjectivité (e.g. (Gallese, 2003)) soulignent l'importance du rôle jouer par l'émotion dans l'émergence du développement du référencement social.

Ce travail suggère que le système robot/partenaire est un système autopoiétique (Mataruna and Varela, 1980) dans lequel le signal émotionnel et l'empathie sont des éléments importants du réseau pour maintenir l'interaction et permettre des capacités d'apprentissage de plus en plus complexes. L'expression émotionnelle est un excellent moyen pour communiquer des signaux non ambigües dans certaines situations qui elles peuvent être ambigües car non connues ou incohérentes par rapport à l'expérience passée du robot. L'ajout du control du bras robotique par un signal issu de la reconnaissance de l'expression faciale d'un humain permet au robot d'apprendre à manipuler un objet sous le regard actif de l'expérimentateur.

Chapitre 8

Référencement social dans un contexte de navigation autonome

8.1 Introduction

Des tâches de navigation sont souvent apprises à un robot de façon supervisée par l'intermédiaire d'un expérimentateur (Hersch et al., 2008). Ces techniques ont l'avantage d'être rapides en terme de temps d'apprentissage mais en contre partie l'expérimentateur doit être un expert du dispositif robotique utilisé, il doit connaître exactement comment le robot fonctionne. En d'autres mots, l'expérimentateur doit fortement s'adapter à l'architecture sous-jacente afin de superviser efficacement le robot pour atteindre des performances d'apprentissage satisfaisantes.

L'autonomie du robot mobile peut être plus facilement atteinte si le robot possède des capacités d'apprentissage en ligne et au travers d'interactions émotionnelles. Le référencement social correspond bien à ce concept car il définit la capacité à reconnaître, comprendre, répondre et altérer son comportement en réponse à des expressions émotionnelles venant d'un partenaire social (Klinnert et al., 1983a; Walden and Ogan, 1988; Russell et al., 1997). Par ailleurs étant dans un contexte non verbal et n'ayant pas besoin de capacités cognitives de haut niveau comme avec le modèle développé dans la chapitre 7, l'ensemble des informations venant des interactions émotionnelles semblent être suffisantes, rapides et efficaces pour déclencher l'apprentissage d'étapes précoces du développement cognitif humain (comparativement à un apprentissage seul sans interaction).

Le but de ce chapitre est de montrer comment l'apprentissage des robots mobiles pourrait être facilité et rendu plus accessible pour des humains non-experts. Nous faisons l'hypothèse que les interactions émotionnelles peuvent être une aide pour des interactions intuitives. Plus précisément, les capacités intrinsèques du référencement social sont idéales pour apprendre aux robots des tâches complexes. Les humains ne doivent pas être des experts du système pour interagir avec lui, contrairement à des techniques d'apprentissage où l'humain doit d'abord apprendre à utiliser la machine. Grâce à l'interaction émotionnelle, l'humain peut envoyer un signal à distance qui soit naturel pour lui (une expression faciale) et qui sans interaction invasive va modifier le comportement du robot au moment où le robot en à besoin. Pour tester cette idée, nous avons couplé deux systèmes: le premier est constitué de notre tête robotique capable d'apprendre et d'imiter les expressions faciales et le deuxième est une plateforme robotique mobile capable d'apprendre des tâches de navigation visuelle de manière autonome dans des environnements réels et complexes. Nous disposons ainsi d'un système robotique complexe capable à la fois d'agir sur son environnement physique (navigation) et d'interagir avec un partenaire humain.

Deux solutions possibles pour le couplage de ces deux systèmes ont été testées. Dans la première solution testée, l'émotion est utilisée pour qualifier le comportement du robot. Le robot montre ces capacités à apprendre comment atteindre un but dans l'environnement en utilisant des signaux venant de l'interaction émotionnelle (individu interagissant avec lui). Ces signaux donnent au robot une information sur la qualité de son comportement et lui permettent d'apprendre des associations lieux-actions dont l'objectif est de construire un bassin d'attraction autour des lieux buts. Dans la deuxième solution testée, l'émotion est utilisée pour qualifier l'environnement immédiat du robot. Le robot montre sa capacité à apprendre comment éviter un lieu de son environnement en l'associant avec l'expression faciale de colère de l'expérimentateur.

La première stratégie permet à l'expérimentateur d'enseigner au robot comment atteindre un lieu spécifique de n'importe quel endroit de son environnement. Cependant, cette stratégie est plus coûteuse en temps d'apprentissage que la seconde stratégie qui semble être inappropriée pour apprendre à atteindre des buts mais qui est très efficace pour les éviter.

Fig. 8-1 – Set-up expérimental: la tête de robot apprend à reconnaître les expressions faciales d'un partenaire interagissant avec elle, alors que la base mobile permet au robot d'explorer son environnement, d'apprendre des lieux et des conditionnements lieux-actions. Les primitives visuelles utilisées pour la localisation du robot sont les mêmes que celles utilisées pour la reconnaissance des expressions faciales. La taille de la pièce est de 7m x 7m.

L'ensemble du système peut fournir au dispositif robotique des informations concernant son environnement et/ou son comportement (par exemple, le renforcement d'une bonne action). Dans ce cas, les associations sensori-motrices contrôlant l'apprentissage du robot sont définies comme la conséquence de l'interaction émotionnelle avec le partenaire social. Cet apprentissage interactif s'appuie sur la capacité de l'expérimentateur à réagir émotionnellement au comportement du robot. Les expressions émotionnelles seront le média de communication entre l'homme et le robot. L'homme et le robot devront s'adapter l'un à l'autre au travers de leurs interactions

faciales d'une manière que nous espérons se rapprocher des interactions parent/bébé.

Dans ce chapitre, nous verrons tout d'abord comment un robot est capable d'apprendre à naviguer de manière autonome dans son environnement en utilisant des informations visuelles et odométriques (Figure 8-1). Puis, nous étudierons le couplage du système de navigation avec une tête de robot expressive fournissant au robot la capacité à apprendre des tâches de navigation via des interactions émotionnelles. La reconnaissance des expressions faciales agit sur le comportement du robot pour fournir des informations sur son environnement et/ou son comportement.

En conclusion, nous discuterons du couplage des deux types d'apprentissage. Nos résultats mettent en évidence l'importance de l'interaction émotionnelle afin d'adapter le comportement du robot mobile vis à vis de son environnement physique. De plus, notre approche peut fournir des données intéressantes sur le développement du nourrisson et plus précisément sur certains aspects du référencement social.

8.2 Le système de navigation

Le système de navigation permet à un robot mobile d'apprendre à naviguer dans des environnements naturels. Il a été développé au laboratoire ETIS au cours des années passées. Nous allons dans ce chapitre réinvestir cette architecture neuronale qui a l'avantage d'être autonome et peu coûteuse en temps de calcul.

Les capacités de navigation du robot mobile sont basées sur une architecture associative sensori-motrice bio-inspirée: l'architecture PerAc (Gaussier and Zrehen, 1995). Cette architecture permet au robot mobile d'apprendre le conditionnement d'une action par l'intermédiaire d'une entrée sensorielle. Plus précisément, le robot utilise un modèle de cellules de lieux inspiré de la neurobiologie (O'Keefe, 1978). Ceci consiste en un réseau de neurones capable d'apprendre à caractériser et donc à reconnaître différents lieux de son environnement. Ces neurones codent l'information sur la localisation des caractéristiques visuelles d'un lieu spécifique de l'environnement (Gaussier and Zrehen, 1995; Gaussier et al., 2000). L'activité des différentes cellules de lieux dépend à la fois de la reconnaissance des caractéristiques locales (aussi appelées landmarks ou amers) et de leur orientation absolue (azimuth). Une cellule de lieu sera d'autant plus active que le robot se rapproche du lieu appris. La zone spatiale qui active une cellule de lieux donnée est appelée champ de lieux. Un simple conditionnement via un réseau de neurones active l'apprentissage d'associations entre un champ de lieux et une action (i.e la direction du robot). Une fois l'apprentissage réalisé, la reconnaissance visuelle d'un champ de lieux déclenchera l'action associée.

Nous expliquerons dans cette section l'architecture neuronale permettant une navigation visuelle et préciserons les différents modules.

8.2.1 Les cellules de lieu visuelles

La stratégie de navigation sensori-motrice est basée sur la capacité à apprendre des cellules de lieux visuelles. Il s'agit d'un modèle biologiquement plausible de cellules de lieux préhippocampiques. La figure 8-2 résume la chaîne de traitements, depuis le flux de la caméra (image brute) jusqu'à la caractérisation des cellules de lieu. La caméra est montée sur un système pan (contrôlé par un servomoteur) permettant au robot de voir dans toutes les directions. En pratique, le robot perçoit son environnement visuel en balayant son panorama horizontalement. La caméra a un angle de vue de 40°, mais afin d'éviter de perdre des informations, elle capture une image tous les 24°(15 positions de caméra pour un panorama complet).

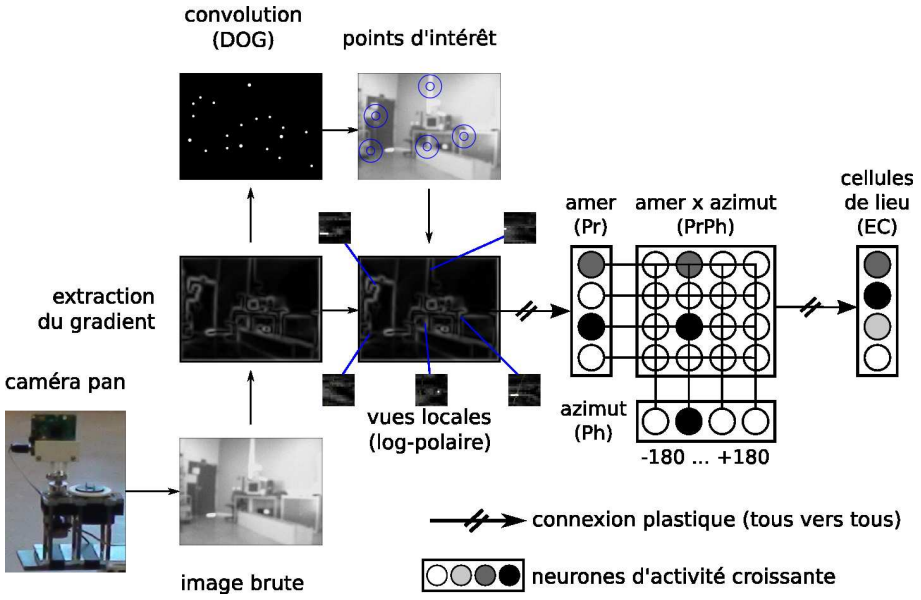


Fig. 8-2 – Architecture permettant l'apprentissage de cellules de lieux. Le système visuel extrait des vues locales en coordonnées log-polaires centrées sur les points à forte courbure de l'image de gradient. Une étape de fusion (PrPh) de l'identité des vues locales (Pr) et de leur azimut (Ph) permet de combiner les informations "où" et "quoi". Une catégorisation de la matrice PrPh permet de caractériser les cellules de lieux (EC). Todo : mettre les images pour les traitements visuels gradient à log-polaire.

Afin de minimiser l'impact des changements d'éclairage, notre système n'utilise pas l'image brute mais l'image du gradient extraite à partir des images capturées. Les traitements visuels qui suivent sont similaires à ceux utilisés pour la reconnaissance des expressions faciales. La figure 8-2 montre les points de focalisation et la zone utile des vues locales sur l'image brute et sur l'image de gradient. Ce processus visuel permet de caractériser simplement et avec une complexité calculatoire faible la scène visuelle. Une transformation logarithme et polaire permet une représentation des amers robustes à des petites rotations et à un certain changement d'échelle. cette transformation mime la projection de l'image rétinienne sur V1 (Schwartz, 1977; Schwartz, 1980).

Ce modèle du système visuel introduit deux flux d'informations distincts. L'information "quoi" correspondant à la reconnaissance d'une vue locale ou amer visuel en coordonnées log-polaires (32 x 32 points dans notre expérience) codé sur un vecteur de neurones de taille L. Il s'agit de la voie Pr (pour cortex périrhinal). L'information "où" correspondant à l'azimut absolu du point de focalisation, codé sur un vecteur de neurones de taille Θ. Il s'agit de la voie Ph (pour cortex parahippocampique). Dans notre cas, cette orientation absolue est obtenu grâce à une boussole électronique mais tout moyen est utilisable (boussole visuelle, proprioceptive, ...).

Ces deux flux d'information sont ensuite fusionnés dans un espace produit, au moyen d'une matrice de neurones appelée $PrPh$. Celle-ci comprend L x Θ unités sigma-pi (Rumelhart and Zipser, 1985) où tous les neurones d'une même ligne sont liés à un des L neurones de reconnaissance des amers (Pr) et tous les neurones d'une même colonne sont liés à un des Θ neurones azimut de Ph, d'où son nom matrice $PrPh$. On notera que l'activité de chaque neurone de PrPh est le produit $S_{ij}^{PrPh} = S^{L_i}$ x S^{Θ_j} , avec S_{ij}^{PrPh} l'activité du neurone ij de PrPh. Ces neurones sont liés à la sortie S^{L_i} du neurone codant pour la vue locale i, et à la sortie S^{Θ_j} du neurone

codant pour son azimut j.[1]

Une mémoire à court terme est ajoutée au niveau de PrPh mémorisant les entrées passées. Avant chaque exploration séquentielle du panorama, toutes les activités des neurones de PrPh sont remises à 0. La mémoire à court terme permet de mémoriser au cours d'une exploration de panorama l'activité de l'ensemble des couples d'information "quoi" et "où".

Le pattern d'activité de la matrice PrPh permet une caractérisation visuelle du lieu dans lequel se trouve le robot. L'activité globale du PrPh est catégorisée comme un pattern codant une représentation invariante du lieu. Il est appris par un groupe de neurones représentant EC (cortex entorhinal). Un signal de neuromodulation reçu par EC permet le recrutement d'un nouveau neurone pour apprendre à caractériser le lieu visuel courant. Ces neurones de EC correspondent donc à des cellules de lieu. L'activité de ces cellules de lieu découle directement de la distance entre le pattern d'activité du PrPh au moment de l'apprentissage et le pattern courant. La variation de cette distance traduit une plus ou moins bonne reconnaissance d'un lieu appris. L'activité P_k d'une cellule de lieu k s'exprime de la manière suivante :

$$P_k = \frac{1}{l_k} \sum_{i=1}^{L} \omega_{ik}.f_s(L_i).(1 - g_d(\theta_{ik}^L - \theta_i)) \tag{8.1}$$

$l_k = \sum_{i=1}^{l} \omega_{ik}$. $\omega_{ik} = \{0, 1\}$ exprime le fait que la vue locale i a servi à coder le lieu k. l_k est le nombre d'amers utilisé pour coder la cellule k. L est le nombre de vues locales apprises. L_i est le niveau de reconnaissance de la vue locale i. $f_s(x)$ est la fonction d'activation du groupe de neurones Pr de reconnaissance des amers. θ_{ik}^L est l'azimut appris par le groupe de neurones codant pour l'azimut de la i^{eme} vue locale pour le k^{eme} lieu. θ_i est l'azimut courant de la vue locale i. d est un paramètre de diffusion de l'information angulaire définissant la pente de la fonction $g_d(x)$. Le rôle de $f_s(x)$ et $g_d(x)$ est d'adapter la dynamique des neurones codant l'information "quoi" et "où". Leur définition est :

$$g_d(x) = \begin{cases} \frac{|x|}{d.\pi} \; si \; |x| \; > \; d.\pi \\ 1 \; sinon \end{cases} \tag{8.2}$$

$$f_s(x) = \frac{1}{1-s}[x - s]^+ \tag{8.3}$$

où $[x]^+ = x \; si \; x > 0$, $et \; 0 \; sinon$. La figure 8-3 montre l'activité de 4 cellules après apprentissage lors d'un parcourt effectué par un robot passant dans 4 lieux sur ce parcourt. On peut observer l'étendue et l'homogénéité de leurs champs d'activité.

8.2.2 Les conditionnements sensori-moteurs

Les cellules de lieu permettent de reconnaitre différents amers dans un environnement mais ne permettent pas de se diriger vers un lieu précis. Il reste donc à associer cette information visuelle à une direction désirée pour qu'une action s'en suive.

L'architecture PerAc permet d'apprendre des conditionnements sensori-moteurs. En utilisant la reconnaissance de cellules de lieux comme sensation et une direction à prendre comme conditionnement. Une sélection de l'action à réaliser peut être apprise en fonction des sensations. La figure 8-4 présente cette architecture. On retrouve sur la partie gauche de cette figure l'ap-

1. Afin d'éviter des problèmes d'explosition combinatoire, les neurones de $PrPh$ ne sont pas précablés à l'avance mais sont recrutés en fonction des configurations à apprendre ce qui permet de limiter le nombre de neurones nécessaires.

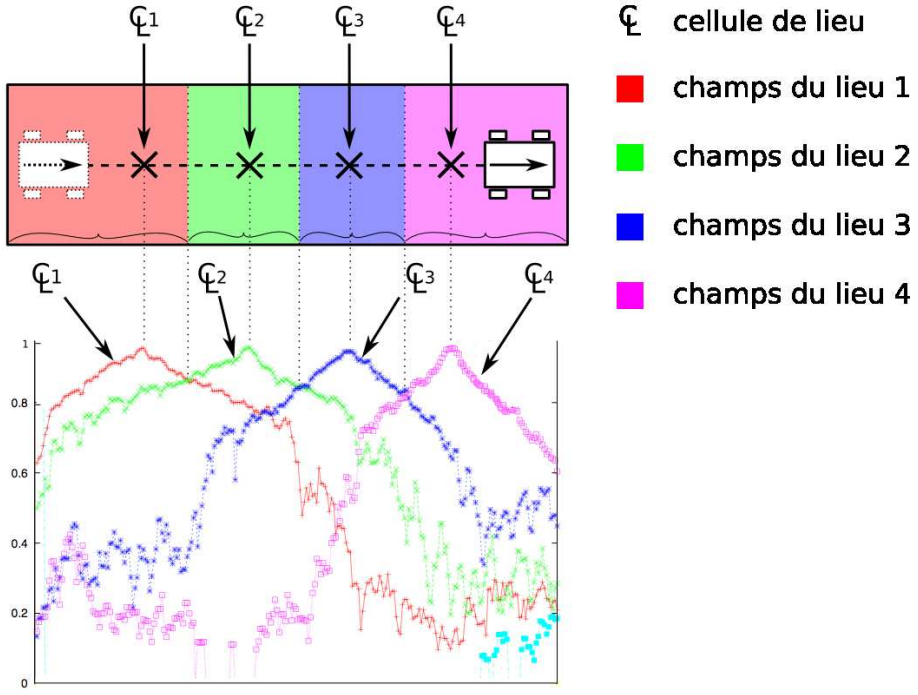


Fig. 8-3 – Activités de 4 cellules de lieu préalablement apprises en ligne. Lorsque le robot refait le parcours, il reconnaît successivement les lieux 1 à 4. L'activité de chaque cellule de lieu dépend de la distance du robot au lieu appris par la sélection de la cellule la plus active à chaque pas. La zone dans laquelle une cellule de lieu a l'activité la plus forte est son champs de lieu.

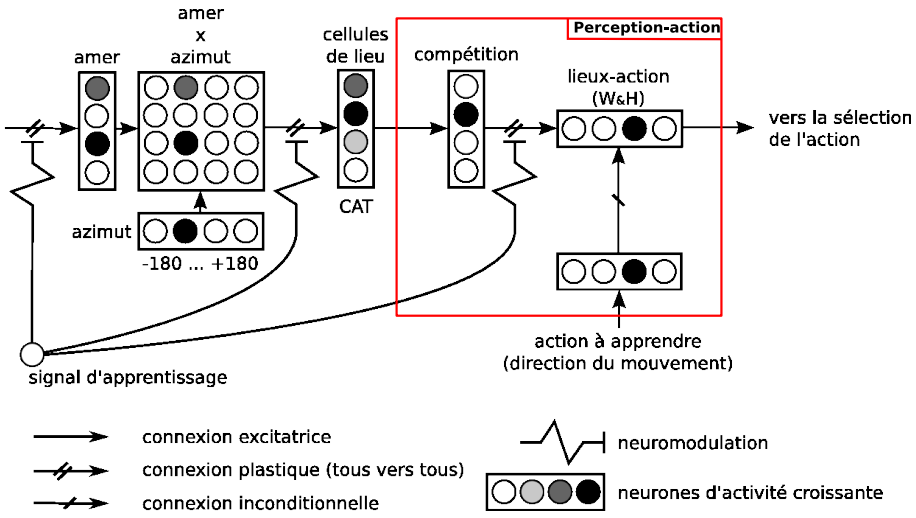


Fig. 8-4 – L'architecture PerAc permet l'apprentissage de conditionnements sensori-moteurs entre la reconnaissance d'une perception (cellule de lieu) et d'une action (direction à prendre). Ces conditionnements permettent ensuite au robot de naviguer dans son environnement de façon autonome.

prentissage des cellules de lieu. Le signal de neuromodulation permet d'une part d'apprendre les amers visuels caractérisant un lieu et d'autre part d'apprendre à catégoriser les différents motifs d'activité de la matrice PrPh correspondant aux différents lieux de l'environnement. Une compétition permet ensuite de déterminer le lieu courant (i.e. correspondant à la cellule de lieu

la plus active). Quand un lieu vient d'être appris, l'activité de la cellule correspondante est proche de 1 (le maximum d'activité). L'apprentissage sensori-moteur à proprement parler est réalisé par un groupe de neurones apprenant successivement les conditionnements entre cellules de lieu et actions à effectuer. Dans notre cas, ces actions correspondent à des directions que le robot doit prendre. Il s'agit donc d'actions liées au comportement de navigation. Ces actions sont représentées sur un champ de neurones. Chacun des neurones code pour une orientation. La précision en orientation dépend de la taille de ce champs. Pour un champs de taille n, la précision de la discrétisation angulaire est donc simplement de $\frac{360}{n}°$. Le groupe de neurones réalisant l'apprentissage du conditionnement sensori-moteur, le champs lieu-action, a la même taille que le champs de neurones des actions à apprendre. Ses neurones utilisant la règle d'apprentissage de Widrow et Hoff. Chaque neurone reçoit une connexion inconditionnelle depuis le neurone correspondant à la direction du mouvement à apprendre. Il s'agit du signal à apprendre (la direction à prendre). Chaque neurone reçoit aussi des connexions conditionnelles (plastiques) depuis chaque cellule de lieu. L'apprentissage consiste à modifier le poids des liaisons entre les cellules de lieu et les neurones du champs lieu-action afin de permettre à la cellule de lieu gagnante d'activer dans le champs lieu-action l'action qui y est associée. Après apprentissage, la reconnaissance d'un lieu active une direction à prendre. Cette commande motrice peut alors être exécutée par le robot.

Pour comprendre comment cette architecture permet à un robot de naviguer, nous prendrons l'exemple d'une tâche de retour au nid (*homing*). Pour rejoindre son but, le robot apprend des couples lieu-action convergeant vers le but. Ces conditionnements sensori-moteur créent un bassin d'attraction qui définit dynamiquement l'attraction vers le but sans nécessiter d'apprendre ou même de définir explicitement celui-ci. Une fois appris un minimum de trois lieux-actions convergeant vers le but, le robot agit en prenant la direction associée au lieu dans lequel il se trouve. Itération après itération, la distance du robot au but se trouve ainsi réduite : le robot est comme attiré par le but. Cependant, pour que ce mécanisme soit opérant, il faut que le robot puisse mettre en compétition les différents lieux qu'il perçoit. C'est ainsi qu'il peut définir dans quel lieu il se trouve. Cette compétition permet une généralisation des apprentissages i.e. la convergence vers le but est possible depuis toute une zone autour de ce dernier. Cette généralisation vient du fait que la décision réelle n'est prise qu'au niveau moteur et doit être comprise en fonction de la dynamique globale du système. Cependant, elle a ses limites. Au delà d'une certaine distance du lieu d'apprentissage, l'activité d'une cellule de lieu n'est plus que du bruit. Le robot ne peut plus alors déterminer quel lieu est le plus actif, donc le plus proche. A une distance suffisante d'un bassin d'attraction, les actions sélectionnées par le robot ne permettent alors plus de converger vers le but. Le bruit sur la perception visuelle est trop important et le robot navigue aléatoirement. La figure 8-5 montre les trajectoires du robot depuis différents points de départ et après avoir appris quatre lieu-action autour du but. Tant que le point de départ du robot est à l'intérieur du bassin d'attraction formé, sa trajectoire converge vers le but. Depuis des positions de départ plus lointaines, le robot navigue aléatoirement. Cette navigation aléatoire n'exclue pas que le robot finisse pas "tomber" dans le bassin d'attraction, cependant, le bassin d'attraction définit les limites de la convergence assurée vers le but.

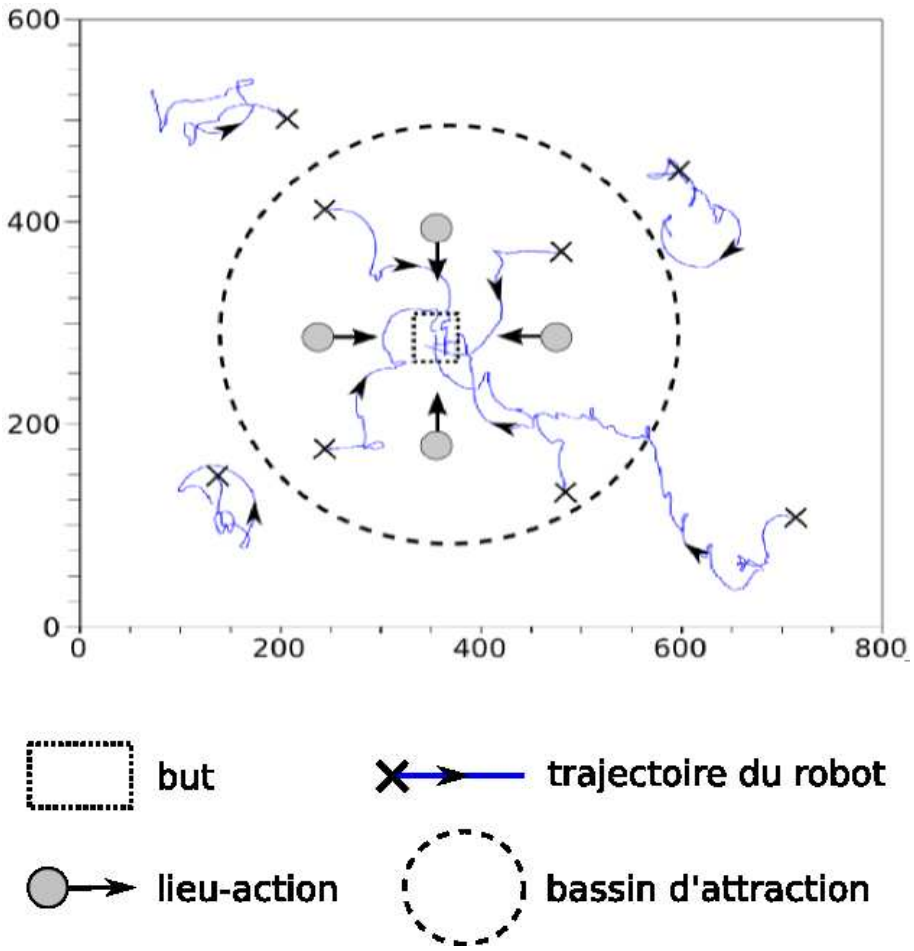


Fig. 8-5 – Trajectoires du robot ayant appris un bassin d'attraction autours d'un but. Le robot a appris quatre couples lieu-action convergeant vers un but. Quand il est dans le bassin d'attraction, ses actions le rapprochent de son but. En dehors du bassin d'attraction, le robot navigue aléatoirement. Il est possible que le robot rejoigne le bassin d'attraction mais cela repose sur le hasard et non sur un apprentissage.

8.3 Couplage environnemental ou comportemental entre le système de navigation et le système de reconnaissance d'expressions faciales

Le référencement social peut se référer à un objet, une personne, une action, un lieu de l'environnement et sans nul doute à d'autres évènements de la vie de tous les jours, montrant que la reconnaissance des expressions émotionnelles peut être interprétée et utilisée dans des tâches très différentes. Dans notre cas, si nous prenons l'exemple d'un expérimentateur qui affiche une expression de joie, le robot peut utiliser cette expression comme un signal qualifiant son comportement. Dans ce cas précis, l'action réalisée par le robot dans un lieu précis doit être apprise comme ayant une valeur positive. Mais on pourrait très bien imaginer aussi que le robot utilise ce signal émotionnel pour qualifier son environnement, indiquant ainsi l'intérêt d'un lieu que le robot devra éventuellement rejoindre. Nous étudierons ces deux couplages entre la navigation visuelle et l'interaction émotionnelle.

8.3.1 Couplage comportemental

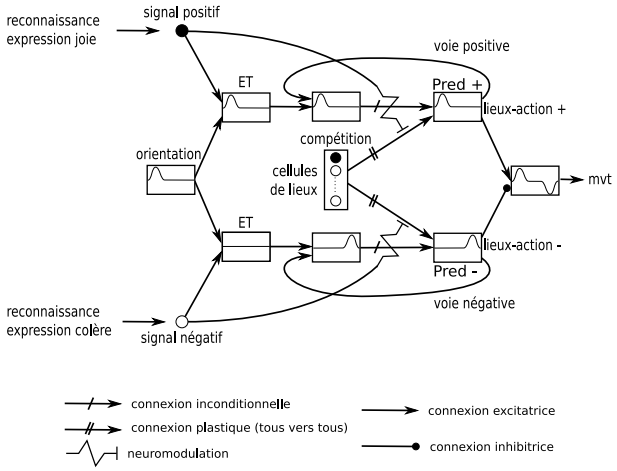


Fig. 8-6 – Modèle du couplage comportementale. Quand l'un des groupes de conditionnement $Pred-$ ou $Pred+$ reçoit une neuromodulation venant de la reconnaissance des expressions faciales (dans le cas décrit, il s'agit de l'expression de joie), l'association entre la localisation courante du robot (perception d'une cellule de lieux gagnante spécifique) et sa direction courante donnée par le compass (sommée avec ce qui a déjà été appris par ce groupe de neurone dans le passé) est apprise. Les neurones correspondant aux signaux "positifs" et "négatifs" sont associés à la reconnaissance de l'expression faciale de joie ou de colère du partenaire social interagissant avec la tête de robot.

Le couplage comportemental se rapporte à la situation où la reconnaissance de l'expression faciale de l'expérimentateur est utilisée afin de qualifier le comportement du robot. Dans le cas d'une interaction entre l'humain et le dispositif robotique, l'humain exprime une expression positive dans le but de renforcer positivement le comportement du robot. Au contraire, le visage de l'expérimentateur exprimant de la colère doit pouvoir inhiber le comportement du robot. Dans le but d'atteindre cet objectif, nous avons adapté l'architecture PerAc (figure 8-4) pour être capable d'apprendre aussi bien des actions positives que des actions négatives grâce à un simple conditionnement. L'architecture PerAc est capable de créer des bassins d'attraction sensori-moteurs afin d'amener le robot dans un lieu précis de l'environnement (figure 8-5). Cependant, nous y ajoutons la capacité d'inhiber dans un lieu une action, permettant ainsi de créer des bassins de répulsion sensori-moteurs. Pour assurer ces conditionnements, nous utilisons la règle d'apprentissage de Widrow et Hoff qui utilise la différence entre la sortie du réseau de neurone et la sortie désirée pour calculer les connexions des poids synaptiques qui doivent être modifiés (adaptation des poids par l'intermédiaire du mécanisme de conditionnement):

$$\Delta w_{ij} = \epsilon . I_i . (O_j^d - O_j) \tag{8.4}$$

Δw_{ij} est la variation des poids w_{ij} à appliquer, ϵ est la vitesse d'apprentissage équivalente à une neuromodulation du réseau, I_i est la ième entrée visuelle (cellule de lieux), O_j est la jème sortie du réseau de neurone et O_j^d est la jème sortie désirée.

Une interaction positive conditionne un mouvement en direction du but (atteindre une zone de l'environnement), tandis qu'une interaction négative inhibe le mouvement (une direction) vers le but pour par exemple éviter un lieu menaçant. Cependant, l'architecture neuronale décrite sur la figure 8-4 peut apprendre exclusivement des conditionnements positifs. Pour cette raison,

nous utilisons un premier réseau de neurones permettant l'apprentissage de toutes les associations positives tandis qu'un second réseau de neurones est utilisé pour l'apprentissage de tous les conditionnements négatifs. Un dernier groupe de neurones est utilisé pour calculer la somme des deux sorties afin de choisir le bon mouvement à réaliser (Figure 8-6). Le groupe de conditionnement positif a des connexions positives avec le groupe de sommation tandis que le groupe de conditionnement négatif a des connexions négatives. Cette solution permet de mémoriser beaucoup plus d'information sur ce qui est appris par le robot que des sorties avec des valeurs positives ou négatives. Cette solution connexionniste est biologiquement plausible car des synapses excitatrices et inhibitrices existent. Elles permettent d'imaginer des situations particulières. Par exemple, ayant appris qu'un comportement particulier est bon et qu'un peu plus tard le même comportement est considéré comme mauvais, alors cela peut signifier à la fois que l'environnement à changer ou que les objectifs de l'expérimentateur ont changé. Le comportement du robot est adaptatif, lui permettant de modifier au fur et à mesure des interactions son point de vue. Quand le robot reçoit un signal au travers de l'interaction sociale, l'apprentissage d'une nou-

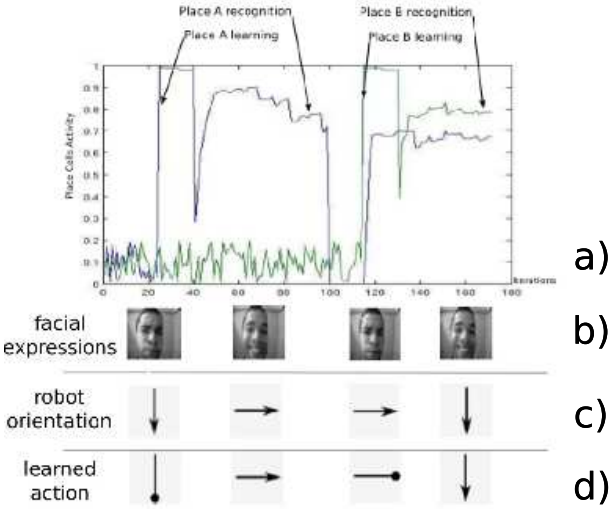


Fig. 8-7 – Expérience du couplage comportemental. a) activité neuronale des cellules de lieux. b) les expressions faciales de l'expérimentateur reconnues par la tête de robot. c) direction du mouvement courante du robot. d) différentes actions apprises par le robot en fonction du lieu dans lequel il se situe. Une flèche signifie une direction renforcée positivement (le robot réalise une bonne action, il peut continuer), tandis que les droites finissant par un point correspondent à des directions renforcées négativement (le robot a réalisé une mauvaise action, son action est inhibé). Les expressions faciales de l'expérimentateur donnent au robot les informations dont il a besoin sur son comportement pour apprendre les associations sensori-motrices nécessaires entre le signal visuel (reconnaissance des lieux) et la direction de mouvement que le robot doit réaliser pour réaliser la tâche. A l'issue de cette apprentissage, le robot sera capable grâce à ces associations d'activer ou d'inhiber certaines directions en fonction du lieu dans lequel il est.

velle cellule de lieux est déclenché ainsi que l'apprentissage instantané entre la nouvelle cellule de lieux et l'action courante. Néanmoins, si une cellule de lieux déjà existante est proche de la position courante du robot (défini par un seuil fixe sur le niveau de reconnaissance des cellules de lieux) l'apprentissage d'une nouvelle cellule de lieux est inhibée et le conditionnement sensorimoteur est appris avec la cellule de lieux la plus proche c'est à dire avec la cellule de lieux la plus active. Ceci complétent l'ancien apprentissage sensori-moteur. Les bassins sensori-moteurs se forment petit à petit jusqu'à convergence du réseau neuronal. Le robot est donc capable d'ap-

prendre progressivement via des interactions avec l'humain qu'elles sont les directions à éviter et qu'elles sont les directions à considérer pour atteindre certaines zones de l'environnement. Nous avons testé cette architecture dans la situation suivante: L'environnement du robot contient une zone d'intérêt et l'expérimentateur veut enseigner au robot comment atteindre cette zone. Le processus d'apprentissage est naturel car à chaque pas de temps l'expérimentateur exprime des expressions émotionnelles pour informer le robot de son bon ou mauvais comportement. Lorsque l'expérimentateur observe le robot et que ce dernier réalise un bon comportement alors l'expérimentateur exprime une expression de joie en direction de la tête expressive. Au contraire quand le robot réalise un mouvement dans une mauvaise direction, l'expérimentateur exprime une expression de colère vers la tête robotique. Le figure 8-7 est une illustration chronologique durant l'apprentissage de l'ensemble de la tâche. La figure 8-7 montre l'apprentissage d'une cellule de lieu à chaque interaction avec le dispositif robotique. Nous observons que le robot réalise des mouvements et que ces mouvements peuvent être renforcés positivement ou inhibés selon le signal émotionnel envoyer par l'expérimentateur. Au même moment, les cellules de lieux sont

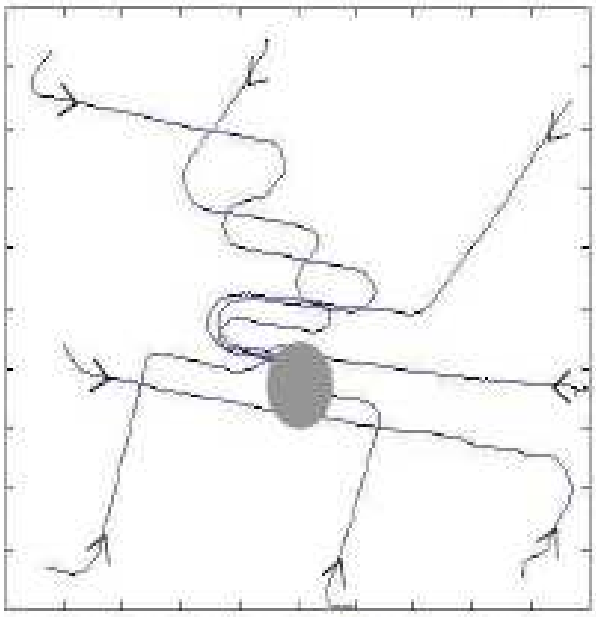


Fig. 8-8 – Les trajectoires du robot provenant de différents points de départ: le robot est capable d'atteindre les lieux associés à l'expression faciale de joie. La zone grise représente le lieu but. Ces trajectoires sont obtenues par un tracking vidéo. La taille de la zone où se déroule l'expérience est de 3m x 3m.

associées aux actions et seules celles nous menant au but seront suffisamment actives pour entrainer le robot dans la zone à rejoindre. Un bassin d'attraction est par conséquent creusé. La figure 8-8 montre les trajectoires du robot pour rejoindre la zone d'intérêt après l'apprentissage. Cette figure montre la capacité du robot à atteindre la zone d'intérêt en partant de positions initiales différentes dans l'environnement. Il est également important de prendre en compte le fait que le robot apprend beaucoup plus d'information sur la tâche à réaliser lorsque le comportement est qualifié positivement par le partenaire social que quand le comportement est qualifié de négativement. Les interactions positives sont un moyen rapide de faire converger le robot dans un bassin d'attraction contrairement à des interactions négatives qui vont inhiber certains

mouvements. L'apprentissage du bassin d'attraction autour de la zone d'intérêt (i.e. un ensemble de lieux-actions qui assure une dynamique de navigation convergente vers un but) prend entre 3 et 5 minutes.

8.3.2 Couplage environnemental

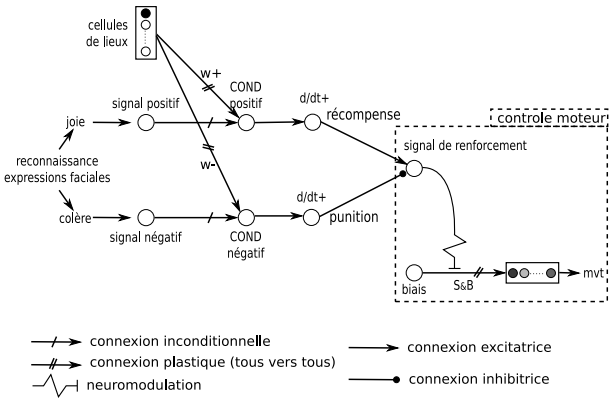


Fig. 8-9 – Modèle du couplage environnemental. En utilisant la règle d'apprentissage des moindres carrés, le conditionnement des neurones permettent l'association entre une cellule de lieux (une zone de l'environnement) et l'expression faciale de l'expérimentateur (modification des poids w^+ et w^-). La dérivée temporelle du signal expressif prédit est utilisée comme un signal de renforcement (règle d'apprentissage de Sutton&Barto) pour maintenir ou changer la direction du robot mobile. Le lien modifiable entre le biais et le groupe $S\&B$ permet l'apprentissage du mouvement à prendre par le robot.

Le couplage environnemental se réfère à la situation où la reconnaissance d'une expression faciale émotionnelle est utilisée pour qualifier l'environnement immédiat du robot. L'expérimentateur affiche par exemple un visage de joie dans le but d'enseigner au robot que la zone est "bonne" (ressource dont le robot à besoin) et au contraire pour une expression de colère que la zone est "mauvaise". Le robot atteind les lieux lui procurant du "plaisir" et évite les lieux menaçant pour lui. Afin de résoudre cette tâche, nous utilisons une nouvelle fois une version modifiée de l'architecture PerAc.

Dans ce paradigme expérimental, le signal émotionnel est vue comme une récompense associée au lieu courant, il permet de qualifier certains lieux de l'environnement. Ainsi au moment de recevoir ce signal, le robot doit apprendre une nouvelle cellule de lieu caractérisant sa localisation courante et doit également savoir prédire le signal de l'interaction (Figure 8-9). Ainsi au moment où, le robot se rapprochera d'un lieu appris, la réponse de la cellule de lieux augmentera ainsi que la récompense prédite associée. A l'inverse lorsque le robot s'éloignera d'un lieu appris, la réponse de la cellule décroira . Au lieu d'utiliser un apprentissage par conditionnement entre une perception visuelle et une action, nous avons utilisé un modèle d'apprentissage par renforcement utilisant la règle de Sutton et Barto (Barto et al., 1981) où la dérivée de la récompense prédite sert de signal de renforcement:

$$\Delta R = \frac{dPredH}{dt} - \frac{dPredA}{dt} \tag{8.5}$$

$$\Delta w_{ij}^{+/-} = \epsilon.\Delta R.\frac{dO_j}{dt}.I_i \tag{8.6}$$

ΔR est le signal de renforcement, $\frac{dPredH}{dt}$ est la dérivée du signal de joie prédit (Happiness), $\frac{dPredA}{dt}$ est la dérivée du signal de colère prédit (Anger). $\Delta w^{+/-}$ est la différence entre les anciens et nouveaux poids, ϵ est la vitesse d'apprentissage (la neuromodulation du réseau de neurone), $\frac{dO_j}{dt}$ est la variation de la jème sortie du réseau et I_i est la ième l'entrée. Un groupe de

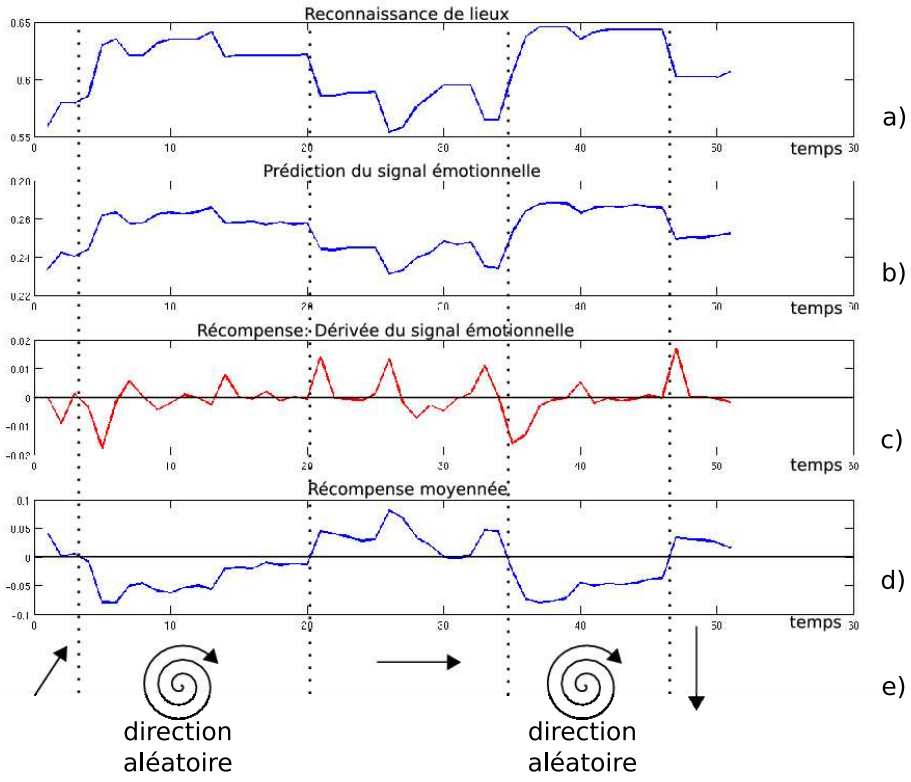


Fig. 8-10 – Figurant montrant les différents signaux utilisés pour réaliser le couplage environnemental. a) activité neuronale des cellules de lieux. b) Signal de prédiction de récompense: nous informe sur l'éloignement ou le rapprochement d'une zone de danger, plus l'activité est forte plus l'on se rapproche de la zone de danger. c) Dérivée de la prédiction de la récompense d) Dérivée de la prédiction de la récompense moyennée, elle est utilisée comme le signal de renforcement pour réaliser le contrôle moteur. e) Lorsque la dérivée est négative, le robot change de direction constamment et au contraire quand la dérivée est positive, le robot maintient sa direction.

neurones est utilisé pour le contrôle moteur du robot mobile, ce groupe est connecté à une entrée constante qui permet le contrôle grâce à un simple signal de renforcement. Sans renforcement, ce groupe moteur produit des sorties aléatoires permettant au robot d'essayer une autre action grâce à l'introduction d'un petit bruit sur chaque neurone de sortie, permettant l'exploration de l'environnement (même aléatoirement) et parfois même de sortir de minimum local. Deux cas sont envisageables durant l'expérience que nous proposons:

- Le renforcement est positif: $\Delta R > 0$ si et seulement si $\frac{dPredH}{dt} > \frac{dPredA}{dt}$. La sortie courante dans le groupe de neurones du Sutton & Barto sera renforcée, permettant au robot de garder la même direction.
- Le renforcement est négatif: $\Delta R < 0$ si et seulement si $\frac{dPredH}{dt} < \frac{dPredA}{dt}$. La sortie courante dans le groupe de neurones du Sutton & Barto sera inhibée, permettant au robot de changer

de direction car son comportement actuel est mauvais.

Pour tester cette architecture, nous attribuons une direction fixe au robot en fonction de son point de départ une fois qu'il a appris au travers de l'interaction avec l'expérimentateur que la zone centrale de son environnement est une zone dangereuse (i.e. associée avec une expression de colère). La figure 8-10 montre comment des directions qui produisent des signaux de récompense positives sont renforcés positivement (éloignement de la zone de danger) tandis que des directions qui produisent des signaux de récompense négatives sont renforcés négativement (rapprochement vers la zone de danger). Cette figure montre certains signaux cruciaux pour le fonctionnement du système. Tout d'abord, nous pouvons observer l'activité de la cellule de lieux qui nous renseigne sur la proximité d'un lieu appris. La prédiction du signal émotionnel nous informe sur la proximité d'une zone de danger, plus l'activité est forte plus l'on se rapproche de celle-ci. Enfin, le renforcement émotionnel correspond à la dérivée du signal de prédiction. Le comportement du signal de renforcement est simple: si le robot s'éloigne de la zone menaçante alors ce signal est positif, et inversement si le robot s'en rapproche le signal de renforcement devient négatif. L'action réalisée par le robot est dépendante du signal de renforcement émotionnel. Le robot renforce la même action si le signal de renforcement est positif tandis qu'il change aléatoirement de direction si le signal de renforcement est négatif. La figure 8-11 montre les différentes trajec-

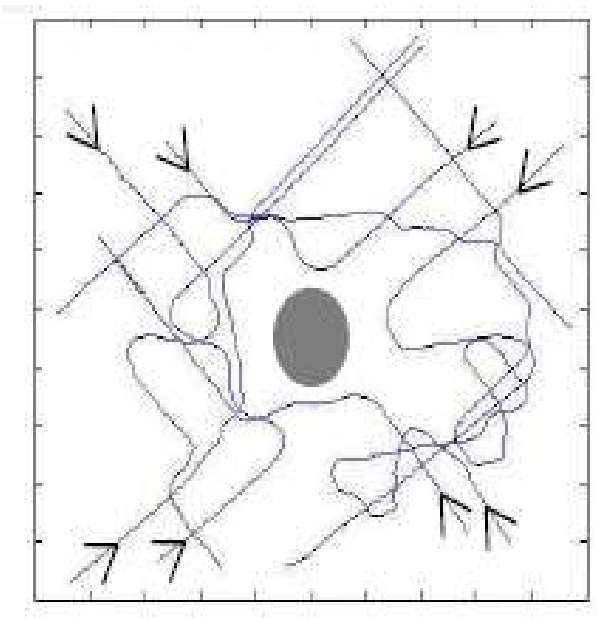


Fig. 8-11 – Les trajectoires du robot ayant pour origine différents points de départ (au départ le robot à une direction fixe) après un apprentissage interactif de la zone grise. La zone grise est associée à une expression faciale négative (l'expression de colère) ce qui implique que le robot est capable de fuir ou d'éviter cette zone. La prédiction d'un renforcement négatif est suffisante pour inhiber un mouvement dans une direction, la zone de danger peut dans ce cas être évitée.

toires que le robot réalise en partant de différents points de départ et avec une direction fixe par point de départ. Nous pouvons observer que le robot au départ garde sa trajectoire fixe et lorsqu'il se rapproche suffisamment de la zone de danger alors il évite ce lieu. Le référencement d'un lieu au travers des interactions avec l'expérimentateur permet au robot d'apprendre rapidement à les éviter. Néanmoins, la tâche devrait être beaucoup plus difficile si nous voulions enseigner

au robot d'atteindre un lieu contrairement à l'éviter. Les signaux fournis par les sensations visuelles sont très bruités, or les signaux de renforcement sont fonctions des signaux visuels. Par conséquent, l'ensemble de ces pertubations empêchent la convergeance du robot vers un but d'où l'utilité d'une architecture PerAc qui elle crée des bassins d'attraction sensori-moteur.

8.4 Conclusion

Les signaux expressifs d'interaction peuvent être utilisé par le robot pour apprendre à qualifier son environnement ou son propre comportement. Le couplage comportemental (association d'une valeur émotionnelle aux actions du robot) et environnemental (association d'une valeur émotionnelle à l'environnement du robot) apportent une grande capacité d'adaptation au robot. Une solution complète et élégante serait de donner au robot la capacité d'effectuer ces deux types couplages en parallèles. Néanmoins, le développement d'une architecture fusionnant les deux modèles est loin d'être trivial du fait même de l'ambiguité du signal d'interaction. L'expression faciale peut signifier à la fois "cette zone de l'environnement est bonne" ou "cette action dans ce lieux est correcte". Dans cette situation, la communication expressive comporte une ambiguïté: "A quel couplage l'humain fait-il référence lorsqu'il interagit avec le robot ?". Cette question est loin d'être triviale.

Une solution simple pour résoudre ce problème serait de donner au robot la capacité de reconnaitre et d'interpréter plusieurs expressions faciales. Par exemple, une expression faciale de colère pourrait signifier que le comportement du robot est "mauvais" tandis qu'une expression faciale de peur pourrait signifier que le robot est dans une partie de l'environnement dangereux. Néanmoins, cette solution trouve rapidement des limitations, la mise à l'échelle parait complexe car cette solution ne fournit pas un réel couplage de ces deux architectures, elle montre simplement que les deux architectures peuvent s'exécuter en parallèle. Cette solution donne plus l'avantage à l'expertise d'un expérimentateur sur l'architecture du robot que sur la capacité naturelle d'interagir émotionnellement avec un dispositif robotique. Une autre solution pourrait être la manière dont le modèle traite les entrées de l'interaction émotionnelle. Le couplage comportemental utilise un signal phasique (le moment où le signal apparait) tandis que le couplage environnemental utilise un signal tonique (le temps global où le signal est présent). De cette manière, les deux couplages devraient fonctionner avec les mêmes entrées avec pour unique différence leur mode de fonctionnement. Bien entendue, la question de la cohérence de l'apprentissage est essentielle: si le robot est entrain de faire quelque chose de "mauvais" (par exemple s'éloigner d'une ressource dont il a besoin) l'expérimentateur affichera un visage de colère et le robot apprendra en même temps que son comportement était mauvais mais aussi que le lieu est à éviter. Une dernière solution pourrait venir de l'apprentissage par renforcement de type Qlearning (Watkins and Dayan, 1992). Si l'expérimentateur exprime de la joie alors les couples lieux-actions seront renforcés positivement. Au contraire, si l'expérimentateur exprime de la colère alors les couples lieux-actions seront renforcés négativement. Dans cette solution, la valeur émotionnelle est attribuée aux couples lieux-actions ce qui implique le référencement des lieux et des actions.

Chapitre 9

Attention Conjointe et Référencement Social

9.1 Introduction

L'attention conjointe est un élément clé de la cognition sociale. Elle nous apprend beaucoup sur les relations sociales et est souvent considérée comme un élément prépondérant de la théorie de l'esprit (Premack and Woodruff, 1978). L'attention conjointe est le mécanisme permettant l'orientation de son propre regard en fonction du regard de l'autre. (Emery, 2000) définit l'attention conjointe comme un interaction triadique, montrant que deux agents portent leur attention sur un même objet. L'agent 1 détecte que le regard de l'agent 2 n'est pas dirigé vers lui et par conséquent suit la direction de son regard afin de regarder "l'objet" d'attention de l'agent 1. Cette définition met en avant un processus uni-directionnel, contrairement à l'attention partagée qui semble être un couplage entre attention mutuelle et attention conjointe. Dans l'attention partagée, l'attention des deux agents porte à la fois sur l'objet mais également sur l'autre agent ("je sais que tu regardes l'objet et tu sais que je regarde l'objet"). La figure 9-1 illustre ces deux mécanismes attentionnels entre deux agents et un objet.

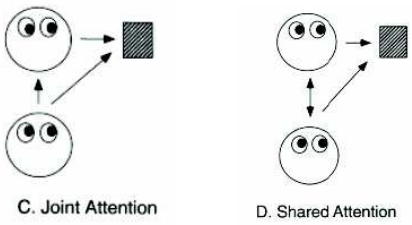


Fig. 9-1 – Attention conjointe - Attention partagéee.

Le domaine de la robotique s'est également intéressé à ce mécanisme puisqu'il semble primordial pour l'interaction sociale afin d'obtenir des robots intéragissant dans un environnement social. (Scassellati, 1998) implémente le modèle de Baron-Cohen (Baron Cohen, 1995) sur un robot humanoide. Selon Baron-Cohen, l'attention conjointe est basée sur deux modules:

– le détecteur de l'intentionnalité (ID) utilise les modalités sensorielles et permet d'interpréter les actions d'autrui en termes de but, d'objectif et de désir.

– le détecteur de la direction du regard (EED) permet de détecter la présence et la direction du regard d'autrui sur soi ou ailleurs. Il permet d'inférer qu'un individu regarde un objet si le regard est dirigé vers cette objet.

Ce modèle suggère que le mécanisme d'attention conjointe se base exclusivement sur ces deux détecteurs. ID permet l'interprétation de la direction du regard comme un état-but, rendant ainsi possible l'interprétation du regard de l'autre comme des intentions. Ce mécanisme permet aux deux agents de focaliser leur attention sur un même objet, agent ou événement.

Les travaux de (Scassellati, 1998; Scassellati, 2001) n'expliquent pas vraiment le développement de ces capacités malgré son approche basée sur le modèle de Baron-Cohen. Leur détecteur de l'intentionnalité est un module inné. De notre point de vue, ce module devrait être appris à travers l'interaction. Le modèle développé par (Yucel et al., 2009) implémente un modèle relativement efficace, consistant à intégrer des algorithmes de traitement d'image robustes comme l'estimation de la pose de la tête et l'estimation de la direction du regard. D'autres auteurs comme (Marin-Urias et al., 2009; Marin-Urias et al., 2008; Sisbot et al., 2007) se sont focalisés sur des capacités importantes de l'attention partagée nommées "mental rotation" et "perspective taking". Ces compétences permettent au robot humanoide HRP2 d'acquérir la représentation de l'environnement venant d'un autre point de vue et d'assimiler la notion génerale de raisonner à partir du point de vue de l'autre (l'expérimentateur humain) afin d'obtenir une représentation de la connaissance de l'autre. Autrement dit, le robot humanoide HRP2 est capable d'orienter son regard sur l'objet captivant l'attention de l'humain même si deux objets sont dans le même champ de vision avec l'un qui occulte l'autre. L'objet d'attention pour l'humain est celui qui est visible, le robot ayant assimilé que l'autre objet est non visible pour l'humain. Néanmoins, ces capacités semblent très intéressantes mais difficile à expliquer d'un point de vue développemental. Le modèle décrit par (Nagai et al., 2003) permet à un robot d'acquérir la capacité d'attention conjointe sans l'évaluation de la tâche. Ce modèle montre comment un robot peut interpréter le regard de l'humain afin de focaliser sur des objets de l'environnement. Cependant, un apriori sur la détection de visage est réalisé pour acquérir cette capacité. Ce modèle est malgré tout très intéressant pour comprendre certains stades du développement de l'attention conjointe et s'inscrit dans la même philosophie générale que nos propres travaux. (Kaplan and Hafner, 2004) résument assez bien le défi de l'attention conjointe en essayant tout d'abord de définir ce mécanisme ainsi que les éléments unitaires le constituant. Il nous paraît intéressant de comprendre comment ce mécanisme peut être développé et comment il peut s'intégrer afin de réaliser d'autres tâches plus complexes. Dans ce chapitre, nous nous intéresserons à comprendre

Fig. 9-2 – Set-up expérimental d'une expérience mêlant des capacités d'attention conjointe et de référencement social. La tête de robot expressive est capable d'apprendre à reconnaitre des expressions faciales de l'expérimentateur et peut également orienter son regard en fonction du regard de l'expérimentateur.

comment un prémices d'attention conjointe peut être construit grâce à une interaction avec un partenaire social dans le but d'intégrer cet aspect dans un cadre de référencement social. Nous décrirons tout d'abord comment un robot est capable d'apprendre des directions de regard ou à défaut des orientations du visage dans un contexte interactif. Et nous finirons par décrire un modéle basé sur des couplages perception-action intégrant un prémices d'attention conjointe afin d'attribuer des valeurs émotionnelles à des objets ayant attirés son attention (figure 9-2).

9.2 Apprentissage de la direction de la tête

9.2.1 Set-up expérimental

L'orientation du regard est apprise de manière autonome par la tête expressive tout comme les expressions faciales ou les primitives motrices par conséquent nous nous plaçons dans le même paradigme expérimental. Durant la phase d'apprentissage, le robot est inititateur dans l'interaction permettant ainsi un apprentissage auto-supervisé. Le robot oriente son regard aléatoirement (droite, gauche, en face) durant 2 secondes en repassant par "regard en face" entre chaque orientation. Il est demandé explicitement à l'expérimentateur d'imiter la tête robotique. Cette première phase d'apprentissage est très rapide entre 1 et 2 minutes, le générateur aléatoire est stoppé. Le robot ne dispose alors plus du signal réflexe permettant le déclenchement d'une action. En revanche lorsque le robot reconnait une vue de l'expérimentateur, il dirige son regard en fonctions des associations apprises et "imite" la direction du regard de l'expérimentateur. Maitenant, le robot est capable de diriger son regard en imitant la direction du regard de l'humain. D'un point de vue développemental, un scénario imaginable serait que le robot agit et interagit dans un environnement naturel par exemple un objet particulier attire l'attention du robot impliquant par conséquent une orientation de son regard vers celui-ci. Imaginons également qu'un expérimentateur interagit avec le robot de manière naturelle, l'attention du robot influencera l'attention de l'expérimentateur: le partenaire humain regarde là où le robot regarde. Grâce à un simple mécanisme associatif, le robot va être capable d'associer ce qu'il est entrain de faire avec ce qu'il voit. Une fois que ces associations sont construites, le robot possédera la capacité à diriger son attention en fonction de du regard de l'expérimentateur (Figure 9-3). Notre set-up expérimental est composé d'une tête de robot capable de focaliser dans des directions privilégiées. Une caméra inscrutée dans l'oeil du robot est capable de détecter la présence d'un objet dans son champ de vision. Nous faisons l'hypothèse qu'un seul objet est présent sur la scène. Une deuxième caméra fixe est capable d'apprendre la direction du regard de l'expérimentateur. La tête expressive contient deux caméras (les yeux) mais nous avons utilisé qu'une seule caméra car il nous aurait fallu traiter le problème du changement de référentiel induit par la rotation de la tête[1]. De plus, les caméras actuelles ayant une résolution limitée, il est difficile lorsque l'on choisit un champ de vision large (nécessaire pour voir à la fois l'objet cible et l'humain) d'avoir une résolution suffisante pour que la reconnaissance visuelle de l'objet soit robuste.

9.2.2 Architecture de controle neuronale

Cette section décrit une architecture sensori-motrice capable d'associer des sensations visuelles avec des mouvements oculaires. Après 1 à 2 minutes d'apprentissage, l'activation de certaines sensations visuelles permettent d'orienter le regard du robot en fonction du regard de l'expérimentateur humain. Comme pour la reconnaissance des expressions faciales ou des primi-

1. Ce problème est très intéressant, il fait l'objet d'une autre thèse

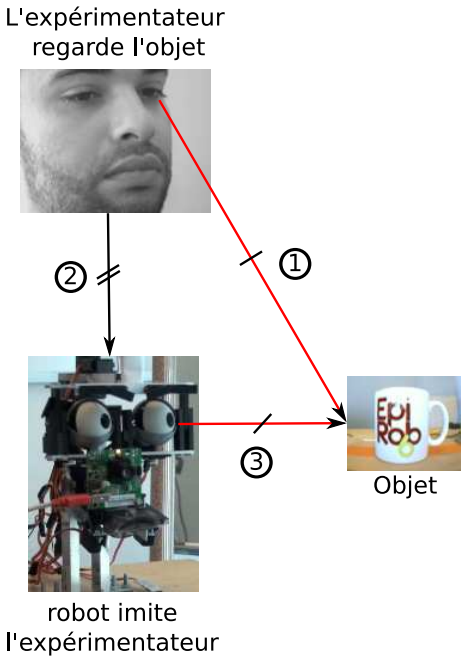


Fig. 9-3 – Attention conjointe entre les deux agents (individu vs robot). L'attention de l'expérimentateur est dirigée vers l'objet impliquant une résonnance motrice du robot. 3 phases sont présentées: 1)l'expérimentateur regarde l'objet, 2) le robot imite la direction du regard de l'humain et 3) le robot stabilise son regard sur l'objet.

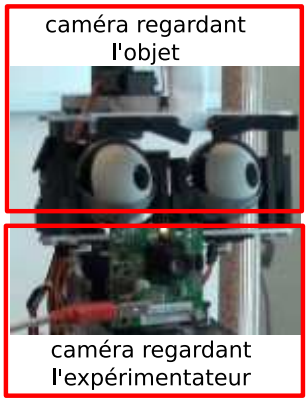


Fig. 9-4 – Matériel utilisé pour l'expérience d'attention conjointe. Une tête de robot composée d'une caméra située dans l'oeil du robot est capable de suivre un objet et d'une seconde caméra firewire fixe est utilisée pour apprendre et imiter l'expérimentateur.

tives motrices, l'apprentissage est effectué en ligne et de manière autonome (pas de cadrage de visage). Le processus visuel reste identique, il n'est pas règlé en fonction des caractéristiques visuelles que l'on veut apprendre montrant ainsi une certaine robustesse du traitement visuel. Les points de focalisation sélectionnés sont suffisants pour caractériser des orientations de regard.

Un groupe de neurones SAW apprend les différentes vues locales (voir chapitre 4). Les vues locales correspondent ici à des caractérisque codant pour l'orientation du regard et/ou de la tête. Pendant l'apprentissage lorsque l'expérimentateur humain imite la direction du regard du

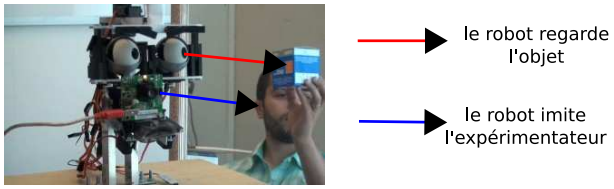


Fig. 9-5 – Set-up expérimental (phase de test): l'humain interagit avec le robot dans un contexte d'attention conjointe. L'humain dirige son regard vers un objet dans le but d'attirer le regard du robot sur cet objet.

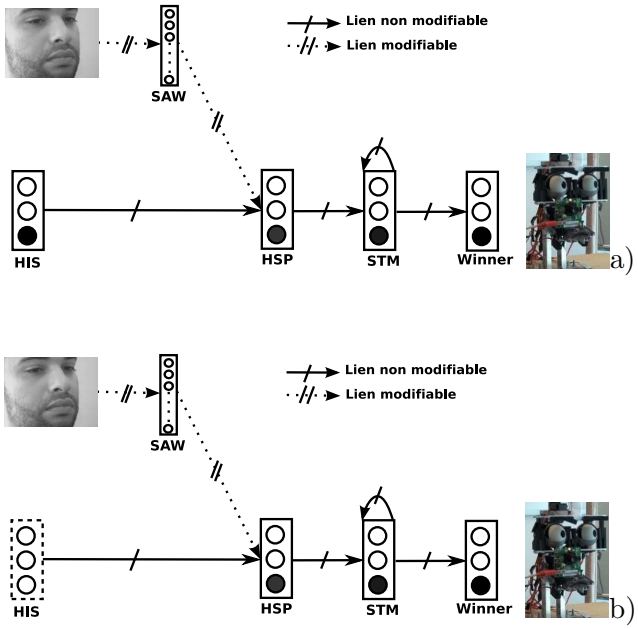


Fig. 9-6 – Architecture capable d'apprendre et de reconnaître la direction du regard de l'expérimentateur. a) représente l'architecture avant l'apprentissage, les flêches pointillés montrent les poids synaptiques n'influençant pas la reconnaissance de l'orientation du regard. Le robot n'a pas encore appris l'association entre ses mouvements oculaires et les caractéristiques visuelles décrivant la direction du regard de l'expérimentateur. b) montre l'architecture après l'apprentissage, la voie correspondant à la reconnaissance de la direction du regard (lien modifiable) est suffisante pour déclencher des mouvements oculomoteurs. La boite en pointillée correspond à l'inhibition de l'orientation réflexe du regard. 2 neurones codent pour deux positions extrêmes (regard à droit et à gauche) et le troisième neurone code pour la position neutre correspondant à l'expérimentateur regarde en face.

robot, la tête de l'expérimentateur bouge en fonction de l'endroit où il regarde. Les points de focalisation sont pris en majorité sur le visage supposé de face comme pour les expressions faciales mais également sur le visage pris de profil (joue, oreille, cheveux sont alors les plus visibles), des distracteurs sont également présents même s'ils sont peu nombreux du fait du focus serré sur la zone occupé par l'expérimentateur. Les neurones appartenant au SAW correspondent à de simple caractéristiques visuelles par exemple un oeil vu de profil. Une fois les vues locales quantifiées, nous devons les associer à une catégorie motrice. HSP correspond à la prédiction de la direction du regard. L'association se fait par un simple mécanisme de conditionnement (Widrow and Hoff, 1960)). HSP associe les sensations visuelles (SAW) avec la direction du regard du robot

(HIS). Ces associations renforcent les caractéristiques visuelles utiles pour la reconnaissance des orientations du regard et au contrairement à inhiber les caratéristiques n'ont discriminante (par exemple le fond de la scène). Une hystérésis temporelle est introduite par l'intermédiaire d'une mémoire à court terme (STM), ayant pour but de stabiliser le système. Sans cette dernière, les mouvements oculomoteurs du robot ont tendance à basculer plus d'une orientation de regard à l'autre. Elle limite les mauvaises reconnaissances.

L'architecture de la figure 9-6a) montre le réseau de neurones avant l'apprentissage où seule la voie réflexe est capable de déclencher des mouvements oculaires car les sensations visuelles n'ont pas encore été associées avec les mouvements oculaires. Au contraire la figure 9-6b) montre le réseau de neurones après l'apprentissage, les sensations visuelles sont capables à elles seules de déclencher des mouvements moteurs. Les liens modifiables sont suffisament actifs pour inhiber la voie réflexe. Le robot est maintenant capable d'imiter la direction du regard de l'expérimentateur humain.

Cependant, le robot est capable d'orienter son regard seulement dans trois directions (gauche, droite et en face). Une solution pour diriger plus finement le regard du robot (sans problème de coût de calcul ou de généralisation) est d'introduire la capacité à interpoler. Cette solution implique la capacité à interpoler les configurations apprises avec un modèle de champs de neurones dynamiques par exemple. L'introduction de bulles d'activités (gaussiennes) et la capacité à les sommer permet de généraliser a priori les apprentissages déjà faits (voir chapitres 6). Les résultats 9-10 mettent cette compétence en évidence.

9.2.3 Résultats

Fig. 9-7 – Base de donnée des orientations du regard et/ou de la tête.

Après l'apprentissage, la voie correspondant à la reconnaissance de la direction de regard est suffisante pour produire des mouvements occulomoteurs grâce aux interactions avec l'expérimentateur humain (figure 9-6b). Les résultats montrent la capacité du robot à orienter son regard en fonction du regard de l'individu interagissant avec lui. Le couplage entre l'ambiguité de la perception

et un système homéostatique permet l'émergence d'un suivi de la direction du regard en recherchant simplement un état d'équilibre. Le robot regarde dans la direction de l'expérimentateur en essayant de faire correspondre ses sensations visuelles avec un état moteur oculaire spécifique.

La figure 9-8 montre les taux de réussites de notre modèle sur 14 personnes ayant interagit avec lui. Les différents expérimentateurs ont dans un premier temps imité le robot durant 1 minute chacuns ("je" regarde là où le robot regarde). Dans un second temps, il est demandé à ces mêmes expérimentateurs de produire des orientations de regard que le robot doit imiter. Ces résultats montrent qu'en un temps très court d'apprentissage (1 à 2 minutes par personnes), le robot est capable de reproduire les orientations des différents expérimentateurs (figure 9-7). le taux de réussites est de 70% pour chacune des 3 orientations apprises. Ces résultats sont la sortie brute du système sans l'hystérésis temporelle ce qui montre la robustesse du système sans le mécanisme de mémoire (STM). La mémoire à court terme (STM) améliore grandement les résultats en pratiques en convergeant sur la sortie la plus fréquemment active au cours des itérations précédentes (le temps de traitement est 5 images/seconde). Nous avons également

Robot's orientation

Human's orientation	left	middle	right
left	**71%**	14%	16%
middle	6%	**67%**	27%
right	5%	23%	**72%**

Fig. 9-8 – Tableaux montrant le taux de confusion de l'orientation du regard. La base de donnée est constituée de 1260 images correspondant aux différentes orientations du regard. 14 personnes constituent la base de donnée, les expérimentateurs ont préalablement imité la tête de robot pour annoter les images en fonction de l'orientation du robot. Le système a appris 14 individus durant la phase d'apprentissage (1 minute par personnes) et par la suite la tête de robot a dû imiter les orientations du regard de ces 14 individus.

testé la robustesse du modèle sur des personnes n'ayant jamais interagit avec le dispositif durant la phase d'apprentissage. La figure 9-9 montre le taux de confusion sur ces nouveaux individus. Comme notre base de donnée ne contient que 14 personnes, nous utilisons une méthode de cross-validation. Le même processus est réalisé pour chaque expérimentateur à savoir on apprend l'orientation du regard sur 13 personnes puis l'on teste sur la 14eme. Les résultats mettent en évidence la capacité du système à généraliser (56% de réussite sur des individus non appris). Ces résultats laissent sous entendre qu'avec un nombre réduit d'individus, le robot imite la direction du regard d'individus encore jamais vus. Bien entendu, les résultats seront améliorés si le nombre de personnes appris augmente et les faux exemples appris durant l'apprentissage diminuent comme pour la reconnaissance des expressions faciales. La figure 9-10 montre les activités neuronales de la reconnaissance des différentes orientations. Nous remarquons que la catégorisation se fait assez facilement (figure 9-10a montre une activité élevée pour l'orientation gagnante) mettant en évidence le peu d'ambiguité entre les différentes orientations du regard.

Human's orientation \ Robot's orientation	left	middle	right
left	**53%**	23%	24%
middle	12%	**49%**	39%
right	10%	24%	**66%**

Fig. 9-9 – Test de généralisation. Le tableau montre le taux de confusion de l'orientation du regard sur des personnes non apprises par le robot. 14 personnes constituent la base de donnée. Ces statiques sont obtenues grâce à la méthode de cross validation car nous disposons d'un nombre de personnes réduites. On retire à chaque fois une personne de la base d'apprentissage pour pouvoir la tester en généralisation.

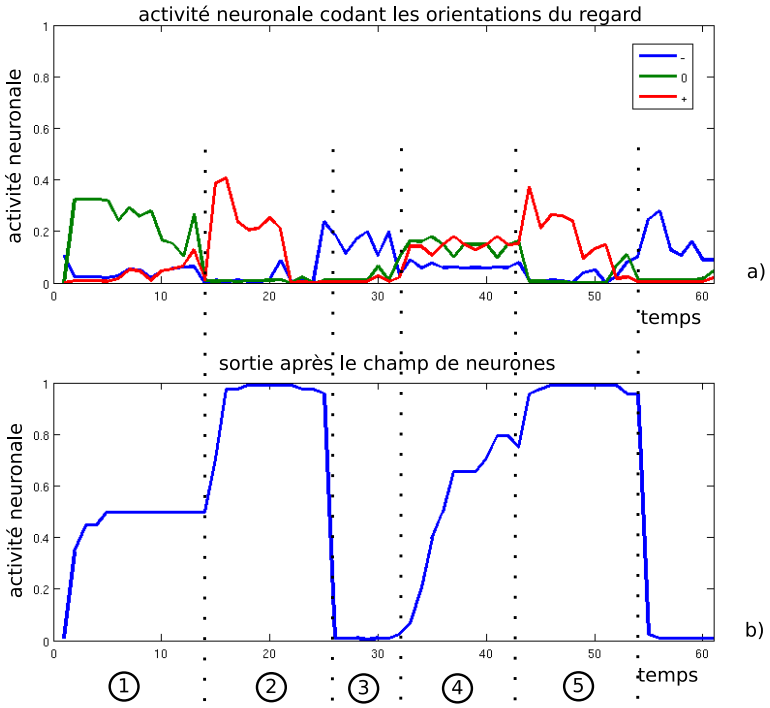


Fig. 9-10 – Orientation du regard durant la phase de reproduction de la tête expressive. Ces courbes montrent l'activité analogique de plusieurs neurones qui sont impliqués dans l'orientation du regard. a) montre l'activité neuronale des 3 neurones codant chacun pour une direction (à droite, à gauche et en face). b) montre l'effet du neural field pour la reconnaissance de l'orientation du regard (interpolation pour une reconnaissance plus précise grâce au champ de neurones).

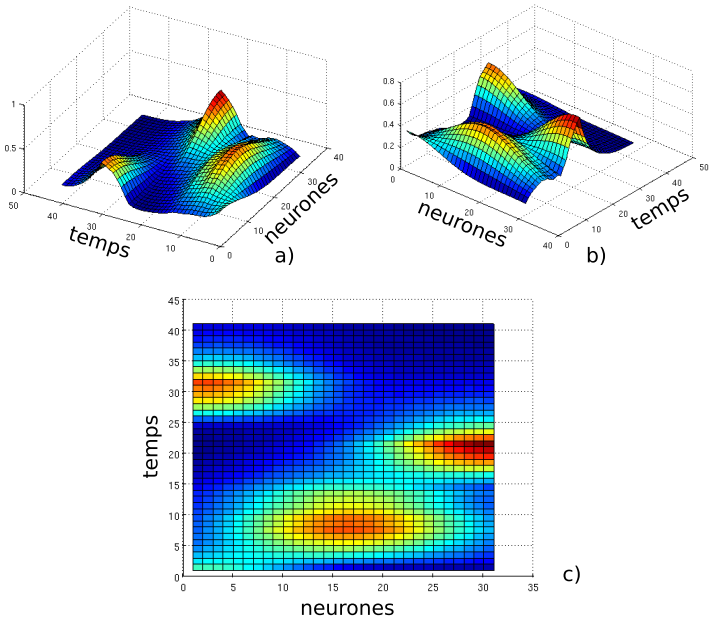


Fig. 9-11 – Activité temporelle du champ de neurone. Les courbes a) b) c) montrent l'activité temporelle du champ de neurones lors d'une séquence d'orientation de regard.

Dans cette figure, l'expérimentateur a dirigé son regard à plusieurs endroit. Les phases 1, 2, 3, 5 montrent que le robot imite correctement les positions extrêmes ainsi que la position centrale (regard à droite, à gauche et en face). Mais le système possède à une certaine finesse améliorant l'orientation du regard par l'intermédiare du champ de neurones (figure 9-10b). Nous sommes capables comme pour les primitives motrices d'avoir une notion "d'intensité", c'est à dire si l'expérimentateur dirige son regard légérement vers la droite alors le robot est capable d'imiter ce comportement. Cette capacité est illustrée par la phase 4 : l'expérimentateur regarde vers la droite sans être à la position extrême. Par conséquent, le robot dirige son regard entre deux positions. Les trois neurones codant pour les positions regard à gauche, à droite et en face sont suffissants pour une orientation du regard grossière mais l'introduction du champ de neurones ajoute la possibilité d'obtenir des positions intermédiaires.

La figure 9-11 illustre le comportement du champ de neurones dans le temps. Des bulles d'activités sont créées, permettant de prendre des décisions plus fines. Les propriétés du neural field sont intéressantes pour justement fusionner des informations et également de prendre en considération des informations appartenant au passé impliquant une stabilité dans le temps.

9.2.4 Conclusion

Nous avons montré dans cette section qu'une architecture sensori-motrice peut apprendre des orientations de regard par l'intermédiaire de jeu d'imitation. Le paradigme expérimental est identique à la reconnaissance des expressions faciales et la reconnaissance des groupements musculaires. De plus, un temps d'interaction très court avec l'expérimentateur est suffisant pour que le robot acquière la capacité à "lire" la direction du regard de son partenaire. En effet, 1 minute d'apprentissage par individu est nécessaire pour que le robot soit capable d'orienter son regard là où l'individu regarde. Généralement l'expérimentateur oriente son regard sur des zones de l'environnement qui l'intéresse ou qu'il a envie de faire découvrir au robot. Autrement

dit, l'expérimentateur peut orienter son regard en direction d'un objet pour signifier au robot l'intêret qu'il porte à ce dernier. Un sourire ou une grimace de l'expérimentateur donne un sens à l'objet.

9.3 Vers une attention conjointe

La capacité d'orienter le regard du robot en fonction du regard de l'expérimentateur est maintenant disponible. Cette capacité prend un sens réel dans un contexte d'interaction triadique (homme-robot-objet). Nous faisons l'hypothèse qu'un seul objet est présent dans l'espace péri-corporel du robot. L'expérimentateur peut regarder un objet lui procurant du plaisir en affichant un sourire, dans le même temps le robot oriente son regard sur ce même objet et comprend grâce à l'expression faciale du partenaire que cette objet est intéressant. Le contraire est également possible en exprimant une expression négatif en présence d'un objet dangereux où menaçant.

9.3.1 Suivi d'un objet

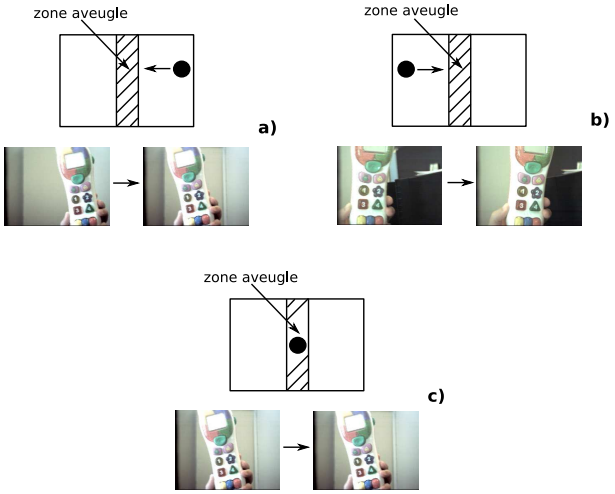


Fig. 9-12 – Illustration du centrage d'un objet. a) et b) montre un objet respectivement à droite et à gauche de l'image puis le centrage de celui-ci. c) montre un objet dans la zone aveugle impliquant la stabilité du centrage

Une fois la zone à suivre est détectée, une manière simple de réaliser le contrôle occulaire est d'utiliser un modèle proche du véhicule de Braitenenberg (Braintenberg, 1984). La solution est de trouver l'état d'équilibre (homéostat) permettant de mettre l'objet au centre de l'image. Ainsi si l'objet est à droite ou à gauche dans le repère image, le contrôle du mouvement de l'oeil essayera de le placer au centre. Le suivie d'objet est une propriété émergente du système puisque ce modèle a pour but d'amener l'objet au centre de l'image (zone aveugle). La dynamique du système simule un suivi de cible par le simple fait qu'à chaque itération l'objet est attiré au centre. La figure 9-12 illustre cette aspect de suivi d'objets.

Le modèle décrit figure 9-13 montre la possibilité de réaliser un suivi d'objet de manière neuronale. Trois étapes sont indispensables:

– Une projection horizontale est nécessaire pour savoir dans quelle zone de l'image l'objet se trouve (figure 9-13a)

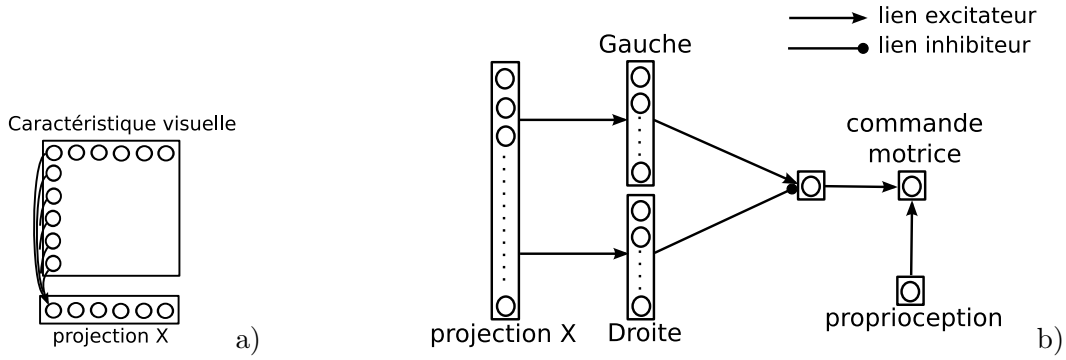


Fig. 9-13 – Architecture contrôlant le suivi d'une cible. a) montre la projection horizontale qui est réalisée grâce aux connexions synaptiques (projection neuronale). Les caractéristiques visuelles utilisées sont la couleur et les filtres de Gabor. b) montre l'architecture contrôlant le centrage d'une cible. La commande motrice correspond au degré de liberté contrôlant l'orientation du regard du robot (angle moteur correspondant au *Pan*).

- Une différence entre la projection horizontale contenant la partie gauche et celle contenant la partie droite est calculée pour obtenir le sens du vecteur direction. L'oeil du robot bouge à droite ou à gauche suivant le signe de cette différence.
- Lorsque la cible se trouve dans la zone aveugle (la différence calculée est nulle) impliquant ainsi la stabilité du système. Il est important de noter que la zone aveugle est l'intersection entre la projection horizontale contenant la partie gauche et celle contenant la partie droite.

Le modéle décrit permet au robot de suivre un objet complexe dans son environnement. La solution est triviale puisqu'on pense à stabiliser le système dans une position d'équilibre. Le comportement du robot peut être vu comme la recherche d'un état homéostatique. La figure 9-14 illustre le suivi d'un objet par la tête robotique, l'objet bouge de la gauche vers la droite et les yeux du robot suivent constamment l'objet afin de le mettre au centre de l'image. Nous pouvons observer que lorque l'objet ne bouge plus, la commande motrice reste stable.

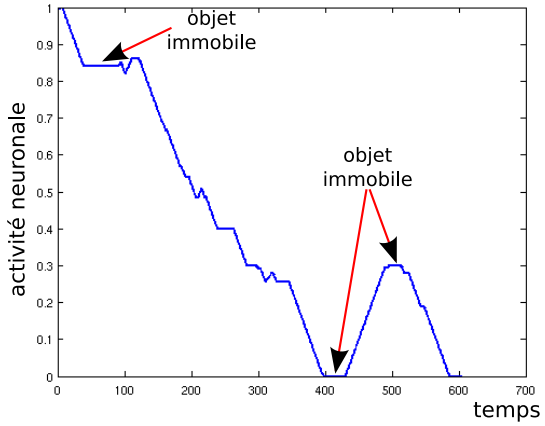


Fig. 9-14 – Activité neuronale correspondant à la position angulaire de l'oeil du robot lors d'un suivie de cible. La cible bouge de la gauche vers la droite. L'activité égale à 1 correspond au cas où la cible est à gauche et l'activité égale à 0 correspond au cas où la cible est à droite.

9.3.2 L'interaction émotionnelle donne un sens à l'objet

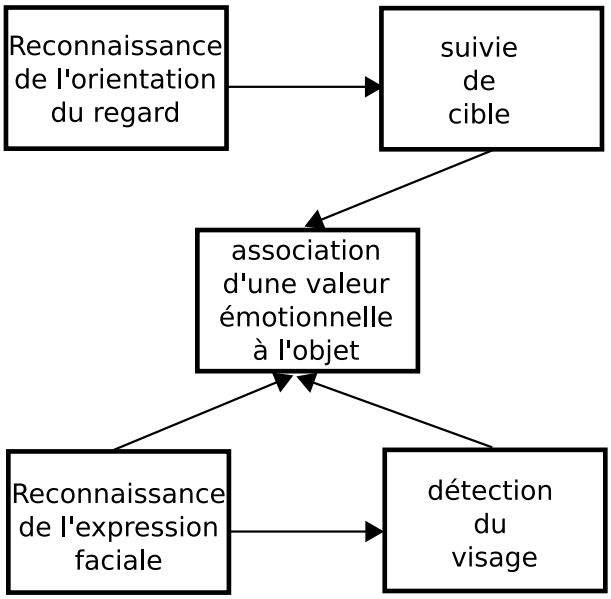


Fig. 9-15 – Schèma bloc: prémices d'attention conjointe dans un contexte de référencement social. La reconnaissance des expressions faciales, l'attribution de valeur émotionnelle à des objets et la reconnaissance de la direction du regard permet de donner un sens émotionnel à l'objet lors d'une interaction avec un expérimentateur.

Jusqu'à présent, nous avons décrit deux architectures complétement déconnectées. L'une permet d'orienter le regard du robot selon celui de l'expérimentateur et l'autre permet de suivre un objet qui bouge dans l'environnement. L'intégration de ces deux modules permet au robot de focaliser son attention sur l'objet que l'expérimentateur regarde. En d'autres termes, il suffit simplement d'utiliser l'architecture correspondant à la reconnaissance de la direction de regard afin d'orienter le champ de vision de la tête robotique d'un côté ou de l'autre de la scène visuelle. Si un objet est détecté alors le suivi d'objet est activé, permettant de mettre l'objet au centre de l'image. Cette dynamique est vue comme un prémices d'attention conjointe puisque l'objet cible est la source d'attention des deux agents. L'objet cible capte l'attention du premier agent qui par conséquent attire le regard du second agent sur cet objet.

Une fois que l'attention du robot est orientée sur l'objet, l'expérimentateur peut exprimer son ressenti vis à vis de l'objet (Figure 9-15). L'expression faciale du partenaire humain est utilisée pour évaluer l'intérêt de l'objet. Le modèle présenté (figure 9-16) montre une cascade d'architectures sensori-motrices dont la finalité est de focaliser sur un objet que l'expérimentateur regarde tout en essayant de lui attribuer une dimension émotionnelle.

Le modèle ainsi développé met en évidence la capacité du robot à pouvoir orienter son regard sur un événement (objet) et de reconnaître l'expression de l'expérimentateur dans le but de communiquer une information. Ce modèle montre que les informations émotionnelles transmises par l'humain permettent au robot d'apprendre des associations entre des caractéristiques visuelles appartenant à l'objet et un état émotionnel venant de l'environnement social. En présence de l'expérimentateur, un prémices d'attention conjointe est disponible permettant aux deux agents d'orienter leurs regards sur un même objet de l'environnement. Au départ, l'objet n'a pas de

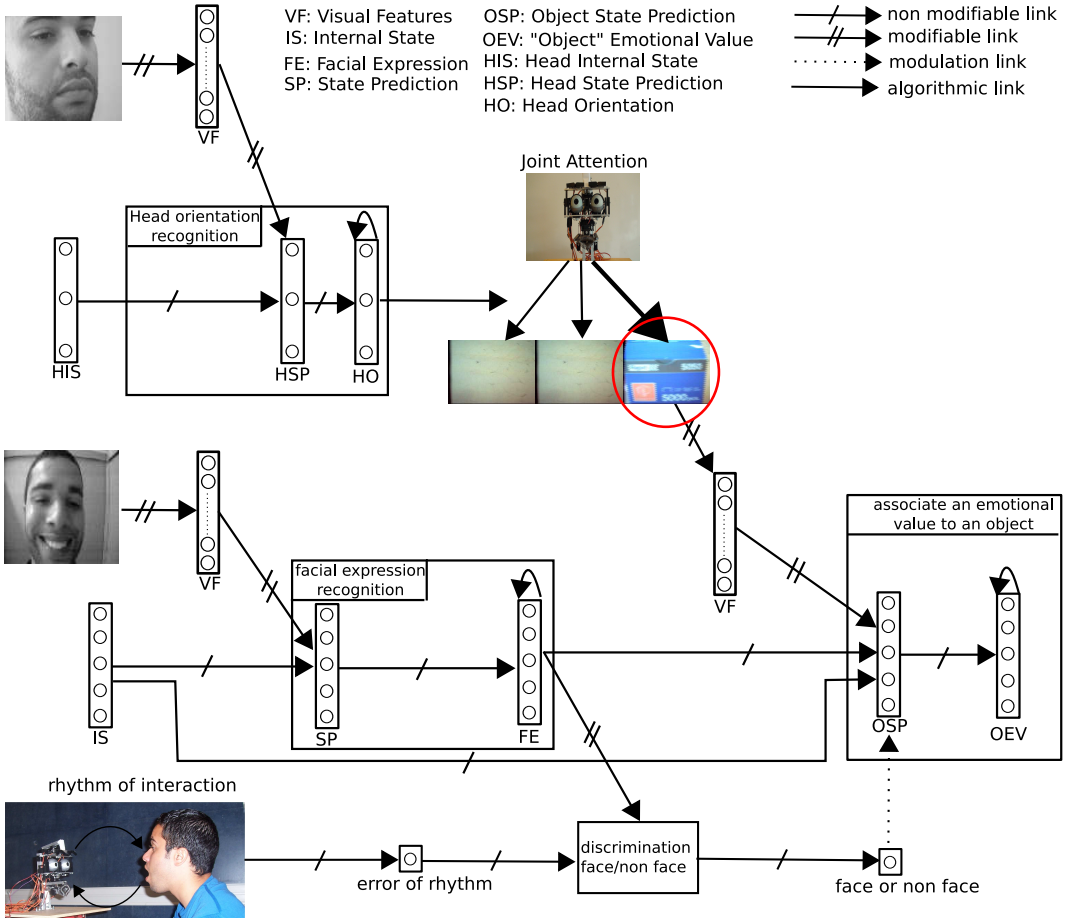


Fig. 9-16 – Modèle intégrant un prémices d'attention conjointe dans un contexte de référencement social. Les interactions sensori-motrices entre la reconnaissance des expressions faciales, l'attribution de valeur à des objets et la reconnaissance de la direction du regard permet d'une part d'orienter l'attention du robot sur un objet que regarde l'expérimentateur et d'autre part donner un sens émotionnel à l'objet. Le robot dirige son regard sur un objet en suivant celui de l'expérimentateur. Une fois que l'objet est dans le champ de vision du robot, l'expérimentateur peut communiquer de l'information grâce à l'expressivité faciale. Par conséquent, un objet peut avoir une valeur émotionnelle positive ou négative selon l'expressivité de l'expérimentateur.

signification pour le robot, il ne posséde pas de valeur émotionnelle. L'expérimentateur peut décider de communiquer son propre ressenti envers l'objet. L'expérimentateur est un enseignant qui permet au robot d'acquérir des connaissances sur l'environnement. Le robot a la capacité d'apprendre que tel objet est positif tandis que tel autre est négatif grâce à un partenaire humain capable d'exprimer des états émotionnels. Ce modèle est basée sur des mécanismes perception-action, permettant l'association des sensations visuelles caractérisant l'objet avec un état émotionnel transmis par l'humain. A peine, ces associations créées, un objet pourra procurer au robot un état de "plaisir" si sa valeur émotionnelle est positive et au contraire un état de "déplaisir" si sa valeur émotionnelle est négative. L'expression faciale du robot est fonction de l'expression de l'expérimentateur et de la valeur émotionnelle de l'objet. En présence d'un objet positif, le robot affichera un sourire et au contraire lorsque l'objet est négatif, le robot exprimera de la colère.

9.3.3 Visualisation de la dynamique du système

L'analyse des résultats pour ce type d'architecture est généralement confronté à un manque de statistique. Nous avons fait le choix de quantifier les différentes briques élémentaires de notre architecture:

- les sections précédentes montrent le taux de reconnaissance de la direction du regard (figure 9-9 et 9-8), montrant ainsi la capacité du dispositif robotique à orienter son regard en fonction du regard de l'expérimentateur.
- le chapitre portant sur la reconnaissance des expressions faciales montre quand à lui, la faculté du robot à reconnaître l'expressivité d'un partenaire humain.
- le chapitre sur le référencement social dans le cadre d'une manipulation d'objet montre que la valeur émotionnelle d'objet peut être apprise par l'intermédiare de l'interaction émotionnelle. Ces résultats ont montré la capacité du modèle à être invariant aux changements d'échelle et à la position de l'objet dans l'image.

Il est difficile d'obtenir des résultats quantitatifs sur l'expérience globale. Cependant, il nous semble important de montrer des résultats qualitatifs. La figure 9-17 illustre le fonctionnement d'une séquence d'interactions entre l'humain et la tête expressive. Les différents signaux présentés, montrent la capacité du robot à apprendre de manière autonome la valeur émotionnelle d'un objet grâce à l'interaction. Cette figure montre plusieurs phases lors d'une interaction triadique entre robot-objet-humain. Une première séquence montre la réaction du robot lorsque l'humain regarde un objet se situant à sa gauche et affiche une expression faciale négative. Durant les phases 1 et 2, l'expérimentateur exprime un visage de colère vis à vis de l'objet. Au départ, l'objet n'est pas encore dans le champ de vision du robot. Le robot doit tourner ses yeux vers la gauche pour mettre l'objet dans son champ de vision. L'objet qui était neutre au debut de l'interaction, devient négatif au fur et à mesure que l'humain exprime une expression de colère. La phase 3 montre qu'en l'absence de l'expérimentateur, la valeur émotionnelle de l'objet est reconnue comme négative impliquant une expression de colère du robot. L'association entre les sensations visuelles et l'expression négative transmise par l'humain a été renforcée. La phase 4 et 5 montre que la valeur de l'objet est reconnue comme étant négative même si l'objet est déplacé vers la droite (c'est bien l'objet qui a été associé à une valeur émotionnelle et non le lieu). Le robot s'engage par conséquent dans un suivi de cible permettant de garder l'objet dans le centre de son champ de vision et les caractéristiques visuelles permettent de reconnaître la valeur émotionnelle de l'objet en mouvement. La phase 6 met en évidence la capacité du robot à réaliser des associations similaires lorsque l'expérimentateur exprime un visage de joie. Il transmet une valeur émotionnelle positive à un nouvel objet captivant son attention. D'un façon similaire, l'humain oriente son regard sur l'objet tout en réalisant une expression positive. Par conséquent, en imitant le sujet humain, le robot oriente son regard en direction de l'objet en lui attribuant une valeur positive grâce à l'expression faciale reconnue. La phase 7 présente la capacité du système à ne pas oublier ce qu'il a appris. Le premier objet est remis sur la scène. L'expérimentateur oriente son regard dans sa direction, sans exprimer d'expression faciale. Une fois, l'objet dans le champ de vision du robot, le robot reconnait l'objet comme négative et exprime une expression de colère. Cela met en évidence la robustesse de l'apprentissage dans l'espace et dans le temps.

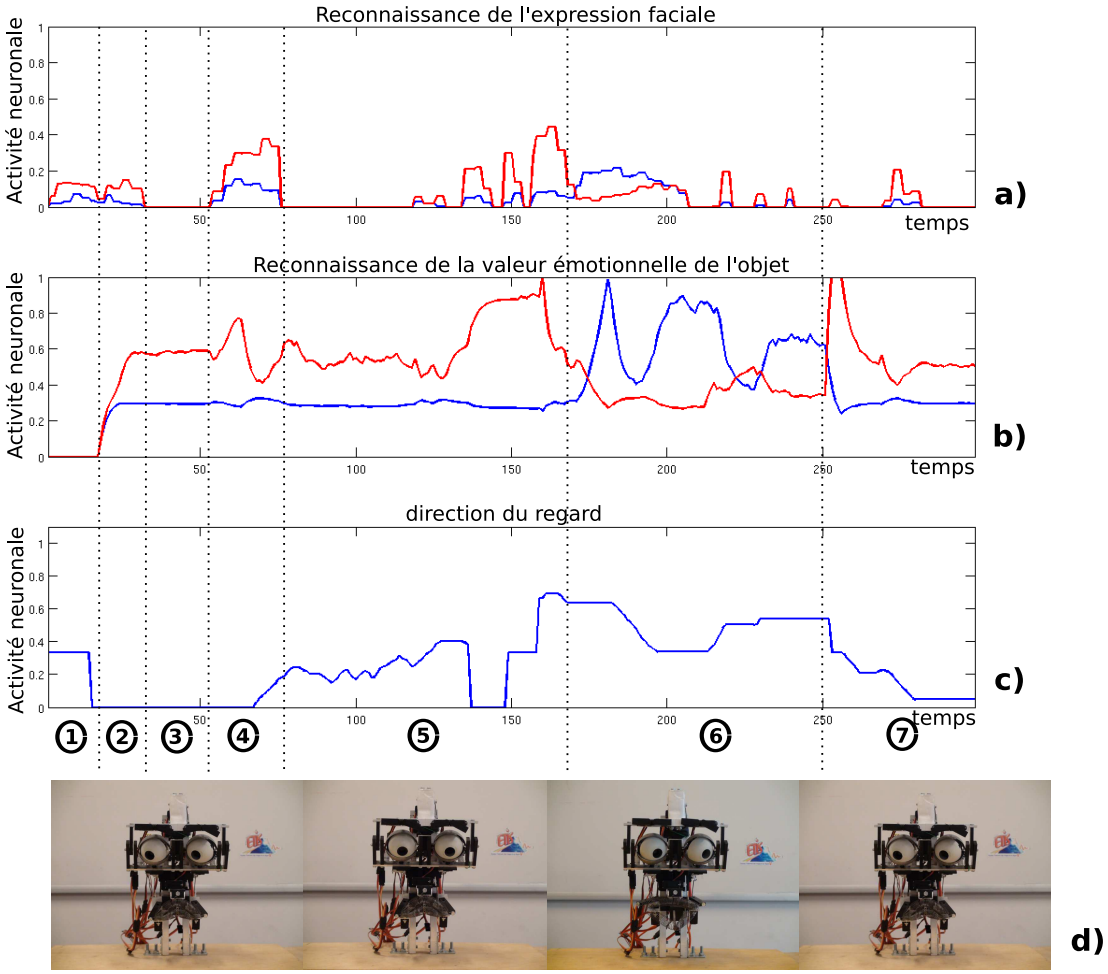


Fig. 9-17 – Résultat d'un référencement social dans un cadre d'attention conjointe. Les différents signaux montrent l'influence de l'interaction sociale pour orienter le regard du robot vers un objet et lui attribuer une valeur émotionnelle. a) montre la reconnaissance de l'expression faciale. b) montre la reconnaissance de la valeur émotionnelle de l'objet. c) montre la reconnaissance de la direction du regard (l'activité neuronale correspond à l'angle de rotation de l'oeil du robot). d) illustre la reconnaissance de l'orientation du regard ainsi que l'expressivité reconnue. Les courbes en rouge correspondent à la reconnaissance de la colère tandis que les courbes en bleu correspondent à la reconnaissance de la joie.

9.4 Conclusion

Ce chapitre insiste sur plusieurs aspects mettant en avant les principes de bases pour le développement d'un robot dans son environnement physique et social. Les premiers résultats montrent que la tête robotique est capable d'apprendre à diriger son regard dans la même direction que l'expérimentateur. Une architecture PerAc est capable de résoudre ce type de tâche si dans un premier temps l'humain regarde dans la même direction que le robot, sa propre action est corrélée avec ses sensations visuelles. Un prémices d'attention conjointe peut être amorcé dans le cas où un objet devient la source d'attention des deux agents. Cependant, dans le cadre expérimental que nous proposons, le robot est capable d'orienter son regard en fonction d'autrui sur un objet dans le but d'y attribuer une valeur émotionnelle. Dans ce cas

précis, avons nous introduit un amorçage d'attention partagée comme définit par Emery (Emery, 2000) dans l'introduction de ce chapitre? On se retrouve dans une situation du type: je sais que le robot regarde l'objet et le robot "sais" que je regarde l'objet. Dans cette expérience, l'interaction émotionnelle permet de communiquer des états émotionnels afin de donner un sens à l'objet. Une perspective qui semble une piste de recherche intéressante est sans doute le rôle de l'attention conjointe dans le développement du langage (Kaye, 1976; Bruner, 1987). Il semble évident que notre cadre expérimental est suffisamment souple pour obtenir ce genre de capacité cognitive. Dans notre cas précis, le robot possède un vocabulaire réduit à "bien" ou "pas bien" correspondant respectivement à l'expression positive et négative. Cependant, l'acquisition d'un vocabulaire plus riche semble réalisable. Tomasello (Tomasello and Todd, 1983) montre qu'il existe un lien très fort entre l'attention conjointe et la richesse du vocabulaire à 18 mois. Nous espérons que notre approche pourra se généraliser pour développer un vocabulaire plus riche (voir les travaux sur les "talking heads" (Kaplan, 2000)).

Chapitre 10

Conclusions et persectives

En focalisant cette thèse sur les interactions entre un expérimentateur humain et un dispositif robotique, nous avons proposé une nouvelle manière de comprendre comment des tâches simples comme la reconnaissance des expressions faciales, la détection du visage ou encore la reconnaissance de l'orientation du regard, peuvent être apprises de manière autonome.

Nous avons montré qu'une architecture sensori-motrice se prête bien à la résolution de ce type de tâche. Une seule condition est nécessaire pour le développement de ces compétences, le robot doit être initiateur dans l'interaction. En d'autres termes, le robot doit produire des actions que l'humain doit imiter. La cohérence entre l'imitant et l'imité permet l'acquisition de ces comportements, émergeant de cette interaction diadique. Notre architecture sensori-motrice étant consituée de mécanismes neuronaux associatifs, le robot associe ses sensations visuelles avec ses propres actions. Dans une première série d'expériences (chapitre 4, chapitre 5 et chapitre 6), le robot produit des actions que l'humain imite, son champ de vision contenant principalement le visage de l'expérimentateur, il est capable d'associer ce qu'il fait à ce qu'il voit. Ce jeu d'imitation permet de montrer qu'un robot peut développer des compétences cognitives en un temps très court (2 minutes) et de questionner certains modèles psychologiques (Meltzoff and Moore, 1977; Meltzoff and Moore, 1983). Dans une deuxième série d'expériences, on passe d'une interaction diadique à une interaction triadique avec l'introduction d'un objet. Ces expériences montrent l'apprentissage de capacités plus complexes comme le référencement social.

Les intêrets de notre approche se situent à différents niveaux conceptuels. Tout d'abord, les modèles développés doivent être à la fois robustes et souples vis à vis de l'environnement contrairement à certaines approches de la robotique classique qui permettent une robustesse indéniable mais sous certaines contraintes d'utilisation. Dans un cadre général, nos modèles s'adaptent à l'environnement au travers de l'apprentissage. La reconnaissance de l'expression faciale est une parfaite illustration. A défaut de présenter un visage expressif devant le robot, on présente un poing fermé pour la colère et la main ouverte pour la joie, le robot apprend à associer ces différentes vues aux états émotionnels correspondants de la même manière qu'il le faisait pour les visages. Les a apriori du modèle sont suffisamment faibles pour le permettre contrairement aux systèmes dédiés à la reconnaissance des expressions faciales.

Une problèmatique importante de l'intelligence artificielle est: comment de nouveaux comportements peuvent émerger au travers de l'interaction? Un premier élément de réponse est fourni dans le chapitre 6. Notre modèle neuronal permet l'émergence de comportements qui ne sont pas triviaux. Par exemple la reconnaissance des primitives motrices qui permet la production d'expressions faciales primaires et secondaires et donc de se dégager des précatégories données initialement au système (i.e les émotions de base selon Ekman).

D'autre part, les modèles décrits dans cette thèse cherchent toujours à être minimaux dans le but de montrer quels sont les mécanismes élémentaires permettant le fonctionnement d'une compétence particulière. Nous avons montré au travers des expérimentations que la détection du visage n'était pas indispensable pour la reconnaissance des expressions faciales. Cependant, l'interaction émotionnelle et plus particulièrement l'émergence de la rythmicité entre l'expérimentateur humain et la tête expressive, permet la détection de visage. Ce modèle montre une nouvelle manière d'aborder cette problèmatique.

Enfin, notre approche nourrit la réflexion vis à vis des modèles proposés en psychologie. (Meltzoff and Moore, 1977; Meltzoff and Moore, 1983) montre que des nourissons sont capables très tôt de reproduire les expressions du visage de leurs parents. Néanmoins pour lui, cette capacité est plus innée qu'apprise. Nos développements montrent que cette compétence peut être apprise très rapidement chez un robot (2 à 3 minutes).

Le défi de notre approche est d'utiliser des architectures neuronales afin d'être robuste aux variations de l'environnement. Nos modèles ont toujours le souci d'être biologiquement plausibles et de permettre l'émergence de comportements constatés chez l'humain. Cependant, généralement les résultats se focalisent sur le comportement souhaité. Les courbes, les pourcentages de réussite, les taux de confusions argumentent en faveur de la compétence implémentée. Ma position serait plus nuancée dans le cas où le modèle doit correspondre à une caractéristique comportementale humaine. Il serait judicieux d'introduire dans le modèle des lésions, montrant le comportement du robot vis à vis de ces pertubations. Ces lésions pourraient "expliquer" ou "être corrélées" avec certains syndrômes et par conséquent justifier davantage la plausibilité du modèle développé. Le chapitre 6 essaie de mettre cette idée en avant.

Dans la seconde partie de cette thèse centrée sur les interactions triadiques, le but était d'obtenir un robot capable de communiquer avec l'expérimentateur afin de découvrir son environnement. Les expériences réalisées se placent dans une période développementale pré-verbale mais supposent que la reconnaissance des expressions faciales émotionnelles est déjà acquise. Les différents modèles sensori-moteurs proposés pour réaliser un référencement social dans différents contextes mettent en évidence plusieurs aspects. Le premier point important à souligner est l'importance des signaux émotionnels. Plus particulièrement, les expressions faciales semblent être un excellent moyen de communication permettant à la fois d'attribuer des valeurs émotionnelles à des objets ou des lieux et d'adapter le comportement du robot. L'émotion est un moyen de communication rapide et efficace. Le second point important est la généricité du modèle. L'architecture perception-action posséde certaines propriétés permettant de résoudre certaines capacités du référencement social. Une cascade d'architectures sensori-motrices non dédiées aux interactions complexes peut résoudre des tâches de cognition sociale en utilisant des mécanismes de bas niveau. Les différentes expériences montrent que le référencement social peut s'expliquer d'abord à un niveau sensori-moteur sans faire appel à une cognition de haut niveau et en favorisant davantage la dimension émotionnelle. Notre approche fournit des hypothèses intéressantes sur le développement du référencement social dans le cadre de dynamiques sensori-motrices.

Notre travail montre que le robot peut acquérir différentes compétences au travers des interactions émotionnelles:

- le bras robotisé a la capacité d'attraper ou de fuir des objets selon l'expressivité de l'expérimentateur humain
- la base mobile du robot posséde la capacité de rejoindre ou d'éviter des lieux de l'environnement
- la tête expressive oriente son regard en fonction du regard de l'expérimentateur tout en

assignant une valeur émotionnelle aux objets selon l'expression du partenaire humain

Ces différents travaux montrent l'importance de l'émotion dans l'apprentissage de comportements sociaux ainsi que la capacité des mécanismes de conditionnement simple à associer l'expressivité de l'expérimentateur avec des objets, des lieux ou des comportements afin d'adapter le comportement du robot et ainsi permettre au travers de l'interaction (passage d'interactions diadiques à triadiques) d'apprendre des comportements sans cesse plus complexes.

Dans de futurs travaux, il sera intéressant d'étudier l'ensemble des tâches qui sont accessibles par les seuls principes développés dans cette thèse pour mieux comprendre l'intêret des structures cérébrales qui ont été volontairement négligées dans cette thèse. Par exemple au chapitre 4, la reconnaissance des expressions faciales est basée exclusivement sur une reconnaissance locale de l'image (points de focalisation) correspond à une analyse précise. Cette stratégie peut être vue comme une implémentation simple de la voie thalamo-cortico-amygdala chez les mammifères (LeDoux, 1996). Dans des travaux précédents (Gaussier et al., 2007), une architecture simple et rapide a été implémentée analysant l'image globale. Elle pourrait correspondre à une voie de traitement courte, la voie thalamo-amygdala (Papez, 1937; LeDoux, 1996) impliquée dans les réactions émotionnelles rapides. De notre point de vue, il serait intéressant de vérifier le couplage de ces deux approches afin de mettre en évidence leurs influences. Le réseau thalamo-cortico-amygdala pourrait être une manière de contrôler l'apprentissage du réseau thalamo-amygdala permettant à la fois une reconnaissance rapide et robustesse des expressions faciales.

De plus, il serait intéressant de construire un modèle du référencement social intégrant des capacités de navigation, d'attention conjointe et de préhension d'objets. Ce type d'intégration n'est pas trivial pour de nombreuses raisons :

- mise en place du set-up expérimental à cause du nombre de robot à intégrer (une solution efficace serait d'utiliser un robot humanoïde)
- changement de référentiel: quels flux informations le robot doit-il utiliser? Le robot doit focaliser sur l'humain, l'objet ou le lieu
- sélection de l'action: quel doit être le comportement du robot afin de montrer son adapatation à l'environnement

Enfin, une perspective qui semble intéressante est sans doute le rôle de l'attention conjointe dans le développement du langage (Kaye, 1976; Bruner, 1987). Notre cadre expérimental est suffisamment souple pour obtenir ce genre de capacité cognitive. Dans notre cas précis, le robot possède un vocabulaire réduit à "bien" ou "pas bien" correspondant respectivement à l'expression positive et négative. Cependant, l'acquisition d'un vocabulaire plus riche semble réalisable à travers ce type d'interaction.

Bibliographie

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87.

Andry, P., Gaussier, P., Moga, S., Banquet, J., and Nadel, J. (2001). Learning and communication in imitation: An autonomous robot perspective. *IEEE transactions on Systems, Man and Cybernetics, Part A*, 31(5):431–444.

Andry, P., P.Gaussier, and Nadel, J. (2002). From sensorimotor coordination to low level imitation. In *Second international workshop on epigenetic robotics*, pages 7–15.

Arnold, M. (1960). Emotion and personality. *Columbia University Press, New Yorkbr*.

B. Abboud, F. Davoine, M. D. (2004). Facial expression recognition and synthesis based on an appearance model. *Signal Processing: Image Communication*, 19:723–740.

Baldwin, D. and L.J., M. (1996). The ontogeny of social information gathering. *Child Development*, 67:1915–1939.

Banquet, J., Gaussier, P., Dreher, J. C., Joulain, C., Revel, A., and Günther, W. (1997). Space-time, order, and hierarchy in fronto-hippocampal system: A neural basis of personality. In Matthews, G., editor, *Cognitive Science Perspectives on Personality and Emotion*, volume 124, pages 123–189, Amsterdam. North Holland.

Bard, K., Leavens, D., Custance, D., Vancatova, M., Keller, H., Benga, O., and Sousa, C. (2005). Emotion cognition: Comparative perspectives on the social cognition of emotion. *Cognitie, Creier,Comportament (Cognition, Brain, Behavior), Special Issue: "Typical and atypical development"*, 8:351–362.

Bard, P. (1928). A diencephalic mechanism for the expression of rage with special reference to the central nervous system. *American journal of psychology*, 84:490–513.

Baron Cohen, S. (1995). Mindblindness. *An essay on autism and theory of mind, MIT Press, Cambridge*.

Barto, A., Sutton, R., and Brouwer, P. (1981). Associative search network: A reinforcement learning associative. *Biological Cybernetics*, 40:201–211.

Bechara, A., Tranel, D., Damasio, H., Adolphs, R., Rockland, C., and Damasio, R. (1995). Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. 269(5227):1115–1118.

Belaroussi, R., Prevost, L., and Milgram, M. (2005). Classifier combination for face localization in color images. *ICIAP (International Conference on Image Analysis and Processing), Lecture Notes in Computer Sciences*, pages 1043–1050.

Bicho, E. and Schoner, G. (1997). The dynamic approach to autonomous robotics demonstrated on a low-level vehicle platform. *Robotics and Autonomous Systems*, 21:23–35.

Billard, A., Calinon, S., Dillmann, R., and Schaal, S. (2008). Survey: Robot programming by demonstration. In *Handbook of Robotics*, volume chapter 59. MIT Press.

Braintenberg, V. (1984). Vehicles, experiments in synthetic psychology. In *Bradfords books*. Cambridge, MIT Press.

Breazeal, C., Buchsbaum, D., Gray, J., Gatenby, D., and Blumberg, B. (2005). Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots. *Artificial Life*, 11(1-2):31–62.

Bretherton, I., Fritz, J., Zahn-Waxler, C., and Ridgeway, D. (1986). Learning to talk about emotions : A functionalist perspective. *Child Development*, 57:529–548.

Brooks, R. (1986). A robust layered control system for a mobile robot. *1EEE Journal of Robotics and Automation*, 2:14–23.

Brown, D., Craw, I., and Lewthwaite, J. (2001). A som based approach to skin detection with application in real time systems. In *in Proc. of the British Machine Vision Conference*.

Brun, P. (1998). La compréhension des émotions: Son développement chez le jeune enfant et chez l'enfant autiste. *Paris*.

Brun, P. and Nadel, J. (1997). Les jeunes enfants face aux émotions. *Revue de Psychologie de l'education*, 2:75–86.

Bruner, J. (1983). Le développement de l'enfant. *Savoir faire, Savoir dire, Paris, Presses Universitaires de France*.

Bruner, J. (1987). Comment les enfants apprennent à parler. *Paris, Retz*.

Bullock, D., Fiala, J., and Grossberg, S. (1994). A neural model of timed response learning in the cerebellum. *Neural Networks*, 7:1101–14.

Bullock, M. and Russell, J. A. (1984). Pre-school children's interpretation of facial expressions of emotion. *International Journal of Behavioral Development*, 1:193–214.

Bullock, M. and Russell, J. A. (1986). Concepts of emotion in developmental psychology. *in C. E. Izard and P. B. Read (Eds.), Measuring emotions in infants and children. Cambridge, England: Cambridge University Press*, 2:203–237.

Butterworth, G. (1991). The ontogeny of joint visual attention. *Natural Theories of Mind, in Whiten, A.*, pages 223–232.

c. Huang, Ai, H., Wu, B., and Lao, S. (2004). Boosting nested cascade detector for multi-view face detection. *International Conference on Pattern Recognition*, pages 415–418.

Calinon, S., Guenter, F., and Billard, A. (2007). On learning, representing and generalizing a task in a humanoid robot. *IEEE transactions on systems, man and cybernetics, Part B. Special issue on robot learning by observation, demonstration and imitation*, 37(2):286–298.

Campos, J. J. (1983). The importance of affective communication in social referencing: A commentary on feinman. *Merill-Palmer Quarterly*, 29(1):83–87.

Campos, J. J., Campos, R. G., and Caplovitz-Barret, K. (1989). Emergent themes in the study of emotional development and emotional regulation. *Developmental Psychology*, 25(3):394–402.

Campos, J. J. and Steinberg, C. (1981). Perception, appraisal and emotion : The onset of social referencing. *in M. E. Lamb and L. R. Sherrod (Eds.), Infant social cognition Empirical and theoretical considerations. Hillsdale, N. J.: Erlbaum*.

Cannon, W. (1927). The james-lange theory of emotions: A critical examination and an alternative theory. *American journal of psychology*, 39:106–124.

Cardinal, R. N., Parkinson, J. A., Hall, J., and Everitt, B. J. (2002). Emotion and motivation: The role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience and Biobehavioral Reviews*, 26(3):321–352.

Carpenter, G., Grossberg, S., , and Rosen, D. (1991a). Art2-a : an adaptive resonance algorithm for rapid category leaning and recognition. *Neural Networks*, 4:493–504.

Carpenter, G. and Grossberg, S. (1987). Invariant pattern recognition and recall by an attentive self-organizing art architecture in a nonstationary world. *Proceeding of Neural Network*, 2:737–745.

Carpenter, G., Grossberg, S., Markuzon, N., Reynolds, J. H., and Rosen, D. (1992). Fuzzy artmap: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Trans. Neural Networks*, 3(5):698–712.

Carpenter, G., Grossberg, S., and Reynolds, J. (1991b). Artmap: A self-organizing neural network architecture for fast supervised learning and pattern recognition. *In Proc. Int. Joint Conf. on Neural Networks*, 1:863–868.

Chai, D. and Ngan, K. (1999). Face segmentation using skin-color map in videophone applications. *IEEE Trans. Circuits Syst. Video Technol*, 9:551.

Chevallier, S. and Tarroux, P. (2008). Covert attention with a spiking neural network. In *International conference on computer vision systems*, volume 5008 of *Lecture notes in computer science*, pages 56–65. Springer.

Conein, B. (1998). Les sens sociaux : coordination de l'attention et interaction sociale. *Intellectica*, 1-2,26-27:181–202.

Cosmides, L. and Tooby, J. (2000). Evolutionary psychology and the emotions. *Handbook of emotions*, pages 91–115.

Damasio, A. (1994). Descartes' error. *Putnam's Sons, New York*.

Darwin, C. (1965). The expression of emotion in man and animals. *Chicago:University of Chicago Press (Originally Published in 1872)*.

De Haan, M. and Nelson, C. A. (1998). Discrimination and categorisation of facial expressions of emotion during infancy. *In A. Slater (Ed.), Perceptual development : Visual, auditory and language perception in infancy.*, pages 287–309.

de Wied, M., van Boxtel, A., Zaalberg, R., Goudena, P., and Matthys, M. (2006). Facial emg responses to dynamic emotional facial expressions in boys with disruptive behavior disorders. *Journal of Psychiatric Research*, 40(2):112 – 121.

Descartes, R. (1649). Les passions de l'âme.

Devouche, E. and Gratier, M. (2001). Microanalyse du rythme dans les échanges vocaux et gestuels entre la mère et son bébé de 10 semaines. *Devenir*, 13:55–82.

Dickstein, S. and Parke, R. D. (1988). Social referencing in infancy: A glance at fathers and marriage. *Child Development*, 59:506–511.

Dimberg, U. (1982). Facial reactions to facial expressions. *Psychophysiology*, 19:643–647.

Ekman, P. (1982). Emotion in human face. *Cambridge University Press*.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6:169–200.

Ekman, P. and Friesen, W. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17:124–129.

Ekman, P. and Friesen, W. (1978). Facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists Press, Palo Alto, California*.

Ekman, P., Friesen, W., and Ellsworth, P. (1972). Emotion in the human face: Guide-lines for research and an integration of findings. *New York: Pergamon Press*.

Emery, N. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews*, 24:581–604.

Fantz, R. (1961). The origin of form perception. *Scientific American*, pages 66–72.

Farroni, T., Menon, E., Rigato, S., and Johnson, M. H.

Farrugia, N., Mamalet, F., Roux, S., Yang, F., and Paindavoine, M. (2009). Fast and robust face detection on a parallel optimized architecture implemented on fpga. *IEEE Trans. Cir. and Sys. for Video Technol.*, 19:597–602.

Fasel, B. and Luettin, J. (2000). Recognition of asymmetric facial action unit activities and intensities. *Pattern Recognition, International Conference on*, 1:1100–1103.

Feinman, S. (1982). Social referencing in infancy. *Merrill-Palmer Quarterly*, 28(4):445–470.

Feinman, S. and Lewis, M. (1983). Social referencing at ten months: A second-order effect on infants' responses to strangers. *Child Development*, 54:878–887.

Feinman, S., Roberts, D., Hsieh, K., Sawyer, D., and Swanson, D. (1992). A critical review of social referencing in infancy. *In S. Feinman (Ed.), Social referencing and the social construction of reality in infancy*, pages 15–54.

Feldman-Barrett, L. and Russell, J. (2009). Circumplex models of affect. *Oxford Companion to Emotion and Affective Sciences, Oxford University Press.*

Feraud, R., Bernier, O., Viallet, J., and Collobert, M. (2001). A fast and accurate face detector based on neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:42–53.

Field, T., Woodson, R., Greenberg, R., and Cohen, D. (1982). Discrimination and imitation of facial expression by neonates. *Science*, 218(4568):179–181.

Field, T. M. and Walden, T. A. (1982). Production and discrimination of facial expressions by preschool children. *Child Development*, 53:1299–1300.

Fontaine, J., Scherer, K., Roesch, E., and Ellsworth, P. (2007). The world of emotion is not two-dimensional. *Psychological Science*, 18:1050–1057.

Fraisse, P. (1963). Les emotions. *Traite de psychologie experimental*, 5:97–181.

Franco, L. and Treves, A. (2001). A neural network facial expression recognition system using unsupervised local processing. *2nd international symposium on image and signal processing and analysis. Cognitive neuroscience*, 2:628–632.

Frijda, N. (1986). The emotions. *Cambridge University Press.*

Frijda, N. (1987). Emotions, cognitive structures and action tendency. *Cognition and Emotion*, 1:115–143.

Fukuyori, I., Nakamura, Y., Matsumoto, Y., and Ishiguro, H. (2008). Flexible control mechanism for multi-dof robotic arm based on biological fluctuation. *From Animals to Animats 10*, 5040:22–31.

G. Gergely, J. W. (1999). Early socio-emotional development: contingency perception and the social-biofeedback model. *In P. Rochat, (Ed.), Early Social Cognition: Understanding Others in the First Months of Life*, pages 101–136.

Gallese, V. (2003). The roots of empathy: The shared manifold hypothesis and neural basis of intersubjectivity. *Psychopathology*, 36:171–180.

Gallet, O., Gaussier, P., and Cocquerez, J. (1998). A model of the visual attention to speed up image analysis. In *ICIP (1)*, pages 246–250.

Garcia, C. and Delakis, M. (2004). Convolutional face finder: A neural architecture for fast and robust face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:1408–1423.

Garcia, C. and Tziritas, G. (1999). Face detection using quantized skin color regions merging and wavelet packet analysis. *IEEE Trans. Multimedia*, 1:264–277.

Gaussier, P. (2001a). Toward a cognitive system algebra: A perception/action perspective. *In European Workshop on Learning Robots (EWRL).*, pages 88–100.

Gaussier, P. (2001b). Toward a cognitive system algebra: A perception/action perspective. *In European Workshop on Learning Robots (EWRL).*, pages 88–100.

Gaussier, P., Boucenna, S., and Nadel, J. (2007). Emotional interactions as a way to structure learning. *epirob*, pages 193–194.

Gaussier, P. and Cocquerez, J. (1991). Utilisation des réseaux de neurones pour la reconnaissance de scènes complexes: simulation d'un système visuel comprenant plusieurs aires corticales. *GRETSI, Saint Martin d'Hères, France*, 8:441–466.

Gaussier, P., Lepretre, S., Quoy, M., Revel, A., Joulain, C., and Banquet, J. (2000). Interdisciplinary approaches to robot learning, chapter experimentsandmodels about cognitive map learning for motivated navigation. *Robotics and Intelligent Systems Series World Scientific*.

Gaussier, P., Moga, S., Quoy, M., and Banquet, J. (1998). From perception-action loops to imitation processes: a bottom-up approach of learning by imitation. *Applied Artificial Intelligence*, 12(7-8):701–727.

Gaussier, P. and Zrehen, S. (1995). Perac: A neural architecture to control artificial animals. *Robotics and Autonomous Systems*, 16(2-4):291–320.

Gewirtz, J. and Pelaez-Nogueras, M. (1992). Social referencing as a learned process. *In S. Feinman (Ed.), Social referencing and the social construction of reality in infancy*, pages 151–173.

Ghamen, K. and Caplier, A. (2008). Estimation of anger, sadness and fear expression intensity based on the belief theory. In *Proceedings ACIT*.

Ghanem, K., Caplier, A., and Stillittano, S. (2008). Estimation of facial expression intensity based on the belief theory. In *VISAPP (1)*, pages 452–460.

Giovannangeli, C. and Gaussier, P. (2008). Autonomous vision-based navigation: Goal-oriented action planning by transient states prediction, cognitive map building, and sensory-motor learning. In *IROS*, pages 676–683.

Giovannangeli, C., Gaussier, P., and Banquet, J.-P. (2006). Robustness of visual place cells in dynamic indoor and outdoor environment. *International Journal of Advanced Robotic Systems*, 3(2):115–124.

Gosselin, P., Roberge, P., and Lavallée, M.-F. (1995). Le développement de la reconnaissance des expressions faciales émotionnelles du répertoire humain. *Enfance*, 4:379–396.

Gross, A. L. and Ballif, B. (1991). Children's understanding of emotion from facial expressions and situations: A review. *Developmental Review*, 11:368–398.

Grossberg, S. (1973). Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, pages 213–257.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, 11:23–63.

Grossberg, S. (1988). Nonlinear neural networks: principles, mechanisms, and architectures. *Neural Networks*, 1:17–61.

Grossberg, S. and Merrill, J. (1992). A neural network model of adaptively timed reinforcement learning and hippocampal dynamics. *Cognitive Brain Research*, 1:3–38.

Grossberg, S. and Mingolla, E. (1985). Neural dynamics of form perception: Boundary completion, illusory figures and neon color spreading. *Psych.Rev*, 92(2):173–211.

Grossberg, S. and Somers, D. (1991). Synchronized oscillations during cooperative feature linking in a cortical model of visual perception. *Neural Networks*, 4:453–466.

Gunnar, M. R. and Stone, C. (1984). The effects of positive maternal affect on infant responses to pleasant, ambiguous, and fear-provoking toys. *Child Development*, 55(4):1231–1236.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

Harris, P. L. (1989). Children and emotion. *Oxford, UK : Blackwell Publishers*.

Hasson, C. and Gaussier, P. (2010). Path integration working memory for multi task dead reckoning and visual navigation. In *Proceedings of Simulation of Adaptive Behavior (SAB'2010)*, pages 380–389.

Hebb, D. (1949). The organization of behavior: A neuropsychological theory. *LEA, Inc.*

Hecht-Nielsen, R. (1987). Counterpropagation networks. *Applied Optics*, 26(23):4979–4984.

Hersch, M., Calinon, S., and Billard, A. (2008). Dynamical system modulation for robot learning via kinesthetic demonstrations. *IEEE Transactions on Robotics*, 24:1463–1467.

Hess, U. and Blairy, S. (2001). Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy. *International Journal of Psychophysiology*, 40(2):129 – 141.

Hirshberg, L. M. and Svejda, M. (1990). When infants look to their parents: I. infants' social referencing of mothers compared to fathers. *Child Development*, 61:1175–1186.

Hodgkin, A. and Huxley, A. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol (Lond)*, 117:500–544.

Hornik, R., Risenhoover, N., and Gunnar, M. (1987). The effects of maternal positive, neutral, and negative affective communications on infant responses to new toys. *Child Development*, 58:937–944.

Hubel, D. and Wiesel, T. (1977). Ferrier lecture: Functional architecture of macaque monkey visual cortex. *Proc. Roy. Soc. Lond. B. Biol. Sci.*, 198:1–59.

Ikegami, T. (1993). Ecology of evolutionary game strategies. In *ECAL 93*, pages 527–536.

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

Izard, C. (1971). The face of emotion. *Appleton Century Crofts*.

Izard, C. (1979). The maximally discriminative facial movement coding system (max). *Newark DE: University of Delaware Instructional Resources Center*.

Izard, C. E. (1990). Facial expressions and the regulation of emotions. *Journal of Personality and Social Psychology*, 58(3):487–498.

Izard, C. E., Fantauzzo, C. A., Castle, J. M., Haynes, O. M., Rayias, M. F., and Putnam, P. H. (1995). The ontogeny and significance of infants' facial expressions in the first 9 months of life. *Developmental Psychology*, 31:997–1013.

J. Bastiaansen, M. Thioux, C. K. (2009). Evidence for mirror systems in emotions. *Phil. Trans. R. Soc. B*, 364:2391–2404.

James, W. (1884). What is an emotion. *Mind*, 9:188–205.

James, W. (1892). Psychology. *Henry Holt and Company, New York*.

Jasso, H., Triesch, J., and Deak, G. (2008). A reinforcement learning model of social referencing. *Development and Learning. ICDL*, pages 286–291.

Jennifer M. Jenkis, Keith Oatley, N. L. S. (1998). *Human Emotions*, chapter The communicative theory of emotions. Blackwell edition.

jier Lien, J. J. and Cohn, J. F. (1998). Subtly different facial expression recognition and expression intensity estimation. In *in Proc. Of the IEEE Int. Conference on Computer Vision and Pattern Recognition*, pages 853–859.

Joulain, C., Gaussier, P., and Revel, A. (1997). Learning to build categories from perception-action associations. In *International Conference on Intelligent Robots and Systems - IROS'97*, pages 857–864, Grenoble, France. IEEE/RSJ.

Kanade, T. (1973). Picture processing by computer complex and recognition of human faces. *PhD thesis, Kyoto Univ.*

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24:881–892.

Kaplan, F. (2000). L'émergence d'un lexique dans une population d'agents autonomes. *Thèse de l'université Paris VI.*

Kaplan, F. and Hafner, V. (2004). The challenges of joint attention. *Proceedings of the Fourth International Workshop on Epigenetic Robotics*, pages 67–74.

Kaye, K. (1976). Infants'effect upon their mothers'teaching strategies. *In J.C. Glidwell (ed.). The social context of learning and development, New York, Gardiner Press.*

Kestenbaum, R. and Nelson, C. A. (1990). The recognition and categorization of upright and inverted emotional expressions by 7-month-old infants. *Infant Behavior and Development*, 13(4):497 – 511.

Kim, J., Kee, S., and Kim, J. (2005). Fast detection of multiview face and eye based on cascaded classifier. *Machine Vision and Application*, pages 116–119.

Klinnert, M. (1984). The regulation of infant behaviour by maternal facial expression. *Infant Behavior and Development*, 7:447–465.

Klinnert, M., Campos, J., Sorce, J., Emde, R., and Svejda, M. (1983a). The development of the social referencing in infancy. *Emotion in early development*, 2:57–86.

Klinnert, M., Campos, J. J., Sorce, J., Emde, R. N., and Svejda, M. (1983b). Emotions as behaviour regulators: Social referencing in infancy. *in R. Plutchik and H. Kellerman (Eds.), Emotion : Theory, research, and experience. Emotions in early development*, 2:57–86.

Klinnert, M. D., Emde, R. N., Butterfield, P., and Campos, J. J. (1986). Social referencing: The infant's use of emotional signals from a friendly adult with mother present. *Developmental Psychology*, 22(4):427 – 432.

Knudsen and Konishi (1979). Mechanisms of sound localization in the barn owl (tyto alba). *Journal of Compartive Physiology*, 133:13–21.

Kohonen, T. (1989a). Self-organization and associative memory. *Heidelberg: Springer-Verlag, Berlin, 3rd edition.*

Kohonen, T. (1989b). Self-organization and associative memory. *Heidelberg: Springer-Verlag, Berlin, 3rd edition.*

Kotropoulos, C. and Pitas, I. (1997). Rule-based face detection in frontal views. In *in Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 97), vol. 4*, pages 2537–2540.

LaFreniere, P. J. (2000). Emotional development : A biosocial perspective. *Belmont, CA : Wadsworth.*

Lagarde, M., Andry, P., and Gaussier, P. (2008). Distributed real time neural networks in interactive complex systems. In *CSTST*, pages 95–100.

Lange, C. (1885). The emotions. *Williams and Wilkins*.

Lapicque, L. (1907). Recherches quantitatives sur l'excitation électrique des nerfs traitée comme une polarisation. *J Physiol Pathol Gen*, 9:620–635.

LeDoux, J. (1996). *The Emotional Brain*. Simon & Schuster, New York.

Lee, K. K. and Xu, Y. (2003). Real-time estimation of facial expression intensity. In *ICRA*, pages 2567–2572.

Li, S. and Zhang, Z. (2004). Floatboost learning and statistical face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:1112–1123.

Liang, D., Yang, J., Zheng, Z., and Chang, Y. (2005). A facial expression recognition system based on supervised locally linear embedding. *Pattern recognition Letter.*, 26:2374–2389.

Lippman, R. (1987). An introduction to computing with neural nets. *IEEE ASSP Magazine*, pages 4–22.

Littlewort, G., Bartlett, M. S., Fasel, I., Kanda, T., Ishiguro, H., and Movellan, J. R. (2004). Towards social robots: Automatic evaluation of human-robot interaction by face detection and expression classification. volume 16, pages 1563–1570. MIT Press.

Loveland, K. A., Tunali-Kotoski, B., Pearson, D. A., Brelsford, K. A., Ortegon, J., and Chen, R. (1994). Imitation and expression of facial affect in autism. *Development and Psychopathology*, 6:433–444.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2:91–110.

Ludemann, P. M. and Nelson, C. A. (1988). Categorical representation of facial expressions by 7-month-old infants. *Developmental Psychology*, 24(4):492 – 501.

Magnée, M., Stekelenburg, J., Kemmer, C., and De Gelder, B. (2007). Similar facial electromyographic responses to faces, voices and body expressions. *Cognitive neuroscience and neuropsychology*, 18:369–372.

Maillard, M., Gapenne, O., Hafemeister, L., and Gaussier, P. (2005a). Perception as a dynamical sensori-motor attraction basin. In *in Advances in Artificial Life (8th European Conference, ECAL*, pages 37–46.

Maillard, M., Gapenne, O., Hafemeister, L., and Gaussier, P. (2005b). Perception as a dynamical sensori-motor attraction basin. In et al., C., editor, *Advances in Artificial Life (8th European Conference, ECAL)*, volume LNAI 3630 of *Lecture Note in ArtificialIntelligence*, pages 37–46. Springer.

Marin-Urias, L., Sisbot, E., and Alami, R. (2008). Geometric tools for perspective taking for human-robot interaction. pages 243–249.

Marin-Urias, L., Sisbot, E., Pandey, A., Tadakuma, R., and Alami, R. (2009). Towards shared attention through geometric reasoning for human robot interaction. *Humanoid 2009*.

Masahiro, M. (1970). The uncanny valey. *Energy*, 7:33–35.

Mataruna, H. and Varela, F. (1980). *Autopoiesis and Cognition: the realization of the living*. Reidel, Dordrecht.

Matsumoto, D. and Ekman, P. (2009). Basic emotions. *Oxford Companion to Emotion and Affective Sciences, Oxford University Press*.

McCulloch, W. S. and Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys*.

Meltzoff, A. N. and Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, 198:75–78.

Meltzoff, A. N. and Moore, M. K. (1983). Newborn infants imitate adult facial gestures. *Child Development*, 54:702–709.

Moga, S., Gaussier, P., and Banquet, J. (2003). Sequence learning using the neural coding. In *IWANN (1)*, pages 198–205.

Moody, J. and Darken, C. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1:281–293.

Muir, D., Kang, L., Hains, C., and Hains, S. (2005). Infant perception and production of emotions during face-to-face interactions with live and virtual adult. *Emotional development*, pages 207–233.

Muir, D. and Nadel, J. (1998). Infant social perception. In Slater, A., editor, *Perceptual development*, pages 247–285. Hove: Psychology Press.

Mumme, D., Fernald, A., and Herrera, C. (1996). Infants's responses to facial and vocal emotional signals in a social referencing paradigm. *Child Development*, 67:3219–3237.

Murray, L. and Trevarthen, C. (1985). Emotional regulation of interaction between two-month-olds and their mother's. In *Social perception in infants*, pages 177–197. N.J. Norwood: Ablex.

Nadel, J., Carchon, I., and Kervella, C. (1999). Expectancies for social contengency in 2-month-olds. *Developmental Science*, 2:164–174.

Nadel, J., Simon, M., Canet, P., Soussignan, R., Blancard, P., Canamero, L., and Gaussier, P. (2006a). Human responses to an expressive robot. In *Epirob 06*.

Nadel, J., Simon, M., Canet, P., Soussignan, R., Blanchard, P., Canamero, L., and Gaussier, P. (2006b). Human responses to an expressive robot. *In Epirob 06*.

Nagai, Y., Hosoda, K., Morita, A., and Asada, M. (2003). A constructive model for the development of joint attention. *Connect. Sci.*, 15(4):211–229.

Nelson, C. A. (1987). The recognition of facial expressions in the first two years of life: Mechanisms of development. *Child Development*, 58:889–909.

Nelson, C. A. and Dolgin, K. M. (1985). The generalized discrimination of facial expressions by seven-month-old infants. *Child Development*, 56:58–61.

O'Keefe, J. (1978). The hippocampus as a cognitive map. *Clarenton Press, Oxford*.

Oudeyer, P. Y., Kaplan, F., and Hafner, V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evolutionary Computation*, 11(2):265–286.

P. Gaussier, K. Prepin, J. N. (2004). Toward a cognitive system algebra: Application to facial expression learning and imitation. *In Embodied Artificial Intelligence, F. Iida, R. Pfeiter, L. Steels and Y. Kuniyoshi (Eds.) published by LNCS/LNAI series of Springer*, pages 243–258.

Papez, J. (1937). A proposed mechanism of emotion. *Archives of Neurology and Psychiatry*.

Paton, J., Belova, M., Morrison, S., and Salzman, C. (2006). The primate amygdala represents the positive and the negative value of visual stimuli during learning. *Nature*, 439:865–870.

Plutchick, R. (1980). A general psychoevolutionary theory of emotion. *Emotion: Theory, research and experience*, pages 3–33.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Prepin, K., Gaussier, P., Revel, A., and Nadel, J. (2005). A formal approach of developmental robotics and psychology : application to the study of a joint attention architecture. *Fifth International Conference on Epigenetic Robotics*, 5:163.

Repacholi, B. (1998). Infants's use of attentional cues to identify the referent of another person's emotional expression. *Developmental Psychology*, 34:1017–1025.

Rime, B. and Giovannini, D. (1986). The physiological patterns of reported emotional states. *Experincing emotion: A cross-cultural study*, pages 84–97.

R.L Hsu, M. Abdel-Mottaleb, A. J. (2002). Face detection in color images. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, 24:696–706.

Rosen, W., Adamson, L., and Bakeman, R. (1992). An experimental investigation of infant social referencing: Mothers's messages and gender differences. *Developmental Psychology*, 28:1172–1178.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408.

Rougier, N. (2009). Implicit and explicit representations. *Neural Networks*, 22(2):155–160.

Rougier, N. and Vitay, J. (2006). Emergence of attention within a neural population. *Neural Networks*, 19:573–581.

Rowley, H. A., Baluja, S., and Kanade, T. (1998). Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:23–38.

Rumelhart and al. (D.E) (1986). Parallel distributed processing. *Cambridge, MIT Press.*

Rumelhart, D. E. and Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, 9:75–112.

Russell, C., Bard, K., and Adamson, L. (1997). Social referencing by young chimpanzees (pan troglodytes). *journal of comparative psychology*, 111(2):185–193.

Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

Russell, J. and Feldman-Barrett, L. (1999). Core affect, prototypical emotional episodes, and others things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76:805–819.

Saarni, C. (1999). The development of emotional competence. *New York, NY: Guilford Press.*

Sander, D., Grandjean, D., and Scherer, K. (2005). A system approach to appraisal mechanisms in emotion. *Neural Networks*, 18:317–352.

Sato, W. and Yoshikawa, S. (2007). Spontaneous facial mimicry in response to dynamic facial expressions. *Cognition*, 104(1):1 – 18.

Scassellati, B. (1998). Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. *In C. Nehaniv, ed., Computation for Metaphors, Analogy and Agents. Lecture Notes in Artificial Intelligence*, 1562:176–195.

Scassellati, B. (2001). Foundations of a theory of mind for a humanoid robot. *Ph.D. dissertation, MIT Department of Computer Science and Electrical Engineering.*

Schachter, S. and Singer, J. (1962). Cognitive, social and physiological determinants of emotional state. *Psychological Review*, 69:379–399.

Scherer, K. (1989). Les emotions: fonctions et composantes. *Dans B. Rime et K. Scherer (dir), les emotions*, pages 97–134.

Scherer, K. (1993). Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition and Emotion*, 7:1–41.

Schöner, G., Dose, M., and Engels, C. (1995). Dynamics of behavior: theory and applications for autonomous robot architectures. *Robotics and Autonomous System*, 16(2-4):213–245.

Schwartz, E. (1977). Spatiale mapping in the primate sensory projection: Analytic structure and relevance to perception. *Biol Cybernetics*, 25:181–194.

Schwartz, E. (1980). Computational anatomy and fonctional architecture of striate cortex: a spacial mapping approach to perceptual coding. *Vision Research*, 20:645–669.

Schwartz, G., Fair, P., Salt, P., Mandel, M., and Klerman, G. (1976). Facial muscle patterning to affective imagery in depressed and non depressed subjects. *Science*, 192:489–491.

Serrano, J. M., Iglesias, J., and Loeches, A. (1992). Visual discrimination and recognition of facial expressions of anger, fear, and surprise in 4- to 6-month-old infants. *Developmental Psychobiology*, 25(6):411–425.

Sigman, M. and Kasari, C. (1994). Social referencing, shared attention, and empathy in infants. *Paper presented at the Ninth International Conference on Infant Studies, Paris: France.*

Silva, L. D., Aizawa, K., and Hatori, M. (1995). Detection and tracking of facial features by using a facial feature model and deformable circular template. *IEICE Trans. Inform. Systems*, pages 1195–1207.

Simon, M., Canet, P., Soussignan, R., Gaussier, P., and Nadel, J. (2007). L'enfant face à des expressions robotiques et humaines. *Enfance*, pages 59–70.

Sinha, P. (1994). Object recognition via image invariants: A case study. *Investigative Ophthalmology and Visual Science*, 25:1735–1740.

Sisbot, E., Marin, L., and Alami, R. (2007). Spatial reasoning for human robot interaction. pages 2281–2287.

Smithers, T. (1995). On quantitative performance measures of robot behaviour. *Robotics and Autonomous Systems*, 15:107–133.

Sokolov, E. (1963). Perception and the conditioned reflex. *New York: Macmillan.*

Sorce, J., Emde, R. N., Campos, J. J., and Klinnert, M. (1985). Maternal emotional signalling : Its effect on the visual cliff behavior of 1 year olds. *Developmental Psychology*, 21(1):195–200.

Sorce, J. F. and Emde, R. N. (1981). Mother's presence is not enough: Effect of emotional availability on infant exploration. *Developmental Psychology*, 17(6):737–745.

Spinoza, B. (1677). Ethique, partie iii : Concernant la nature et l'origine des emotions.

Steels, L. (1994). Mathelmatical analysis of behavior systems. In Gaussier, P. and Nicoud, J., editors, *From Perception to Action: PerAc'94*, pages 88–95. IEEE computer society press.

Stenberg, G. and Hagekull, B. (1997). Social referencing and mood modification in 1-year-olds. *Infant Behavior and Development*, 20:209–217.

Susskind, J., Lee, D., Cusi, A., Feiman, R., Grabski, W., and Anderson, A. (2008). Expressing fear enhances sensory acquisition. *Nature neuroscience*, 11(7):843–850.

Thomaz, A. L., Berlin, M., and Breazeal, C. (2005). An embodied computational model of social referencing. *In IEEE International Workshop on Human Robot Interaction (RO-MAN).*

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381:520–522.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(05):675–691.

Tomasello, M. and Todd (1983). Joint attention and lexical acquisition style. *First language*, 4:197–212.

Trevarthen, C. (1993). The function of emotions in early infant communication. *in J. Nadel and L. Camaioni (Eds.), New perspectives in early communicative development. London, U. K.: Routledge.*

Tronick, E., Als, H., Adamson, L., Wise, S., and Brazelton, T. B. (1978). The infant's response to entrapment between contradictory messages in face-to-face interaction. *Journal of American Academy of Child Psychiatry*, 17:1–13.

Turk, M. and Pentland, A. (1991). Face recognition using eigenfaces. *In IEEE Conference on computer Vision and Pattern Recognition.*, pages 586–591.

Vezhnevets, V., Sazonov, V., and Andreeva, A. (2003). A survey on pixel-based skin color detection techniques. In *in Proc. Graphicon-2003*, pages 85–92.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *CVPR*, pages 511–518.

Viola, P. and Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154.

Walden, T. and Baxter, A. (1989). The effect of context and age on social referencing. *Child Development*, 60:1511–1518.

Walden, T. and Ogan, T. (1988). The development of social referencing. *Child development*, 59:1230–1240.

Walker-Andrews, A. S. (1986). Intermodal perception of expressive behaviors: Relation of eye and voice? *Developmental Psychology*, 22(3):373 – 377.

Walker-Andrews, A. S. (1997). Infants' perception of expressive behaviors : Differentiation of multimodal information. *Psychological Bulletin*, 121:437–456.

Walkers, A. (1982). Intermodal perception of expressive behavior by human infants. *Journal of Experimental Child Psychology*, 33:514–535.

Wallon, H. (1987). Les origines du caractère chez l'enfant. *Paris: PUF, coll. "Quadrige" (Originally Published in 1934).*

Watkins, C. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8:279–292.

Weyers, P., Muhlberger, A., Hefele, C., and Pauli, P. (2006). Electromyographic responses to static and dynamic avatar emotional facial expressions. *Psychophysiology*, 43:450–453.

Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. In *IRE WESCON*, pages 96–104, New York. Convention Record.

Wiskott, L. (1991). Phantom faces for face analysis. *Pattern Recognition*, 30:586–191.

Wood, D., Bruner, J., and Ross, G. (1976). The role of tutoring in problem solving. *J. Child Psychol. Psychiat*, 17:89–100.

Wu, T., Butko, N. J., Ruvulo, P., Bartlett, M. S., and Movellan, J. R. (2009). Learning to make facial expressions. *International Conference on Development and Learning*, 0:1–6.

Wundt, W. (1897). Outlines of psychology. *Engelman, Oxford.*

Yang, F., Paindavoine, M., and Malasne, N. (2003). Localisation et reconnaissance de visages en temps réel avec un réseau de neurones rbf : Algorithme et architecture. *Revue Traitement du Signal*, 20:353–373.

Yang, G. and Huang, T. S. (1994). Human face detection in complex background. *Pattern Recognition*, 27:53–63.

Yang, J., Lu, W., and Waibel, A. (1997). Skin-color modeling and adaptation. In *ACCV '98: Proceedings of the Third Asian Conference on Computer Vision-Volume II*, pages 687–694.

Young-Browne, G., Rosenfeld, H. M., and Horowitz, F. D. (1977). Infant discrimination of facial expression. *Child Development*, 48:555–562.

Yow, K. and Cipolla, R. (1996). Feature-based human face detection. *Image and Vision Computing*, 15:713–735.

Yu, J. and Bhanu, B. (2006). Evolutionnary feature synthesis for facial expression recognition. *Pattern Recognition Letters*.

Yucel, Z., Salah, A., Mericli, C., and mericli, T. (2009). Joint visual attention modeling for naturally interacting robotic agents. *International Symposium on Computer and Information Sciences*, pages 242–247.

Yuille, A., Hallinan, P., and Cohen, D. (1992). Feature extraction from faces using deformable templates. *Int. J. Comput. Vision*, 8(2):99–111.

Yuste, R. and Tank, D. W. (1996). Dendritic integration in mammalian neurons. *a century after Cajal, Neuron*, 16:701–716.