Apprendre à un robot à reconnaître des objets visuels nouveaux et à les associer à des mots nouveaux : le rôle de l'interface

Pierre Rouanet

INRIA BORDEAUX SUD-OUEST

ÉCOLE DOCTORALE MATHÉMATIQUES ET INFORMATIQUE
UNIVERSITÉ BORDEAUX 1 SCIENCES ET TECHNOLOGIES

T H È S E

pour obtenir le titre de

Docteur en Sciences

de l'Université de Bordeaux 1

Mention : Informatique

Présentée et soutenue par

Pierre Rouanet

Apprendre à un robot à reconnaître des objets visuels nouveaux et à les associer à des mots nouveaux : le rôle de l'interface

---

Thèse dirigée par Pierre-Yves Oudeyer
préparée à l'INRIA Bordeaux Sud-Ouest, Équipe FLOWERS
soutenue le 4 avril 2012

Jury :

Rapporteurs :	Mohamed Chetouani	-	Mc	ISIR (Paris)	HdR
	Peter Ford Dominey	-	Dr	CNRS (Lyon)	HdR
Examinateurs :	François Chaumette	-	Dr	INRIA (Rennes)	HdR
	David Filliat	-	Ec	ENSTA - INRIA (Paris)	HdR
	Rodolphe Gelin	-	Dr	Aldebaran Robotics (Paris)	HdR
Directeur :	Pierre-Yves Oudeyer	-	Cr	INRIA - ENSTA (Bordeaux)	HdR

Remerciements

Je tiens tout d'abord à remercier Pierre-Yves Oudeyer, mon directeur de thèse, avec qui travailler durant ces trois années fût un enrichissement et un épanouissement au quotidien. En plus d'avoir toujours su m'orienter, me conseiller et me re-motiver lorsque ce fût nécessaire, sa bonne humeur et son enthousiasme ont grandement contribué à faire des heures passées au laboratoire un réel bonheur.

Je tiens aussi à remercier tout particulièrement Fabien Danieau. Sa collaboration efficace est pour beaucoup dans la réalisation des expériences décrites dans cette thèse. Les longues heures passées à concevoir des études utilisateurs écologiquement valides en ont fait un ami.

Je remercie également David Filliat pour son aide et ses conseils. Sa disponibilité et sa réactivité furent d'une grande aide. Je remercie aussi Louis ten Bosch pour ces quelques mois passés à Bordeaux, où il m'a fait découvrir l'approche NMF.

Je remercie bien sûr Jérôme Béchu pour son aide et son support au quotidien. Nos discussions sur les cubes et les heures passées à conduire sur les routes de Californie resteront de magnifiques souvenirs.

Je remercie également tout particulièrement Marie Sanchez et Nathalie Robin, nos assistantes successives, dont l'enthousiasme et la motivation m'ont permis de partir en mission un peu partout dans le monde, même lorsque je m'y prenais à la dernière minute...

Durant ces trois dernières années, j'ai eu la chance de voir l'équipe FLOWERS naître et grandir. Je tiens à remercier toutes les personnes qui ont participé à en faire un cadre de travail aussi enrichissant qu'agréable : Alexandre, Adrien, Blaise, Bérenger, Clément, Damian, Damien, les Fabien, Franck, mon pti Fred, Haylee, Hong Li, Jérémy, Jonathan, Jonas, Manuel, Mai, les Matthieu, Ming Li, les deux Olivier, Paul, Timothée et tous les Thomas. Je remercie aussi tous nos collègues de l'ENSTA, même si je n'ai pas souvent eu l'occasion de les croiser.

Je tiens aussi à remercier l'ensemble du personnels de Cap Sciences pour leur disponibilité et leur bonne humeur. Ils ont fait des journées passées là-bas un réel plaisir.

Je remercie enfin toutes les personnes qui ont accepté de relire ce manuscrit de thèse, afin de le corriger. Leurs suggestions et commentaires ont contribué à l'améliorer.

Et bien sûr, je veux dire un grand merci à ma famille et à mes amis. Ils ont toujours été derrière moi pendant ces trois ans. Ils m'ont apporté le soutien et la motivation nécessaire qui m'ont permis de continuer lors des moments de doute.

Résumé

Cette thèse s'intéresse au rôle de l'interface dans l'interaction humain-robot pour l'apprentissage. Elle étudie comment une interface bien conçue peut aider les utilisateurs non-experts à guider l'apprentissage social d'un robot, notamment en facilitant les situations d'attention partagée. Nous étudierons comment l'interface peut rendre l'interaction plus robuste, plus intuitive, mais aussi peut pousser les humains à fournir les bons exemples d'apprentissage qui amélioreront les performances de l'ensemble du système. Nous examinerons cette question dans le cadre de la robotique personnelle où l'apprentissage social peut jouer un rôle clé dans la découverte et l'adaptation d'un robot à son environnement immédiat. Nous avons choisi d'étudier le rôle de l'interface sur une instance particulière d'apprentissage social : l'apprentissage conjoint d'objets visuels et de mots nouveaux par un robot en interaction avec un humain non-expert. Ce défi représente en effet un levier important du développement de la robotique personnelle, l'acquisition du langage chez les robots et la communication entre un humain et un robot. Nous avons particulièrement étudié les défis d'interaction tels que le pointage et l'attention partagée.

Nous présenterons au chapitre 1 une description de notre contexte applicatif : la robotique personnelle. Nous décrirons ensuite au chapitre 2 les problématiques liées au développement de robots sociaux et aux interactions avec l'homme. Enfin, au chapitre 3 nous présenterons la question de l'interface dans l'acquisition des premiers mots du langage chez les robots. La démarche centrée utilisateur suivie tout au long du travail de cette thèse sera décrite au chapitre 4. Dans les chapitres suivants, nous présenterons les différentes contributions de cette thèse. Au chapitre 5, nous montrerons comment des interfaces basées sur des objets médiateurs peuvent permettre de guider un robot dans un environnement du quotidien encombré. Au chapitre 6, nous présenterons un système complet basé sur des interfaces humain-robot, des algorithmes de perception visuelle et des mécanismes d'apprentissage, afin d'étudier l'impact des interfaces sur la qualité des exemples d'apprentissage d'objets visuels collectés. Une évaluation à grande échelle de ces interfaces, conçue sous forme de jeu robotique afin de reproduire des conditions réalistes d'utilisation hors-laboratoire, sera décrite au chapitre 7. Au chapitre 8, nous présenterons une extension de ce système permettant la collecte semi-automatique d'exemples d'apprentissage d'objets visuels. Nous étudierons ensuite la question de l'acquisition conjointe de mots vocaux nouveaux associés aux objets visuels dans le chapitre 9. Nous montrerons comment l'interface peut permettre d'améliorer les performances du système de reconnaissance vocale, et de faire directement catégoriser les exemples d'apprentissage à l'utilisateur à travers des interactions simples et transparentes. Enfin, les limites et extensions possibles de ces contributions seront présentées au chapitre 10.

Mots clés : interaction humain-robot, attention partagée, acquisition du langage, conception d'interface, robotique personnelle et sociale

Abstract

This thesis is interested in the role of interfaces in human-robot interactions for learning. In particular it studies how a well conceived interface can aid users, and more specifically non-expert users, to guide social learning of a robotic student, notably by facilitating situations of joint attention. We study how the interface can make the interaction more robust, more intuitive, but can also push the humans to provide good learning examples which permits the improvement of performance of the system as a whole. We examine this question in the realm of personal robotics where social learning can play a key role in the discovery and adaptation of a robot in its immediate environment. We have chosen to study this question of the role of the interface in social learning within a particular instance of learning : the combined learning of visual objects and new words by a robot in interactions with a non-expert human. Indeed this challenge represents an important an lever in the development of personal robotics, the acquisition of language for robots, and natural communication between a human and a robot. We have studied more particularly the challenge of human-robot interaction with respect to pointing and joint attention.

We present first of all in Chapter 1 a description of our context : personal robotics. We then describe in Chapter 2 the problems which are more specifically linked to social robotic development and interactions with people. Finally, in Chapter 3, we present the question of interfaces in acquisition of the first words of language for a robot. The user centered approach followed throughout the work of this thesis will be described in Chapter 4. In the following chapters, we present the different contributions of this thesis. In Chapter 5, we show how some interfaces based on mediator objects can permit the guiding of a personal robot in a cluttered home environment. In Chapter 6, we present a complete system based on human-robot interfaces, the algorithms of visual perception and machine learning in order to study the impact of interfaces, and more specifically the role of different feedback of what the robot perceives, on the quality of collected learning examples of visual objects. A large scale user-study of these interfaces, designed in the form of a robotic game that reproduces realistic conditions of use outside of a laboratory, will be described in details in Chapter 7. In Chapter 8, we present an extension of the system which allows the collection of semi-automatic learning examples of visual objects. We then study the question of combined acquisition of new vocal words associated with visual objects in Chapter 9. We show that the interface can permit both the improvement of the performance of the speech recognition and direct categorization of the different learning examples through simple and transparent user's interactions. Finally, a discussion of the limits and possible extensions of these contributions will be presented in Chapter 10.

Keywords : human-robot interaction, joint attention, language acquisition, interface design, personal and social robotic

Table des matières

Chapitre 1

Introduction

Sommaire

Résumé du chapitre

La robotique personnelle est un domaine en pleine expansion, avec des enjeux très forts, mais soulevant également un ensemble de défis scientifiques, technologiques et culturels importants. Nous allons les présenter ici et nous intéresser plus particulièrement à un défi transverse : comment un robot peut s'adapter à son environnement. Nous montrerons qu'une des solutions possibles à cette question est l'apprentissage social où l'humain guide l'apprentissage d'un robot. Nous nous proposons ici d'étudier le rôle de l'interface dans cet enseignement. Nous avons choisi d'étudier cette problématique dans le cadre plus particulier de l'enseignement conjoint d'objets visuels et de mots nouveaux à un robot.

1.1 Un domaine à forts enjeux économiques et sociétaux : la robotique personnelle et sociale

De nombreux indicateurs semblent montrer que l'arrivée de robots personnels dans nos maisons et nos vies quotidiennes pourrait être un des événements majeurs du XXI^e siècle [Gates 2007][Rahimi 2009]. Ce type de robots pourrait jouer un rôle de plus en plus important dans nos sociétés et notamment dans l'accompagnement du vieillissement de la population et le maintien des personnes à domicile. Des études prédisent que la robotique personnelle connaitra une très forte croissance ces prochaines années. La robotique personnelle pourrait connaître le même développement au XXI^e siècle que celui connu par l'informatique au siècle précédent, passant d'une informatique rare et très chère à une informatique personnelle présente partout dans notre quotidien. La robotique personnelle constitue donc un enjeu majeur des années à venir aussi bien du point de vue de la société que du point de vue économique où elle représente un marché à très fort potentiel. Parmi les applications possibles de la robotique personnelle nous pouvons citer la robotique de service, l'aide à la personne, la télé-présence ou télé-surveillance, l'éducation, la thérapie assisté par robot ou encore le divertissement. Le lecteur pourra se référer à [Fong 2003a] pour avoir une revue plus complète des applications possibles.

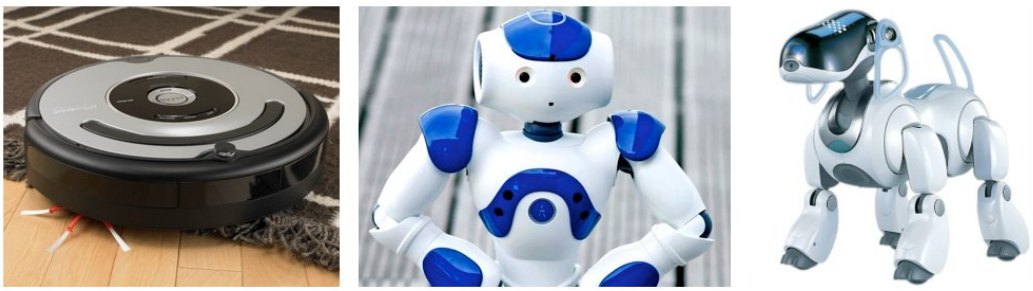


FIGURE 1.1 – Exemples de robots personnels et sociaux (de gauche à droite : Roomba, Nao et l'Aibo).

1.2 Des défis scientifiques et technologiques

Pour permettre à la robotique personnelle de connaître un véritable essor, de très nombreux défis scientifiques et technologiques restent à résoudre. En effet, contrairement à la robotique industrielle, les robots personnels devront pouvoir évoluer dans un milieu non-structuré sur lequel très peu de suppositions peuvent être faites. Nous avons choisi de présenter un panorama des défis liés à la robotique personnelle et quotidienne selon six grandes catégories : perception, navigation, conception, interaction avec l'humain, langage et acceptation. Chacun de ces défis sera découpé en sous défis détaillant les principales questions de recherche ainsi que les applications

possibles. Les questions plus particulièrement liées au problème de recherche étudiée dans cette thèse seront décrites plus en détails par la suite.

1.2.1 Perception et analyse de l'environnement

Un des défis majeurs de la robotique personnelle est de doter les robots de la capacité de percevoir leur environnement et de l'analyser afin de pouvoir y évoluer. Les travaux réalisés dans ce domaine s'intéressent principalement aux questions de perception visuelle, auditive et haptique.

Pour pouvoir évoluer dans nos maisons, le robot doit tout d'abord être capable de percevoir visuellement son environnement immédiat et il faut notamment pouvoir permettre au robot de percevoir et reconnaître les objets présents dans son environnement immédiat [Lowe 1999][Meger 2008] et lui permettre d'éviter les obstacles. Le robot doit aussi pouvoir percevoir les humains qui l'entourent, leur visage, leurs gestes [Nickel 2004][Correa 2009]. Il doit également être capable de reconnaître et de comprendre leurs comportements, leurs états émotionnels et leurs activités afin de pouvoir s'y adapter [Cowie 2001].

Le robot doit aussi pouvoir entendre et comprendre les bruits produits dans son environnement immédiat, e.g. entendre lorsqu'un humain lui parle ou l'appelle. Il doit donc être capable de savoir d'où vient un son, et d'identifier qui parle [Chetouani 2009a][Kazuhiro 2009]. Il faut aussi qu'il puisse reconnaître des commandes dans un flux de parole et percevoir l'état émotionnel des humains à travers leurs paroles [Breazeal 2002a][Chetouani 2009b].

Un autre défi technologique et scientifique important est de permettre à ces robots d'être dotés du sens du toucher. Des capteurs tactiles ont, par exemple, été développés pour permettre à un robot d'éviter les chocs importants [Alami 2006]. L'interaction tactile a aussi été étudiée pour faciliter la communication entre un humain et un robot [Miyashita 2005].

1.2.2 Navigation et manipulation

Un autre défi crucial pour le développement de la robotique personnelle concerne la navigation. Un robot doit pouvoir se déplacer dans une maison qu'il ne connait pas a priori. Il doit pouvoir se déplacer dans un environnement encombré, sur un sol inégal et donc avoir une locomotion robuste. Pour permettre à un robot d'évoluer dans un environnement conçu pour les humains (e.g. avec des escaliers, des passages étroits) faut-il construire des robots bipèdes ayant une marche inspirée de la marche humaine ? Permettre à un robot bipède de se déplacer dans un environnement non-contraint de manière robuste soulève encore de nombreuses difficultés [Raibert 2008] D'autres recherches visent à permettre à un robot de se construire une carte de son environnement afin de pouvoir planifier ses déplacements. En particulier, elles s'intéressent à comment permettre à un robot de cartographier une maison qu'il ne connait pas à l'avance et où l'environnement peut être modifiée à tout moment [Filliat 2003][Durrant-Whyte 2006].

En plus de pouvoir permettre à un robot de se déplacer dans une maison, il faut aussi que le robot puisse attraper des objets et les transporter. Pour cela, le robot doit planifier les mouvements de son corps et plus particulièrement de son bras pour attraper un objet. Il faut qu'il puisse attraper des objets ayant des formes très différentes. Comment peut-il y arriver lorsqu'il ne connait pas les objets à l'avance ? La construction de mains robotisées inspirées des mains humaines a aussi été explorée pour permettre une plus grande robustesse de préhension [Bicchi 2002][Edin 2008].

1.2.3 Interactions sociales avec les humains

Les robots personnels évolueront dans un monde qu'ils partageront avec les humains et devront donc pouvoir interagir avec eux. Ils devront être capables d'adopter un comportement qui soit socialement acceptable par les humains. Il est donc nécessaire de modéliser un tel comportement et de permettre à un robot de s'y conformer. Le robot doit aussi pouvoir adapter son comportement à la situation et aux attentes des utilisateurs. Il faut également qu'il puisse exprimer son état interne ou des « émotions » pour permettre à l'humain de mieux comprendre son comportement. Différents canaux peuvent être utilisés pour permettre une interaction fluide et intuitive par l'humain : e.g. les expressions du visage ou les mouvements [Scassellati 2001][Breazeal 2002a].

Il faut aussi permettre à l'humain d'interagir facilement et directement avec un robot. De nombreux travaux étudient comment reconnaître et analyser les interactions naturelles utilisées par les humains telles que les gestes de pointages, le suivi de regard ou encore la reconnaissance de la parole [Scassellati 1996][Haasch 2004]. D'autres chercheurs étudient comment la conception d'interfaces peut permettre de canaliser l'interaction entre un humain et un robot afin de la rendre plus robuste et/ou plus intuitive [Goodrich 2007]. La conception d'études permettant d'évaluer l'utilisabilité réelle et perçue d'une interaction humain-robot a été particulièrement étudiée. Différentes méthodologies ont été proposées pour garantir que les systèmes développés soient effectivement utilisables et appréciés par les humains [Walters 2005][Weiss 2009].

1.2.4 Langage naturel

Le langage naturel est très utilisé par les humains pour communiquer lors des interactions du quotidien. Un robot personnel devrait donc pouvoir reconnaître et utiliser ce langage. En particulier, le robot devrait être capable de reconnaître les commandes prononcées par les humains. Pour cela, il devrait pouvoir identifier les mots dans un flux de parole. Il est aussi intéressant de s'intéresser à la manière dont un robot peut-il acquérir le sens de mots nouveaux et comment ces mots peuvent-ils être reliés au monde réel. L'humain peut aussi guider un robot dans son acquisition du langage [Steels 2000][Iwahashi 2007][Cangelosi 2010].

1.2.5 Acceptabilité et éthique

Comme les robots personnels sont amenés à partager notre quotidien et notre intimité, ils vont avoir un impact fort sur nos sociétés. Il est donc important de faciliter leur acceptabilité par les humains et plus particulièrement par les personnes non-technophiles. Des recherches s'intéressent à la question de l'intégration naturelle de ces robots dans notre quotidien sans provoquer de heurts ni de craintes [Heerink 2006].

L'utilisation de robot personnel soulève aussi des questions d'éthique. Par exemple, un robot peut-il être utilisé pour l'assistance de personnes fragiles ? Comment évaluer réellement l'efficacité de ces robots lors de leur utilisation dans un cadre thérapeutique ? Comment garantir que le robot ne causera ni dommages physiques ni dommages moraux aux utilisateurs ? Les robots vont-ils permettre de renforcer les liens sociaux de personnes fragiles ou au contraire renforcer leur isolement ? [Tapus 2007][Dautenhahn 2002a]

1.2.6 Conception, morphologie et apparence

En plus des défis présentés ci-dessus, se posent également des défis techniques et d'ingénierie, pour la conception de robots personnels qui aient en même temps une morphologie suffisamment complexe pour pouvoir réaliser une grande variété d'actions, mais suffisamment petite et légère pour pouvoir évoluer sans problème dans une maison « ordinaire ». Il est important de pouvoir satisfaire ces critères tout en réalisant un robot à un coût permettant son achat par le grand public. Il faut aussi concevoir des robots dont l'interaction physique ne présente pas de danger pour les utilisateurs [Haddadin 2010].

L'influence de la morphologie et du design esthétique sur le modèle que l'utilisateur se fait du robot a aussi été très étudiée [Li 2010]. La conception de robots inspirés de la morphologie des animaux/humains a été étudiée comme un facteur permettant de faciliter leur évolution dans un monde conçu pour les humains. Le design d'un robot (e.g. sa forme, ses mouvements) peut également faciliter la compréhension de son comportement par l'utilisateur [Bartneck 2009]. Des recherches cherchent à copier aussi fidèlement que possible la morphologie humaine alors que d'autres s'intéressent à la conception de morphologies caricaturales présentant clairement aux utilisateurs les capacités du robot [Walters 2008].

1.2.7 Un défi transverse : s'adapter à son environnement

Nous nous intéressons dans cette thèse à une famille de défis transverses à ces différentes catégories : permettre à un robot de s'adapter à son environnement. Contrairement à la robotique industrielle, où l'environnement est très structuré et connu à l'avance, la plupart des robots personnels devront évoluer dans un monde a priori inconnu, changeant et non-contraint. Ils devront de plus interagir avec des personnes ayant des attentes, des besoins et des compétences différents. Ces robots devront donc être capables d'adapter leur comportement à leur environnement, à

la situation et à leurs utilisateurs. Il n'est en effet pas possible de pré-programmer l'ensemble des comportements d'un robot lui permettant d'interagir dans un milieu aussi peu contraint qu'une maison.

Cette adaptation peut se faire à travers deux grandes familles d'apprentissage. Dans le première de ces familles, le robot explore par lui-même son environnement et essaie d'acquérir de nouveaux savoir-faire. En particulier, se posent ici des questions sur la manière dont un robot peut acquérir des savoir-faire nouveaux pour lui permettre de répondre adéquatement à une situation nouvelle, ou encore de savoir comment il peut apprendre par lui-même une nouvelle tâche. Les questions de représentation et de ré-utilisation de ses tâches ont été largement étudiées. Certaines approches cherchent à permettre à un robot d'explorer de nouvelles actions seul. Le robot doit, dans ce cas, pouvoir réaliser qu'il a fait une erreur ou qu'il fait des progrès [Sutton 1998][Oudeyer 2007]. L'apprentissage peut aussi se faire à travers les interactions avec les humains. Cependant, il faut s'intéresser à la manière dont un humain non-expert peut efficacement et intuitivement guider un robot dans son apprentissage. En particulier, un utilisateur doit pouvoir montrer à un robot comment réaliser une tâche. Le robot doit lui pouvoir extraire les informations pertinentes dans une démonstration humaine, et également pouvoir généraliser ces démonstrations [Breazeal 2005][Thomaz 2008][Billard 2008].

1.3 Le rôle de l'interface pour l'apprentissage social en robotique

1.3.1 L'apprentissage social en robotique

Parmi l'ensemble des problématiques étudiant comment un robot peut s'adapter à son environnement, nous nous intéressons ici à l'apprentissage social robotique, où l'humain guide l'apprentissage d'un robot. De nombreuses études ont montré que l'apprentissage social était un moyen efficace pour permettre le transfert de compétences de l'humain vers le robot. En jouant le rôle d'enseignant, l'humain peut à travers des démonstrations, ou grâce à des instructions, réduire considérablement l'espace des possibilités à explorer et accélérer fortement l'apprentissage et faciliter une meilleure généralisation des savoir-faire acquis [Billard 2008].

L'apprentissage social en robotique peut se présenter sous des formes différentes : l'observation de l'humain par le robot, le tutorat (e.g. [Lockerd 2004][Thomaz 2007]) ou encore l'apprentissage par démonstration [Billard 2008][Argall 2009]. Ces formes d'apprentissages s'inspirent directement des nombreuses études sur l'apprentissage social chez l'humain ou chez certains animaux [Piaget 1945][Nehaniv 2004]. Ces approches diffèrent par le degré d'autonomie laissé au robot lors de l'apprentissage (guidage-exploration) [Thomaz 2007].

Une des formes les plus utilisées d'apprentissage social en robotique est l'apprentissage par démonstration, ou apprentissage par imitation, où un humain effectue des démonstrations d'une tâche à un robot [Billard 2008][Argall 2009]. L'observation

de ces exemples, qu'ils soient bons ou mauvais, permet de réduire significativement l'espace d'apprentissage possible, c'est à dire l'ensemble des combinaisons sensori-motrices possibles. Le robot peut en effet directement explorer les solutions possibles pour réaliser une action « autour » des démonstrations réussies de cette action qu'il a observées. Les démonstrations peuvent être réalisées directement par l'humain, dans ce cas le robot doit les observer et les transposer à son propre corps pour pouvoir les reproduire, ou l'humain peut agir directement sur le robot, par interaction kinesthésique (e.g. [Calinon 2007a]) ou via télé-opération (e.g. [Evrard 2009]). Ce type d'apprentissage soulève cependant des questions fondamentales souvent résumées en : *Que faut-il imiter ? Comment imiter ? Quand et qui imiter ?* [Billard 2008]. Calinon et al. ont, par exemple, exploré l'utilisation de méthodes statistiques basées sur des modèles à base de mélange de gaussiennes (*GMM : gaussian mixture model*) pour encoder des trajectoires de mouvements démontrés par l'humain et la régression de mélange de gaussienne (*GMR : gaussian mixture regression*) pour les généraliser et permettre au robot de les reproduire [Calinon 2007a]. D'autres travaux ont étudié le rôle de l'interaction humain-robot dans l'apprentissage par imitation et notamment comment des indices sociaux peuvent souligner les composantes les plus pertinentes d'une démonstration de l'humain [Breazeal 2005].

Dans l'approche d'apprentissage par tutorat, l'utilisateur fournit directement des instructions au robot (e.g. en dirigeant son attention) lui permettant de réaliser une tâche ou d'acquérir de nouveaux savoir-faire [Lockerd 2004]. Lauria et al. ont, par exemple, montré comment le langage naturel pouvait être utilisé par un utilisateur pour apprendre à un robot à naviguer à travers une ville miniature [Lauria 2002].

Une autre forme d'apprentissage social en robotique est l'exploration socialement guidée (*Socially guided exploration*). Ici, le robot explore par lui-même comment réaliser une tâche nouvelle, mais les indices sociaux détectés lors de l'interaction avec un humain vont être utilisés pour guider cet apprentissage [Thomaz 2007]. Le robot est ici doté de mécanismes de motivations intrinsèques le poussant à explorer le monde par lui-même (e.g. [Nguyen 2011]), ainsi que de mécanismes de détection d'indices sociaux tels que le suivi de regard ou les gestes de pointage. Ces indices sociaux peuvent être utilisés comme une récompense d'un apprentissage par renforcement. D'autres techniques inspirées du dressage animal ont aussi parfois été utilisées [Kaplan 2001].

Comme souligné par les travaux décrits ci-dessus, l'apprentissage social en robotique met en jeu deux aspects différents. D'une part, la mise en place d'algorithmes d'apprentissage permettant, par exemple, d'encoder des connaissances ou des savoir-faire nouveaux, de les généraliser et d'en construire de nouveaux à partir de ceux précédemment acquis. D'autre part, il est également nécessaire de développer des techniques d'interaction naturelles et intuitives entre l'humain et un robot, et également permettre au robot d'utiliser efficacement les indices fournis par l'humain lors de l'interaction [Breazeal 2002a][Revel 2004][Thomaz 2008]. Ces deux défis sont complémentaires et peuvent se renforcer mutuellement : l'interaction humain-robot permet de faciliter l'apprentissage social qui peut alors permettre une plus grande robustesse dans l'interaction.

Le rôle spécifique de l'interface dans l'apprentissage social a cependant été très peu étudié jusqu'à présent. Ces mécanismes sont pourtant connus pour jouer un rôle majeur dans l'apprentissage social chez l'humain et chez certains animaux [Miller 2001]. Thomaz et Breazeal ont montré l'importance de comprendre la relation enseignant humain/élève robotique pour concevoir des algorithmes d'apprentissage adaptés à la manière d'enseigner de l'humain et ainsi améliorer l'apprentissage du robot [Thomaz 2008]. Cakmak et al. ont étudié comment un partenaire social peut influencer l'apprentissage d'un robot [Cakmak 2009]. Calinon et Billard ont montré qu'un système d'apprentissage par imitation devait prendre en compte le scénario d'interaction et le rôle actif de l'utilisateur [Calinon 2007b]. Dans cette thèse, nous allons étudier le rôle de l'interface et des facteurs humains dans l'apprentissage social en robotique. Cette étude sera faite dans le cadre de l'enseignement conjoint d'objets visuels et de mots nouveaux à un robot. Plus précisément, nous allons étudier comment une interface bien conçue peut d'une part faciliter l'interaction entre l'humain et un robot mais aussi faciliter les situations d'attention partagée et ainsi permettre à un utilisateur de fournir de meilleurs exemples d'apprentissage qui vont améliorer l'apprentissage du robot.

1.3.2 Robotique développementale : inspiration fonctionnelle des mécanismes d'attention partagée

Notre travail se situe aussi dans le cadre de la robotique développementale. Cette branche de la robotique tente de reconstruire le procédé de développement du cerveau d'un enfant et ainsi permettre à un robot d'apprendre à la manière d'un enfant. Pour atteindre cet objectif ambitieux, les travaux en robotique développementale s'inspirent des concepts et théories de la psychologie développementale et tentent de les formaliser afin de les transposer à la robotique [Weng 2001][Lungarella 2003][Johnson 2005]. Cette approche cherche d'une part, à modéliser le développement de l'humain afin de mieux comprendre le vivant et d'autre part, à permettre une plus grande adaptabilité et une plus grande robustesse de l'apprentissage afin de développer des machines, et en particulier des robots, plus polyvalents. C'est ce deuxième aspect qui nous intéresse particulièrement ici. Cette thèse n'a, en effet, pas pour objectif d'éclairer certains aspects du développement dans le vivant. Par contre, nous cherchons en effet à nous inspirer fonctionnellement des mécanismes d'interaction existants chez l'humain, connus pour jouer un rôle fondamental dans l'apprentissage social et à les transposer à l'interaction humain-robot afin d'en étudier l'impact.

Des recherches en psychologie développementale ont, en particulier, souligné l'importance de l'attention partagée dans l'apprentissage social et le développement du langage [Tomasello 1995][Tomasello 2004]. Les enfants et certains animaux sont dotés de mécanismes attentionnels leur permettant de se concentrer sur des évènements saillants (e.g. mouvement, bruit). Kaplan et Hafner ont identifié les défis sous-jacents au développement de l'attention partagée entre un humain et un robot à partir de l'identification des principales étapes du développement de ces mécanismes atten-

tionnels chez l'enfant [Kaplan 2004]. Les enfants acquièrent, lors des deux premières années de leur développement, la capacité de percevoir, manipuler et coordonner les mécanismes attentionnels. Kaplan et Hafner ont proposé un modèle d'attention partagée basé sur les quatre capacités suivantes :

- pouvoir détecter l'attention d'un autre agent (e.g suivre son regard)
- pouvoir manipuler l'attention d'un autre agent (e.g. à l'aide de gestes de pointage),
- pouvoir se coordonner socialement (e.g. prise de parole),
- et pouvoir comprendre l'intentionnalité des autres agents.

En détaillant la chronologie de l'acquisition des capacités nécessaires à l'attention partagée chez l'enfant, Kaplan et Hafner ont montré qu'il existe différentes étapes clés dans l'acquisition interconnectée de ces quatre pré-requis [Kaplan 2004]. Par exemple, concernant la détection de l'attention, l'enfant peut, dans un premier temps, simplement échanger des regards avec un adulte puis est capable de suivre le regard d'une personne de plus en plus précisément et enfin peut interpréter les gestes de pointage. De même, pour l'acquisition de la manipulation de l'attention, l'enfant commence par utiliser des gestes de pointage d'abord impératifs (requête d'obtention d'un objet) puis déclaratif (pouvant désigner des objets même hors de vue) et enfin accompagne ses gestes de mots pour spécifier un aspect de l'objet désigné.

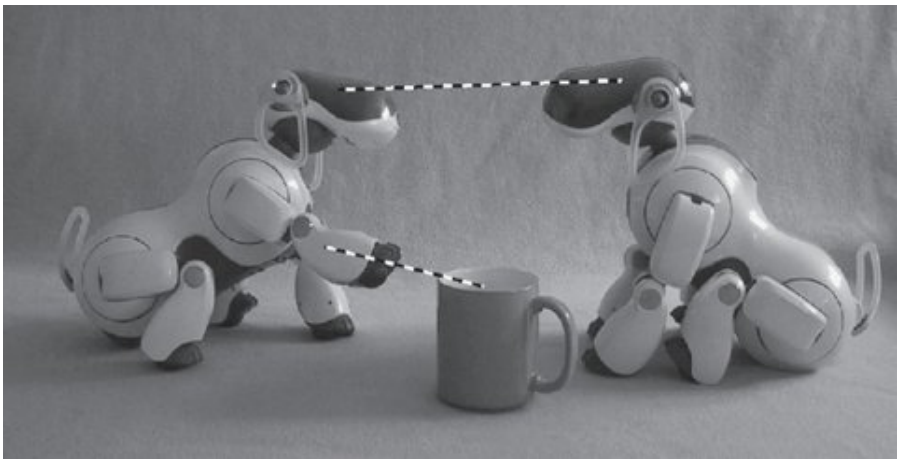


FIGURE 1.2 – L'établissement de l'attention partagée à travers des mécanismes tels que le suivi de regard ou la reconnaissance de geste joue un rôle fondamental dans l'apprentissage social chez l'humain mais aussi en robotique (source [Kaplan 2004]).

Le développement de modèles robotiques d'attention partagée peut permettre de mieux comprendre le développement de ces mécanismes chez l'enfant. Par exemple, Scasselati a présenté des mécanismes d'attention partagée basés sur le suivi de regard [Scassellati 1996]. Nagai et al. ont montré comment un robot pouvait apprendre à suivre le regard d'un humain pour trouver des objets [Nagai 2003]. Cependant, ici nous ne cherchons pas à reproduire un modèle compatible avec les observations du

développement de l'enfant, mais à nous inspirer fonctionnellement de ces mécanismes attentionnels, afin d'en étudier l'impact sur l'apprentissage social en robotique. En particulier, nous allons principalement nous concentrer sur la détection et la manipulation de l'attention afin de permettre à l'élève robotique et à son enseignant humain de se concentrer conjointement sur ce qui est le plus important à apprendre.

1.3.3 Interaction humain-robot

L'étude du rôle de l'interface dans l'apprentissage social et notamment dans la facilitation de situations d'attention partagée entre un humain et un robot se place également dans le domaine de recherche de l'interaction humain-robot, souvent désignée par le sigle HRI (*human-robot interaction*). L'étude de l'interaction entre un humain et un robot est un domaine de recherche vaste et en pleine expansion. En effet, pour pouvoir intégrer des robots en douceur à notre quotidien il est nécessaire que ces robots puissent interagir naturellement, efficacement et sans heurt avec des humains même non-experts. Ce défi soulève une très grande variété de questions de recherches dont nous allons présenter une vue globale ici. Les problématiques plus particulièrement liées à cette thèse et à notre contexte applicatif seront détaillées dans la section 3.3.

Pour permettre à un robot d'interagir naturellement et sans heurt avec un humain, il faut pouvoir répondre aux questions suivantes : Comment modéliser un comportement social adéquat ? Comment l'adapter à la situation et aux différents utilisateurs ? [Scassellati 2001][Breazeal 2002a] Pour faciliter la communication, il faut aussi permettre au robot d'exprimer son état interne aux utilisateurs. Comment permettre à un robot d'exprimer ses émotions ? Quelles modalités sont les plus adaptées (e.g. les expressions faciales) ? [Breazeal 2002a] La morphologie et l'apparence des robots ont aussi été étudiées afin d'en évaluer l'impact sur les attentes des utilisateurs. Faut-il imiter aussi précisément que possible une apparence humaine ou au contraire s'inspirer de personnages caricaturaux utilisés en animation par exemple ? [Dautenhahn 2002b][Fong 2002]

Le robot et l'humain devront aussi pouvoir travailler en collaboration. Comment un robot et un humain peuvent-ils se coordonner ? Quels mécanismes de synchronisation utiliser ? Comment permettre au robot d'inférer le comportement de l'humain pour qu'il puisse s'y adapter ? [Hinds 2004][Sofge 2004]

En plus de ces questions, il est aussi indispensable d'étudier les différentes interfaces, c'est-à-dire les différents moyens possibles pour transférer les informations entre l'humain et un robot. Certaines interfaces sont des transpositions directes des interfaces humaines (e.g. les gestes, la parole, le regard) alors que d'autres s'inspirent des interfaces humain-machine et sont par exemple basées sur des terminaux mobiles communicants. Quel type d'interface permet une interaction intuitive et robuste ? [Goodrich 2007]

L'évaluation de l'interaction entre un humain et un robot a aussi fait l'objet de nombreuses recherches. Quels critères permettent de caractériser la robustesse de l'interaction ? Son intuitivité ? Mais aussi comment évaluer l'expérience utilisateur ?

[Steinfeld 2006][Weiss 2009] Quelle méthodologie utiliser pour permettre une évaluation écologiquement valide ? Comment concevoir des études utilisateurs réalistes et intéressantes pour les participants ? Comment évaluer l'interaction humain-robot sur le plus long terme ? [Walters 2005]

1.4 Un contexte applicatif : l'apprentissage social du langage

Comme nous l'avons présenté dans la section précédente, l'étude du rôle de l'interface pour l'apprentissage social en robotique se positionne au carrefour de trois grands domaines de recherches : la robotique développementale, l'apprentissage social en robotique et l'interaction humain-robot. Dans cette thèse, nous avons choisi d'étudier cette question pour un apprentissage social particulier : l'apprentissage social du langage. L'acquisition du langage représente en effet un défi majeur pour la robotique personnelle et sociale. Le langage peut permettre à un humain et un robot de communiquer naturellement et est donc une brique de base à l'acquisition et au développement de nombreux autres savoir-faire variés. De plus, cette activité est prépondérante pour permettre à des utilisateurs de faire découvrir à leur robot son environnement afin de lui permettre d'y évoluer et d'interagir avec les différents objets qui l'entourent [Tomasello 1995][Tomasello 1999].

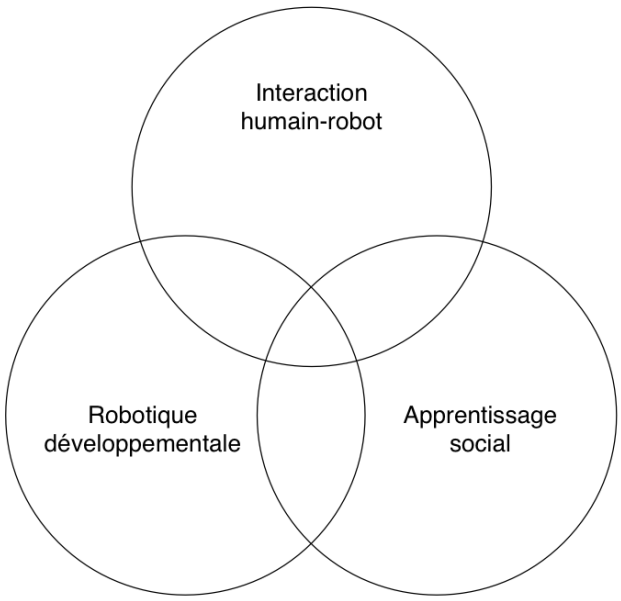


FIGURE 1.3 – La question de recherche étudiée dans cette thèse, le rôle de l'interface pour l'apprentissage social en robotique, se positionne au carrefour des trois domaines de recherches suivants : la robotique développementale, l'apprentissage social et l'interaction humain-robot.

L'acquisition du langage en robotique représente un défi très vaste et nous allons donc ici nous restreindre à un problème plus spécifique : l'acquisition des premiers mots du langage. Plus précisément, nous allons étudier le rôle de l'interface dans l'enseignement de mots nouveaux associés à des objets visuels nouveaux à un robot. Comme nous allons le voir par la suite, bien que contournant certains des défis importants liés à l'acquisition du langage chez les robots, ce défi soulève déjà de très nombreuses questions notamment d'interaction et d'interface entre l'homme et le robot, qui nous intéressent particulièrement dans cette thèse. En effet, de même qu'il a été montré que des phénomènes tels que l'attention partagée jouaient un rôle fondamental dans l'acquisition du langage chez l'enfant [Tomasello 1995][Tomasello 2004], nous pensons que ces mécanismes, bien que peu étudiés en robotique, ont aussi un rôle prépondérant à y jouer.

Dans les chapitres suivants, nous allons présenter un système complet permettant à un utilisateur de guider son robot, d'attirer son attention vers des objets nouveaux et de les nommer afin de lui permettre de les reconnaître plus tard. Nous allons bien sûr présenter les algorithmes de vision et d'apprentissage utilisés, cependant la véritable contribution de ce travail réside dans l'étude des différentes interfaces développées, pour permettre à la fois aux utilisateurs d'enseigner intuitivement et efficacement des objets visuels et mots nouveaux à leur robot, mais également de les pousser à collecter de bons exemples d'apprentissage, dont la qualité aura un impact positif sur les performances de l'ensemble du système. Ainsi dans les chapitres suivants, nous montrerons que l'interface peut avoir un impact sur les performances globales du système d'apprentissage, d'un ordre de magnitude bien supérieur aux améliorations typiques obtenues lors de l'utilisation d'algorithmes de vision ou d'inférence statistiques plus perfectionnés.

La conception de ces différentes interfaces a été inspirée fonctionnellement des mécanismes et contraintes de l'acquisition du langage chez l'enfant (notamment le partage de l'attention) mais transposés à la robotique et à l'interaction homme-robot. En effet, du fait des espaces sensori-moteurs très différents entre un humain et son robot personnel, il n'est pas possible de copier directement les mécanismes utilisés par l'humain mais il est nécessaire de les adapter à l'interaction homme-robot de la même manière qu'il a été nécessaire de développer de nouvelles manières d'interagir pour permettre de communiquer avec des singes bonobos. Savage-Rumbaugh et al. ont, en effet, conçu un tableau sur de grandes feuilles de papier présentant différents symboles correspondant à des objets de la vie courante, à des activités ou encore à des concepts simples. Grâce à cet intermédiaire, Kanzi, un mâle bonobo, a pu apprendre le sens de ces symboles, à les combiner en structure grammaticale simple ainsi qu'à les associer avec les mots anglais correspondants. Un tel résultat n'avait pas pu être obtenu par les méthodes d'enseignement du langage communément utilisées chez l'humain [Rumbaugh 1996]. Dans le chapitre 2, nous décrirons tout d'abord les problématiques liées à la robotique sociale et aux interactions entre un humain et un robot. Nous présenterons ensuite, au chapitre 3, la question de l'interface dans l'acquisition des premiers mots du langage chez les robots.

Nous présenterons ensuite la démarche de développement centrée-utilisateur sui-

FIGURE 1.4 – De la même manière qu'il a fallu développer de nouveaux outils pour permettre la communication entre l'homme et le bonobo, nous avons développé des interfaces basées sur des objets médiateurs permettant à un humain d'enseigner de manière robuste des mots nouveaux associés à des objets visuels à son robot (ici une interface basée sur un iPhone).

vie tout au long du travail de cette thèse, afin de garantir que les interfaces proposées et étudiées ici permettent réellement à un utilisateur non-expert d'enseigner à son robot des mots nouveaux associés à des objets visuels nouveaux (chapitre 4). Les chapitres suivants présenteront les différentes itérations de ce cycle de développement, notamment des interfaces permettant de guider le robot (chapitre 5), de lui montrer des objets pour qu'il apprenne à les reconnaître (chapitre 6) ainsi que leur évaluation à travers une étude à grande échelle conçue sous forme de jeu robotique (chapitre 7). Nous présenterons finalement une extension de ces interfaces permettant à l'humain l'utilisation de mots vocaux nouveaux. Plutôt qu'utiliser un système de reconnaissance vocale « *off-the-shelf* », transformant les sons en symboles, dont les performances sont limitées en milieu non-contraint, nous proposerons notre propre méthode de comparaison de mots vocaux. Nous montrerons comment l'interface, couplée à ce système, peut, d'une part, permettre de faire participer l'utilisateur au processus de reconnaissance vocale afin d'en améliorer la robustesse, et d'autre part, à l'association des différents exemples d'apprentissage et ainsi améliorer très significativement les performances du système dans son ensemble (chapitre 9).

Nous présenterons ensuite les limites et extensions possibles du travail présenté ici (chapitre 10) et conclurons.

Chapitre 2

Robots sociaux et interactifs : un domaine de recherche en pleine expansion

Sommaire

Résumé du chapitre

Dans ce chapitre, nous présentons une revue des enjeux et défis scientifiques et technologiques liés à la robotique sociale et interactive. Nous parlerons en particulier de l'influence sur l'interaction de la perception du robot par l'utilisateur, des différentes interfaces humain-robot existantes, de méthodologie d'évaluation de l'interaction ainsi que d'acceptation de ces robots par les humains. Les thématiques de recherche directement liées au sujet d'étude de cette thèse, le rôle de l'interface dans l'apprentissage social, seront plus particulièrement mises en avant.

2.1 Un domaine en pleine expansion

La robotique sociale est un champ de recherche émergent visant spécifiquement à permettre à des robots de reconnaître un agent (un humain ou un autre robot social), d'engager et de maintenir des interactions sociales avec lui [Dautenhahn 1999]. Ce champ d'étude se trouve donc à la croisée de nombreux domaines de recherche

parmi lesquels nous pouvons citer la robotique, les sciences sociales, les sciences cognitives ou encore l'interaction humain-robot. En plus des problématiques liées à la robotique traditionnelle, Fong et al. ont défini un ensemble de caractéristiques nécessaires au développement d'un robot social interactif [Fong 2002]. Ces robots doivent notamment pouvoir :

- établir et maintenir des relations sociales,
- exprimer et percevoir des émotions,
- communiquer et adopter un comportement social acceptable : e.g. respecter les conventions sociales,
- se construire un modèle des autres agents,
- utiliser les indices sociaux, tels que le regard, les gestes ou la parole, et
- manifester une personnalité.

Nous nous limiterons ici à l'étude des interactions de type pair-à-pair humain-robot : i.e. entre un seul robot et un humain et nous nous concentrerons sur les interactions de type co-location, c'est-à-dire où le robot et l'humain sont physiquement présents l'un à côté de l'autre. En effet, cette configuration représente les conditions favorisées de l'apprentissage social de type élève/enseignant qui nous intéresse dans cette thèse. Nous ne parlerons donc pas d'interaction mettant en jeu plusieurs robots sociaux (voir e.g. [Mataric 2001]).

La robotique sociale connait à l'heure actuelle un développement important. En effet, manifester des comportement sociaux facilite l'interaction humain-robot, rend l'interaction plus compréhensible par l'humain et en améliore la robustesse [Breazeal 2002a]. Doter les robots de mécanismes sociaux permet de donner des repères au robot et à l'humain, particulièrement s'il est non-expert, sur la manière dont ils peuvent interagir. Il existe aussi un intérêt croissant autour de l'apprentissage social en robotique, où l'humain guide l'exploration d'un robot à l'aide d'indices sociaux tels que des expressions du visage, la parole ou encore des démonstrations. En utilisant les indices fournis par l'humain, le robot peut restreindre son exploration, en se focalisant sur ce que l'humain lui a indiqué comme étant intéressant, et permet donc d'accélérer significativement l'apprentissage [Billard 2008]. Cette thèse se place dans ce cadre de recherche. Nous étudions en effet le rôle de l'interface dans l'apprentissage social, plus précisément du langage. Nous avons déjà présenté un ensemble de défis de recherches liés à l'apprentissage social en robotique dans la section 1.3.1. Parmi l'ensemble de ces défis, l'utilisation de démonstrations pour guider un robot a fait l'objet de nombreuses recherches. En particulier, comment représenter ces démonstrations, comment les transposer au corps du robot ou encore comment les généraliser sont autant de défis majeurs de l'apprentissage social [Calinon 2007a][Calinon 2007b][Billard 2008][Argall 2009]. L'utilisation d'indices sociaux (e.g. suivi de regard, gestes) pour guider socialement l'exploration d'un robot est aussi un sujet d'étude en plein développement [Thomaz 2007][Thomaz 2008][Cakmak 2009][Nguyen 2011].

En plus de rendre l'interaction humain-robot plus naturelle pour l'humain et de jouer un rôle clé dans le guidage de l'apprentissage chez les robots, la robotique sociale cherche aussi à étudier et/ou à valider des modèles sociaux issus de la psycho-

logie développementale ou de la sociologie. Les comportements sociaux d'un robot peuvent aussi servir à influencer le comportement de l'humain, pour le motiver par exemple. Fong et al. ont présenté une liste des applications permises par la robotique sociale [Fong 2002]. Nous pouvons notamment citer : les services et l'assistance à la personne (voir [Tapus 2007] pour une revue détaillée des défis liés au robot sociaux d'assistance), la thérapie (e.g. [Dautenhahn 2002a]), le divertissement, l'éducation ou encore utiliser des robots comme sujets de tests pour l'étude de modèles sociaux humains (e.g. [Scassellati 2001][Breazeal 2002a]).

Afin d'étudier le rôle que l'interface peut avoir dans l'apprentissage social chez les robots, il est important de présenter une revue de l'ensemble des questions de recherche soulevées par la robotique sociale et interactive. Il reste en effet de nombreux défis scientifiques à lever avant qu'un humain et un robot puisse maintenir des interactions sociales robustes et intuitives. Nous allons présenter ces différentes problématiques en détails dans les sections ci-dessous en mettant plus en avant celles dont les thématiques sont directement liées à la question de recherche de cette thèse : le rôle de l'interface dans l'apprentissage social du langage. Nous présenterons les différentes recherches qui ont étudié comment l'incarnation, la morphologie ou les mouvements du robot influençaient les attentes de l'utilisateur et donc leur perception de l'interaction. Nous discuterons aussi des travaux qui se sont intéressés aux différents modes de communication pouvant faciliter une interaction sociale intuitive et robuste. Nous présenterons des méthodologies d'évaluation de l'interaction sociale entre un humain et un robot. Enfin, nous parlerons, de l'acceptation des robots sociaux par les humains non-nécessairement amateurs de technologie.

2.2 Influence de la perception des robots sociaux sur les attentes de l'utilisateur et sur l'interaction

La perception des robots sociaux par les humains influence leurs attentes et l'interaction, plus particulièrement les premiers instants. Différents aspects tels que la morphologie, l'aspect, la personnalité ou encore les émotions manifestées par le robot influencent les attentes des utilisateurs. Il est crucial de modéliser ces attentes de l'utilisateur, d'une part afin de mieux les comprendre et pouvoir les prendre en compte lors de la conception et évaluation des interactions sociales humain-robot, mais aussi afin de permettre au robot de se construire un modèle théorique du comportement de l'utilisateur et ainsi lui permettre de prévoir et s'adapter aux interactions de son partenaire [Hafner 2011].

Vision a priori des robots

La vision a priori des robots sociaux influence fortement la manière dont les utilisateurs perçoivent l'interaction avec un robot. Khan a montré que ce modèle est fortement influencé par la littérature et les films de science-fiction [Khan 1998]. Nomura et al. ont montré que les a priori, plus particulièrement ceux négatifs, avaient

un impact sur la façon dont les utilisateurs interagissaient et se comportaient avec un robot. Ils ont proposé différentes méthodes de mesures de ces attitudes négatives à l'égard des robots et identifié différents facteurs les influençant : les expériences passées d'utilisation de robot, l'éducation ainsi que l'âge [Nomura 2006]. Bumby et Dautenhahn se sont plus particulièrement intéressées à la façon dont les enfants perçoivent les robots et ont montré qu'ils les conçoivent comme des formes géométriques dotées de libre arbitre et d'émotions [Bumby 1999].

Incarnation, morphologie et aspect

L'incarnation du robot (*embodiment*), c'est-à-dire sa capacité à perturber et à être perturbé par l'environnement, joue aussi un rôle important sur l'établissement d'interactions sociales humain-robot [Breazeal 2002a]. Dautenhahn et al. ont proposé différents niveaux d'incarnation pour les robots interagissant socialement [Dautenhahn 2002b]. L'impact de plusieurs types de morphologies sur les attentes sociales des utilisateurs ont été étudiées. L'apparence physique influence particulièrement les premiers instants de l'interaction. Fong et al. ont proposé quatre types d'apparence différentes pour les robots sociaux :

- anthropomorphique : la morphologie de ces robots reproduit structurellement et fonctionnellement l'humain afin de permettre aux humains d'interagir avec ces robots de la même manière qu'ils interagissent avec les autres humains (e.g. [Scassellati 2001])
- zoomorphique : afin d'établir des relations de type humain-créature dont les attentes sont moins élevées que les relations de type humain-humain
- caricaturale : l'accentuation de certains traits du robot peut guider les attentes des utilisateurs (e.g. [Scheeff 2000])
- fonctionnelle : l'aspect du robot reflète directement les tâches qu'il peut réaliser

La morphologie peut aussi servir à indiquer à l'utilisateur quel type d'interaction sont possibles. Par exemple, Breazeal a montré comment la conception d'un visage expressif robotique permettait de contraindre l'interaction à des interactions face à face [Breazeal 2002a].

Personnalité

L'influence de la personnalité, ce qui distingue un individu d'un autre, d'un robot sur les attentes des utilisateur a aussi été étudiée. De nombreuses recherches ont essayé de doter les robots sociaux de personnalité et d'en étudier l'impact sur l'interaction : Le robot doit-il avoir une personnalité imitant celle de l'humain pour faciliter l'interaction ? Cette personnalité doit-elle évoluer ? Quelle influence la personnalité du robot va-t-elle avoir sur l'interaction ? Woods et al. ont exploré la question de la personnalité d'un robot et plus précisément si les humains projetaient leur propre personnalité dans celle du robot [Woods 2005]. Goetz et Kisler ont examiné l'influence de différentes personnalités sur une interaction humain-robot. Ils ont notamment montré qu'une personnalité charmante n'était pas le meilleur moyen pour

obtenir la collaboration de l'humain [Goetz 2002]. Meerbeek et Saerbeck ont proposé une méthodologie de conception et l'évaluation d'une personnalité pour un robot social autonome [Meerbeek 2009].

Mouvements du robot

Les mouvements du robot jouent aussi un rôle important dans la compréhension de son comportement social pour l'utilisateur. Par exemple, Bogdan et al. ont montré comment les mouvements du robot pouvaient améliorer la communication avec un humain [Bogdan 2011]. Plus précisément, ils ont montré que les variations de vitesse d'un chariot robotique, conçu pour suggérer des produits à l'utilisateur dans un supermarché, permettaient de souligner les commentaires vocaux du robot et augmentaient les chances que l'humain suive ses conseils.

D'autres travaux ont, par exemple, étudié la distance à laquelle un robot doit se tenir pour engager une interaction avec un humain en respectant les codes sociaux. Walters et al. ont, à partir des théories de la communication humaine, conçu un modèle de distance à maintenir lors de l'interaction entre un humain et un robot [Walters 2007].

Émotions

Les émotions jouent également un rôle majeur dans la communication et l'interaction sociale. Elles apportent de l'information supplémentaire et peuvent agir comme un mécanisme de contrôle de l'interaction [Armon–Jones 1986] et peuvent aider le robot à détecter l'état émotionnel de l'humain et ainsi adapter son comportement. Le robot peut aussi, à travers ses émotions, présenter à l'humain son état interne de fonctionnement et permettre ainsi une plus grande transparence dans l'interaction et donc une meilleure compréhension mutuelle [Breazeal 2003].

L'utilisation d'émotions en robotique sociale soulève de nombreuses questions : Comment un robot peut-il exprimer ses émotions ? Quelle(s) modalité(s) seront facilement interprétées par l'humain ? par le robot ? Les émotions peuvent être exprimées à travers les expressions du visage (e.g. [Kobayashi 1994]), la prosodie du discours (e.g. [Breazeal 2002a]) ou encore le langage corporel. Une classification des émotions en terme de valence et d'excitation a été introduite par [Russell 1997]. Comment l'humain perçoit-il ces émotions ? Canamero et Fredslund ont, par exemple, évalué les performances d'un humain pour reconnaître des expressions du visage exprimées par un robot [Canamero 2001]. Différentes architectures, plus ou moins inspirées des recherches en théorie des émotions, ont aussi été proposées pour permettre au robot d'exprimer des émotions simples [Nourbakhsh 1999][Schulte 1999].

L'étude des facteurs présentés ci-dessus soulignent l'influence que ces mécanismes jouent sur les attentes des utilisateurs et donc sur leur perception de l'interaction. Bien qu'ils ne soient pas l'objet d'étude central de notre travail, ces critères restent néanmoins à considérer lors de la conception et/ou évaluation d'interaction social humain-robot.

2.3 Interfaces utilisateurs

Une autre composante cruciale de la robotique sociale et interactive est la question de l'interface : i.e. la manière dont les informations peuvent être échangées entre l'humain et le robot. De nombreuses formes d'interfaces humain-robot ont été proposées ces dernières années. Elles se focalisent sur trois des sens humains : la vision, l'audition et le toucher. Goodrich et Schultz proposent de séparer les différentes interfaces humain-robot en quatre groupes [Goodrich 2007] :

- les affichages visuels (e.g. une interface basée sur un smartphone),
- les gestes et mouvements (e.g. les gestes de la main),
- le langage naturel (e.g. les commandes vocales),
- les interactions physiques.

Comme précisé plus haut, nous nous intéressons, dans cette revue de la littérature, aux interactions où l'utilisateur est physiquement présent à côté du robot. Dans cette section nous nous intéresserons donc principalement aux interfaces permettant d'interagir en co-location avec un robot. Dans le cadre de notre contexte applicatif, l'apprentissage social chez les robots, nous avons choisi de présenter les différentes interfaces humain-robot existantes selon deux grandes approches :

- les interfaces transposées directement des interactions humaines, e.g. les gestes, la parole ou le suivi de regard,
- les interfaces basées sur des objets qui vont servir de médiateurs entre l'humain et le robot, e.g. les terminaux mobiles communicants, les interfaces tangibles ou encore les pointeurs lasers.

2.3.1 Transposition directe des interactions humaines

Les humains sont des experts de l'interaction sociale et apprennent très tôt à utiliser des mécanismes tels que les gestes des mains ou des bras, le suivi du regard ou encore la parole pour interagir [Tomasello 1995]. Il est donc intéressant d'étudier s'il est possible de permettre à un humain d'interagir avec un robot de la même manière qu'il interagit avec les autres humains. Cette approche fournit potentiellement des interfaces naturelles dans le sens où elles sont déjà maitrisées par les humains et ne nécessitent donc pas de prise en main particulière. De plus, le développement de ce type d'interface peut aussi permettre d'étudier les modèles d'interactions humaines.

2.3.1.1 Gestes

De nombreux travaux ont étudié comment les gestes humains pouvaient être utilisés pour communiquer avec un robot. Les gestes de pointage jouent un rôle particulièrement important lors des interactions humaines en permettant notamment d'attirer l'attention et de désigner des objets particuliers. La reconnaissance de gestes a fait l'objet de beaucoup de recherches dans le domaine de la perception et de l'apprentissage machine (e.g. [Wu 1999][Nickel 2004]). Hafner et Kaplan ont aussi montré comment un robot pouvait apprendre à reconnaître des gestes de pointage [Hafner 2004].

Ces gestes ont été utilisés dans des scénarios d'interactions variés. Par exemple, Scassellati a développé un système permettant d'attirer l'attention d'un torse robotique assis face à une table à l'aide de geste de pointage [Scassellati 1996]. Haasch et al. ont montré comment ces gestes pouvaient être utilisés pour interagir avec un robot compagnon [Haasch 2004]. Perzanowski et al. ont aussi utilisé les gestes de pointage pour guider un robot [Perzanowski 2001].

2.3.1.2 Suivi de regard

Le suivi de regard est un mécanisme attentionnel très important chez l'humain. Cette modalité a donc naturellement été étudiée dans le cadre de l'interaction avec un robot. Des algorithmes d'estimation visuelle du regard et de la position de la tête ont été développés afin de détecter précisément l'endroit où regarde un humain [Newman 2000][Morimoto 2005]. Atienza et Zelinsky ont par exemple utilisé le suivi de regard pour désigner des objets à un robot afin qu'il les saisisse [Atienza 2003]. Matsumoto et al. ont développé une interface basée sur le suivi de regard pour diriger un fauteuil roulant robotique [Matsumoto 2001]. Haasch et al. ont aussi utilisé la détection de regard pour faciliter l'interaction avec un robot compagnon [Haasch 2004]. Scassellati a utilisé le suivi de regard pour développer des mécanismes d'attention partagée avec un robot humanoïde [Scassellati 1996]. Staudte et Crocker ont montré comment le suivi de regard pouvait faciliter la compréhension mutuelle entre un humain et un robot [Staudte 2009].

2.3.1.3 Reconnaissance de la parole

La reconnaissance de la parole a également été utilisée dans le cadre de l'interaction humain-robot. Lauria et al. ont proposé un système permettant de programmer un robot à travers l'utilisation du langage naturel, afin qu'il parcourt une ville miniature [Lauria 2002]. Dominey et al. ont montré comment le langage parlé pouvait être utilisé pour programmer une interaction en coopération avec un robot humanoïde [Dominey 2007][Dominey 2009]. Roy a enseigné des mots pour des couleurs et des formes à un bras robotisé à travers le langage naturel [Roy 2003]. Perzanowski et al. ont développé une interface humain-robot multi-modale combinant la reconnaissance de la parole avec la reconnaissance de geste pour diriger un robot [Perzanowski 2001]. McGuire et al. ont, quant à eux, développé une interface multimodale combinant les gestes de pointage avec la parole pour faire attraper des objets à un robot [Mcguire 2002].

Les systèmes de reconnaissance vocale « *off-the-shelf* », tels que Siri [1] ou Google Voice [2], bien que de plus en plus performants, souffrent encore de problèmes de robustesse. Plus précisément, ces systèmes se basent sur les les séquences de mots pour améliorer leur performance et éprouvent donc des difficultés à reconnaître les mots isolés. Ils supposent de plus que l'utilisateur parlera juste en face d'un micro, d'un

1. `http://www.apple.com/iphone/features/siri.html`
2. `http://www.google.com/insidesearch/voicesearch.html`

téléphone par exemple, et ne sont que peu robustes au bruit environnant, inévitable lors des interactions humain-robot en milieu ouvert. Des projets importants, tel que le projet européen SERA [3] (*Social Engagement with Robots and Agents*), se sont intéressés à l'utilisation de ce type de système de reconnaissance vocale dans le cadre des interactions entre un humain et un robot en environnement quotidien. Malgré l'utilisation de micros professionnels et de différents systèmes de reconnaissance vocale de l'état de l'art, le fond sonore ambiant ne permettait pas d'obtenir une reconnaissance suffisamment robuste. Ils ont finalement remplacé la reconnaissance vocale par un système basé sur des cartes RFID. La modalité vocale est, cependant, très intéressante pour l'interaction humain-robot et particulièrement dans le cadre de l'acquisition sociale du langage qui nous intéresse dans cette thèse. Nous avons donc développé un système de reconnaissance vocale, basé sur la combinaison d'une mesure de similarité entre des mots acoustiques et d'une interface bien conçue, permettant à l'utilisateur d'intervenir dans le processus de reconnaissance et ainsi d'en améliorer la robustesse. Ce système sera décrit en détails dans le chapitre 9.

Les interfaces transposées de l'interaction humaine présentées ci-dessus fournissent donc potentiellement des interactions naturelles ne nécessitant pas d'apprentissage particulier. Cependant, ces interfaces souffrent de problèmes de robustesse liées au bruit (e.g. changement de lumière, fond sonore ambiant) rendant la reconnaissance et l'interprétation de ces interactions difficiles en milieu ouvert. Leur utilisation est donc souvent limitée à des milieux très contraints (e.g. utilisation du suivi de regard avec un robot fixé à une table), qui vont permettre de contourner certaines de ces difficultés. Par exemple, en fixant le robot face à une table, il est possible de s'assurer que le robot perçoive l'ensemble de la scène, mais aussi de contrôler l'éclairage et le bruit sonore de la pièce.

2.3.2 Interfaces basées sur des objets médiateurs

L'utilisation d'objets de la vie courante a aussi été étudiée pour la conception d'interfaces homme-robot cherchant à canaliser l'interaction et ainsi la rendre plus robuste. Ce type d'interface a été utilisé pour de nombreuses applications robotiques notamment de navigation et pour désigner des objets à un robot. À notre connaissance ce type d'interfaces n'a pas été utilisé pour l'apprentissage social en robotique.

2.3.2.1 Terminal mobile communicant (TMC)

De nombreuses interfaces humain-robot basées sur des affichages visuels ont été développées récemment. En effet, ce type de dispositif permet un transfert d'informations dans les deux sens : du robot vers l'humain (e.g. à travers l'affichage visuel du terminal) et de l'humain vers le robot (e.g. à travers des boutons ou un écran tactile) [Fong 2001]. La plupart de ces interfaces s'inspirent des études en interface

3. `http://project-sera.eu/`

humain-machine et en reprennent les méthodes d'interactions classiques. Cependant, ce type d'interface implique également l'utilisation d'écrans de taille réduite et contraint donc le type d'interactions possible.

Ce type d'interface humain-robot a été largement utilisé pour la navigation, particulièrement en télé-opération. Fong et al. ont par exemple développé un système de télé-opération basé sur un TMC [Fong 2003b]. Ils ont étudié plusieurs modes de navigation : un mode de contrôle direct où l'utilisateur dirige le robot à l'aide de flèches virtuelles, une interaction de type pointer-et-cliquer directement sur l'image issue de la caméra du robot ou via une carte de l'environnement [Fong 2001]. Ils ont aussi montré que ce type d'interface permettait de développer facilement des interactions collaboratives où l'humain peut superviser le robot et où le robot peut tirer profit des capacités de l'humain [Fong 2001]. De même, Kaymaz et al. ont développé une interface de télé-opération où les utilisateurs peuvent diriger un robot, soit à l'aide de flèches affichées directement sur le flux vidéo issu de la caméra du robot, soit sur une représentation de la scène vue du dessus construite à partir des différents capteurs du robot (sonar et laser) [Kaymaz 2003]. Ces auteurs ont souligné l'importance d'avoir des représentations différentes de la scène (e.g. vue du robot ou vue du dessus) pour permettre à l'opérateur d'avoir une meilleure connaissance de la situation. Dalgalarrondo et al. ont également étudié comment ce type d'interface pouvait être utilisé pour ajuster le niveau d'autonomie d'un robot de télé-surveillance [Dalgalarrondo 2004]. Enfin, Skubic et al. ont montré comment l'utilisateur pouvait dessiner directement sur l'écran tactile d'un terminal une carte approximative de l'environnement afin de diriger un robot [Skubic 2002]. Bien que développée pour la télé-opération, ces interfaces peuvent aussi être utilisées en co-location.

FIGURE 2.1 – Nous pouvons différentes utilisations d'interfaces, basées sur des objets médiateurs, pour l'interaction humain robot : un TMC est utilisé pour diriger un robot sur l'image de gauche (source [Fong 2003b]), sur l'image du milieu, un laser est utilisé pour désigner des objets (source [Ishii 2009]) et enfin sur l'image de droite un contrôleur Wiimote permet de faire prendre des positions spécifiques à un robot (source [Guo 2008a]).

Il existe aussi d'autres types d'applications robotiques pour l'utilisation d'interfaces basées sur un terminal mobile communicant. Huttenrauch et Norman ont, par exemple, proposé d'utiliser une interface basée sur un TMC pour contrôler un robot de service [Huttenrauch 2001]. Ils ont présenté différents prototypes permettant de diriger le robot, de contrôler et modifier les tâches qu'il effectue. Sakamoto et al. ont montré comment un TMC pouvait être utilisé pour contrôler un robot de nettoyage [Sakamoto 2009]. L'utilisateur peut dessiner différentes commandes sur l'écran de l'appareil affichant une représentation, vue du plafond, de la pièce. Les terminaux mobiles communicants ont aussi été utilisés pour demander à un robot d'attraper un objet. Choi et al. ont par exemple montré comment un écran tactile affichant le flux vidéo de la caméra du robot pouvait être utilisé par les humains pour désigner les objets à faire attraper par le robot simplement en cliquant dessus [Choi 2008]. Yanco et al. ont également utilisé ce type d'interface pour permettre à un utilisateur de sélectionner des objets qui seront saisis par un bras robotique fixé à un fauteuil roulant [Tsui 2008]. Cependant, dans ces travaux les auteurs supposent que le robot soit capable d'attraper un objet simplement à partir d'une position 3D et donc que le robot soit capable de segmenter visuellement l'objet, ce qui reste un problème difficile (voir la section 6.5.2 pour une description de ce problème). Calinon et Billard ont aussi utilisé un TMC pour fournir différentes modalités d'interaction humain-robot. Ils ont en particulier montré comment à l'aide de la caméra un robot pouvait détecter les mouvements des bras et de la tête d'un utilisateur afin de les imiter et comment le microphone pouvait être utilisé pour enregistrer des commandes vocales [Calinon 2003].

2.3.2.2 Pointeur laser

Les pointeurs lasers sont un moyen couramment utilisé pour désigner des objets ou un endroit précis à d'autres personnes. Ce type d'interface a donc naturellement été transposé à l'interaction humain-robot. Kemp et al. ont proposé une interface de type pointer-et-cliquer pour désigner des objets à un robot afin qu'il les attrape [Kemp 2008]. Ils peuvent ainsi attirer l'attention d'un robot sur des points particuliers. Cependant, ils supposent ici que le robot puisse se saisir d'un objet simplement à partir d'une indication de position. Ishii et al. ont développé une interface basée sur l'utilisation d'un pointeur laser, où les utilisateurs dessinent directement dans l'environnement des esquisses au laser afin de spécifier différentes commandes à un robot de service [Ishii 2009]. Cependant, afin de pouvoir détecter le point du laser, Kemp et al. ont dû utiliser un robot équipé d'une caméra omnidirectionnelle et Ishii et al. ont utilisé des caméras fixées au plafond. Il semble cependant difficile de supposer l'existence de tels systèmes dans le cadre de la robotique personnelle où les robots sont souvent dotés de capteurs très limités.

2.3.2.3 Interface tangible

Les interfaces tangibles, souvent désignées par TUI (*Tangible User Interface*), ont aussi été couramment utilisée en interaction humain-robot. Les utilisateurs in-

teragissent ici physiquement avec un objet dont les mouvements sont transformés en commandes envoyées au robot. Guo et Sharlin ont par exemple utilisé un contrôleur Wiimote, doté d'accéléromètres enregistrant les mouvements, pour diriger un robot à travers un parcours [Guo 2008a]. Des commandes reproduisant les gestes effectués par les cavaliers pour diriger un cheval ont été utilisées. Ils ont également montré que cette interface tangible pouvait aussi permettre de faire prendre des positions particulières à un robot [Guo 2008b]. Gams et Mudry ont utilisé une Wiimote dont les mouvements sont transformés en commande afin de faire jouer du tambour à un robot [Gams 2008].

Un des principaux avantages de ce type d'interface, souligné par les auteurs cités ci-dessus, est de permettre aux utilisateurs de constamment focaliser leur attention sur le robot. Les humains peuvent en effet déplacer leurs mains sans avoir besoin de les regarder. Cependant, ce type d'interface ne permet pas de présenter une importante variété de retours à l'utilisateur. Le retour haptique ne permet pas, par exemple, de rendre compte de ce que le robot perçoit.

2.3.3 Interfaces multi-modales

Des interfaces multi-modales ont également été utilisées en interaction humain-robot. Combiner plusieurs interfaces permet généralement d'améliorer l'utilisabilité, en contrebalançant les faiblesses d'une modalité par les forces d'une autre, ainsi qu'en levant les ambiguïtés possibles. La reconnaissance de la parole a par exemple souvent été couplée avec la reconnaissance des gestes de pointage [Perzanowski 2001][Haasch 2004]. Il pourrait aussi être intéressant de combiner une interface tangible avec une autre interface permettant de fournir des retours à l'utilisateur plus complet. Nous souhaitions, par exemple, utiliser un TMC équipé d'accéléromètres afin de combiner les avantages d'une interface tangible (utilisation très simple) et les avantages des TMC (transfert d'information). Cependant, des tests pilotes ont indiqué que l'inclinaison de l'appareil dégradait trop la visibilité de l'écran pour rendre son utilisation robuste. Cette interface ne sera pas décrite dans cette thèse. Nous présenterons par contre une autre interface multi-modale (section 6.3.3) combinant une interface tangible à une interface basée sur un pointeur laser.

2.4 Méthodologie d'évaluation

Avec le développement de la robotique sociale et interactive, il devient crucial de pouvoir caractériser les interactions entre un humain et un robot. L'interaction sociale humain-robot présente, en effet, des caractéristiques uniques la différenciant des autres interactions avec des machines, et il est donc nécessaire de développer des méthodologies particulières permettant l'évaluation et l'étude des interactions avec les robots sociaux interactifs. En particulier, il est nécessaire de développer des métriques caractéristiques de l'interaction humain-robot permettant de rendre compte de son efficacité, de son intuitivité, mais aussi du niveau d'engagement de l'humain ou de ses attentes.

Walters et al. ont décrit un ensemble de méthodologies et de bonnes pratiques pour le développement d'études de l'interaction humain-robot [Walters 2005]. Steinfeld et al. ont identifié un ensemble de métriques communes à l'interaction humain-robot (e.g. le degré d'autonomie ou la connaissance de la situation), ainsi qu'un ensemble de biais possibles de ce type d'évaluation [Steinfeld 2006]. Weiss et al. ont proposé un cadre d'évaluation centré-utilisateur d'une interaction humain-robot [Weiss 2009]. Les auteurs définissent différents critères permettant d'évaluer l'utilisabilité, l'acceptation sociale, l'expérience utilisateur et l'impact social d'une interaction humain-robot. Différentes méthodes inspirées de l'interaction homme-machine ainsi que de la psychologie sont proposées pour évaluer ces différents critères. Kahn et al. ont présenté cinq approches (psychométrique, littéraire, par modélisation, philosophique et structurelle) pour valider différentes caractérisations des niveaux de sociabilité d'un robot [Kahn 2010]. Ils décrivent chacune de ces approches et proposent des clés permettant de choisir les approches les plus adaptées aux différentes questions de recherches liées à l'interaction humain-robot.

Bien que les robots sociaux et interactifs soient amenés à communiquer directement avec des humains non-experts, peu d'études ont réellement évalué des interactions humain-robot hors laboratoire. Kanda et al. ont présenté une étude de terrain où un robot communicant était déployé dans un centre commercial et renseignait les visiteurs [Kanda 2009]. Ils décrivent notamment une méthodologie de collecte et d'analyse de données permettant de caractériser le niveau de communication d'un robot dans un environnement non-contraint. Les interactions sociales humain-robot à long-terme ont également été très peu étudiées. Huttenrauch et Severinson-Eklund ont réalisé une étude de terrain d'un robot de service dans un bureau. Ils ont montré que la manifestation de comportements sociaux du robot facilitait son acceptation par les humains [Huttenrauch 2002].

Nous décrirons en détails la méthodologie d'évaluation choisie dans cette thèse au chapitre 4.

2.5 Acceptation et sécurité

En plus de l'ensemble des questions de recherche présentées ci-dessus, un autre défi majeur de la robotique sociale est l'acceptation de ces robots par les humains. Ces types de robots devront notamment interagir avec des personnes fragiles (e.g. les personnes âgées) et la sécurité de ces populations devra donc aussi être prise en considération. Tapus et al. ont présenté un ensemble de défis du développement de la robotique sociale d'assistance tels que susciter l'engagement des humains ou encore permettre le transfert de compétences acquises lors de l'interaction avec le robot à l'interaction humain-humain (e.g. dans le cas de l'autisme) [Tapus 2007].

La conception de l'interaction peut aussi améliorer/garantir la sécurité des utilisateurs. Dahl et al. ont, par exemple, montré comment des réflexes de retrait pouvaient être intégrés sur un robot Nao, à l'aide de capteurs tactiles, afin de prévenir les

contacts violents [Dahl 2011]. Le projet européen PHRIENDS [4] (*Physical Human-Robot Interaction : DepENDability and Safety*) s'intéresse à la conception de robots intrinsèquement sûrs en étudiant des algorithmes de contrôle moteurs, l'utilisation de capteurs de force et des méthodes d'évitement de collisions [Haddadin 2010]. Ils ont aussi proposé un ensemble de métriques permettant d'évaluer le niveau de sécurité d'un robot.

En plus de garantir la sécurité des utilisateurs, les robots sociaux doivent aussi pouvoir être acceptés par les humains. Comment évaluer le niveau d'acceptation d'un robot ? Kupferberg et al. ont proposé une méthode d'évaluation de l'acceptation d'un robot basée sur les interférences motrices [Kupferberg 2011]. Heerink et al. ont étudié l'influence de différentes capacités sociales d'un robot sur son acceptation par des personnes âgées [Heerink 2006].

2.6 Discussion

Il est important de séparer les travaux en robotique sociale selon deux grandes familles d'objectifs :

1. La première approche cherche à concevoir des robots dont les mécanismes internes reproduisent aussi fidèlement que possible le fonctionnement social de créatures vivantes existantes, et plus particulièrement de l'homme, afin de pouvoir les étudier et ainsi mieux les comprendre [Brooks 1999][Adams 2000][Scassellati 2001][Breazeal 2002a]. Cette approche nécessite donc de transposer directement les mécanismes existants chez l'homme au robot, dans un souci de justesse et de validité du modèle.
2. La deuxième approche cherche, quant à elle, à concevoir des robots fonctionnels dont seul le comportement social perçu par les utilisateurs aura une importance. Le fonctionnement interne des mécanismes sociaux peut n'avoir aucun fondement biologique tant que le comportement du robot satisfait les attentes des utilisateurs [Ishiguro 2001][Dautenhahn 2002a][Eklundh 2003][Wada 2005].

Bien que pouvant bénéficier l'une de l'autre, ces deux approches amènent des contraintes de développement différentes. Comme expliqué au chapitre précédent, nous nous intéressons au rôle de l'interface dans l'enseignement conjoint de mots nouveaux associés à des objets visuels à un robot et nous souhaitons permettre l'utilisation de tels systèmes par des utilisateurs non nécessairement experts. Nous nous plaçons donc ici très clairement dans la deuxième approche, où le modèle interne d'interaction sociale importe peu, mais où il sera par contre critique que l'interaction soit perçue par l'utilisateur comme satisfaisante et efficace. Nous pouvons également nous autoriser à contourner un certain nombre des problèmes rencontrés lors de la transposition des interactions humaines à l'interaction homme-robot, tant que l'interaction sociale reste intuitive et efficace pour l'utilisateur.

4. `http://www.phriends.eu/`

FIGURE 2.2 – On peut discerner deux approches majeures en robotique sociale : la première cherche à développer des robots reproduisant aussi fidèlement que possible le modèle social de l'humain afin de mieux le comprendre (illustré par le robot Kismet à gauche de l'image (source [Breazeal 2002a])), et la seconde cherche à développer un robot qui satisfasse effectivement les attentes sociales de l'utilisateur, sans se préoccuper de la plausibilité du modèle social utilisé (comme le robot Paro à droite de l'image (source [Wada 2005])).

Les mécanismes de perception et la technologie actuelle ne permettent pas de reproduire suffisamment fidèlement les mécanismes humains de l'interaction sociale pour pouvoir interagir avec un robot de la même manière que nous interagissons avec d'autres humains. En particulier, les interactions humain-robot transposées des interactions humaines ne sont pas, à l'heure actuelle, suffisamment robustes pour permettre une interaction fluide et naturelle en milieu ouvert. Comme souligné par Fong et al. il n'est pas nécessaire que le robot soit aussi compétent socialement que l'humain, mais il doit simplement être compatible avec les attentes et besoins des utilisateurs [Fong 2003a]. En particulier, les interactions sociales nécessaires dépendent très fortement de la tâche à accomplir. Un robot aspirateur n'aura, par exemple, pas besoin de maintenir des interactions sociales complexes. Par contre, il devra pouvoir proposer une interface qui permette à l'humain et au robot de se focaliser sur la tâche en elle même, plutôt que sur l'interaction. Partant de ce constat, les interactions et interfaces proposées et étudiées dans ce travail de thèse ne cherchent pas à reproduire fidèlement les interactions humaines, mais plutôt à permettre une interaction intuitive et efficace pour l'utilisateur, qui s'avérera dans notre contexte d'utilisation, plus pertinente qu'une interaction a priori plus directe.

Chapitre 3

L'acquisition des premiers mots du langage chez les robots : la question de l'interface

Sommaire

Résumé du chapitre

Nous nous intéressons ici au problème de l'acquisition des premiers mots du langage chez les robots et plus précisément au rôle de l'interface. Nous nous restreindrons à un cas particulier : l'enseignement conjoint de mots nouveaux associés à des objets visuels nouveaux à un robot. Nous identifierons les principaux défis de l'interaction humain-robot dans ce cadre et présenterons une revue critique des différentes interfaces utilisées. Nous conclurons en décrivant l'approche technique envisagée dans cette thèse pour s'attaquer à ces défis.

3.1 L'acquisition des premiers mots du langage chez les robots

L'acquisition du langage chez les robots est un domaine de recherche actif depuis de nombreuses années. Ce défi représente en effet un enjeu majeur pour la modéli-

sation et une meilleure compréhension de l'acquisition du langage chez l'humain. De plus, il joue aussi un rôle clé dans la conception de systèmes fonctionnels permettant à un humain d'interagir naturellement avec un robot. Cependant, doter un robot des mécanismes nécessaires à l'acquisition du langage soulève un grand nombre de défis scientifiques. Nous allons maintenant présenter plusieurs des aspects de l'acquisition du langage chez les robot, tels que la création de modèle statistiques et/ou informatiques de l'acquisition du langage, l'apprentissage du langage à travers l'interaction avec l'environnement et le problème de l'ancrage des symboles en perception.

Utiliser les robots comme une plateforme d'expérimentation permet d'étudier et de valider différents modèles d'acquisition du langage [Kaplan 2008]. Ces recherches ont entre autre permis une meilleure compréhension de l'évolution et de l'acquisition du langage chez l'homme et la validation de différents modèles du développement humain. Comme expliqué par Steels et Kaplan, le robot étant un agent physique percevant le monde à travers un appareil sensoriel, son utilisation empêche d'ignorer les problématiques et contraintes du monde physique [Steels 2002]. Le robot permet aussi à un humain d'être physiquement engagé dans l'interaction, afin d'étudier le caractère social de l'acquisition du langage [Tomasello 1999].

Des modèles statistiques et/ou informatiques ont été développés pour étudier et mieux comprendre les différentes étapes inhérentes à l'apprentissage du langage [Kaplan 2008]. La question de la découverte et de l'apprentissage de la segmentation des unités linguistiques de base (e.g. les phonèmes ou les mots) à partir d'un flux de parole a par exemple été largement étudiée. Une des méthodes consiste à extraire automatiquement les mots à partir d'énoncés en cherchant les motifs du signal sonore se répétant [Iwahashi 2003][Park 2008][Aimetti 2009]. L'approche NMF (*non-negative matrix factorization*) peut aussi être utilisée pour segmenter automatiquement les mots à partir d'énoncés [ten Bosch 2008]. La supervision de la segmentation des mots par d'autres modalités a aussi été explorée. Par exemple, Roy a étudié, à travers une installation robotique, la manière dont les nourrissons découvrent les unités linguistiques correspondant aux mots, et comment elles peuvent être automatiquement associées à des catégories sémantiques appropriées [Roy 1999]. Il a montré comment l'utilisation des entrées sensorielles de bas niveaux (son et image) permettait d'extraire automatiquement les unités linguistiques de base. Yu et Ballard ont aussi proposé un système capable de segmenter les mots à partir de la détection de co-occurrences d'énoncés vocaux et d'actions [Yu 2004a]. Ces unités linguistiques de base peuvent ensuite être utilisées pour construire des représentations du langage de plus haut niveau.

Des modèles de construction de lexiques ont aussi été proposés. La construction et classification lexicale permet de classer les associations entre des mots et des sens en différents groupes partageant un ensemble de similarités. Les groupes peuvent être construits directement en fonction de la distance entre les mots, mais aussi en regroupant les mots différents mais jouant un rôle similaire dans le langage. Niehaus et Levinson ont proposé un modèle statistique basé sur les modèles de Markov cachés pour permettre à un robot de se construire conjointement des lexiques de mots vocaux et d'actions [Niehaus 2012]. Vogt et Steels ont aussi montré comment

un lexique pouvait évoluer à travers les interactions entre des agents robotiques [Vogt 1998]. Des études ont montré comment un agent robotique pouvait apprendre un lexique, à travers l'imitation d'un autre robot [Billard 1997], ou à travers des instructions de l'utilisateur [Yanco 1993].

Le problème de conceptualisation, dit « Gavagaï », directement lié à la construction de lexique, est aussi une question de recherche centrale en acquisition du langage chez les robots [Quine 1960]. Ce problème s'intéresse à la question de l'ancrage des symboles en perception, c'est-à-dire comment les symboles acquièrent leur sens à travers l'association avec des données sensori-motrices (e.g. visuelles, actions), ainsi qu'à la question plus profonde de l'émergence de concepts [Steels 2000]. L'association de la parole avec d'autres modalités, notamment des informations visuelles, soulève le problème de la détection de co-occurrences entre les symboles et leur correspondance dans les différentes modalités utilisées. De plus, il faut déjà pouvoir identifier les invariants dans chacune des modalités (e.g. reconnaissance d'objets visuels). Yu et Ballard ont présenté un système d'apprentissage multi-modal (son et image) capable d'ancrer visuellement les noms d'objets visuels et de les reconnaître [Yu 2004a][Yu 2004b]. Roy et Pentland ont développé un système capable de segmenter les mots et de catégoriser des objets visuels à partir d'énoncés vocaux associés à des images d'objets [Roy 2000]. Cangelosi et al. ont présenté un modèle robotique du problème de l'ancrage des symboles et ont montré comment des capacités de haut niveau pouvaient être automatiquement construites à partir d'actions précédemment ancrées [Cangelosi 2006]. Pour une revue plus complète de la littérature du problème de l'intégration du langage et des actions, le lecteur pourra se référer à [Cangelosi 2010].

D'autres travaux se sont intéressés à la découverte et l'utilisation de structures grammaticales dans le discours. Iwahashi a par exemple utilisé une sémantique simple (objet, action, position) pour analyser les énoncés d'un humain à l'aide d'une structure de graphe [Iwahashi 2003][Iwahashi 2007]. Dominey et al. ont montré comment leur système pouvait apprendre la correspondance entre des phrases et leur sens directement à partir d'un flux vidéo de la scène ainsi que de l'enregistrement vocal. La position des classes de mots (e.g. noms, verbes, adjectifs) dans la phrase était utilisée pour analyser la structure grammaticale d'un énoncé [Dominey 2004a][Dominey 2004b]. Leur système d'acquisition du langage, compatible avec les observations de la trajectoire de développement de l'humain, permet de valider la cohérence de certaines observations faites par des études développementales.

Les modèles robotiques ont aussi permis l'étude de la question de l'émergence d'une langue et notamment le rôle des interactions avec l'environnement. Steels et Kaplan ont utilisé des agents robotiques dotés de modèles d'acquisition du langage afin d'étudier comment l'apprentissage social permet d'amorcer le développement de la communication [Steels 2002]. Oudeyer et Kaplan ont aussi étudié le rôle des motivations intrinsèques dans l'acquisition du langage [Oudeyer 2006].

Les recherches en acquisition du langage chez les robots s'intéressent aussi au développement de systèmes d'apprentissage social du langage qui soient effective-

ment fonctionnels et facilement utilisables par les humains. La validité biologique des modèles utilisés n'a donc ici pas d'importance en soi, seul le résultat perçu par l'utilisateur est pris en considération. Ces recherches sont cruciales pour favoriser le développement de la robotique personnelle, en permettant à l'utilisateur de guider un robot dans sa découverte du monde. La question de recherche étudiée dans cette thèse se situe très clairement dans ce cadre. Nous cherchons en effet à étudier le rôle de l'interface dans l'enseignement conjoint de mots nouveaux associés à des objets visuels nouveaux. Ce problème, qui peut sembler simple du point de vue de l'acquisition du langage, soulève néanmoins des questions fondamentales d'interaction homme-robot si la robustesse et l'intuitivité du système sont des critères primordiaux.

D'autres travaux se sont déjà intéressés à l'utilisation du langage naturel pour guider l'apprentissage d'un robot. Par exemple, Lauria et al. ont proposé un système permettant de programmer un robot afin qu'il parcoure une ville miniature à travers l'utilisation du langage naturel [Lauria 2002]. Dominey et al. ont montré comment le langage parlé pouvait être utilisé pour programmer une interaction en coopération avec un robot humanoïde [Dominey 2007][Dominey 2009]. Roy a également montré comment le langage naturel pouvait être utilisé pour enseigner des couleurs et des formes à un bras robotisé [Roy 2003]. Perzanowski et al. ont développé une interface humain-robot multi-modale, combinant la reconnaissance de la parole avec la reconnaissance de geste pour diriger un robot [Perzanowski 2001]. McGuire et al. ont, quant à eux, développé une interface multi-modale combinant les gestes de pointage avec la parole pour faire attraper des objets à un robot [Mcguire 2002]. Des méthodes d'identification des émotions dans le discours ont aussi été proposées et utilisées dans le cadre de l'interaction humain-robot (e.g. [Breazeal 2002b][Oudeyer 2003]).

3.2 Impact des facteurs humains et de l'interface

Comme nous venons de le présenter, le problème de l'acquisition sociale du langage chez les robots a été très étudié ces dernières années. Les problématiques de perception et d'apprentissage machine ont été particulièrement traitées. Cependant, les questions d'interactions telles que l'attention partagée ont aussi été identifiées comme un levier critique de l'apprentissage social du langage [Tomasello 1995][Kaplan 2004]. Ces mécanismes aident l'élève à focaliser son attention sur ce qu'il est important à apprendre [Steels 2000]. Les enfants acquièrent très tôt la capacité d'identifier les indices visuels, tels que les gestes de pointage, le regard ou encore les objets agités devant eux (voir la section 1.3.2 pour plus de détails).

Bien que fondamentales, ces questions ont souvent été contournées ou laissées de côté dans cette littérature. Nous pouvons par exemple remarquer que la plupart des expériences réalisées dans ce domaine ont été faites dans des conditions très contrôlées de type « laboratoire ». L'installation expérimentale est souvent contrainte : le robot est par exemple assis immobile face à une table sur laquelle sont disposés quelques objets [Scassellati 1996][Dominey 2004a]. Le protocole expérimental est

généralement lui aussi très précis et contraignant, rendant l'interaction inaccessible aux utilisateurs ne connaissant ni les mécanismes internes de l'apprentissage ni les limites du robot. Roy a, par exemple, présenté un système robotique d'acquisition de mots nouveaux où l'utilisateur doit placer un objet en face du robot puis le décrire [Roy 2003].

Ces expériences très contrôlées permettent d'évaluer les performances obtenues par un système d'apprentissage dans des conditions idéales. Cependant, ces méthodes utilisées en laboratoire peuvent-elles être transposées directement à une utilisation quotidienne dans un environnement réaliste ? Pour permettre de telles utilisations, ces systèmes doivent pouvoir satisfaire les critères suivants :

- Le système doit pouvoir fonctionner dans un environnement du quotidien — i.e. un environnement apriori non-contraint, inconnu et/ou changeant — et être utilisé intuitivement et de manière non-contraignante par des utilisateurs non-experts.
- Le système doit aussi pouvoir fonctionner avec un robot personnel ayant des capacités limitées (e.g. capteurs peu nombreux et bruités, petite taille).
- Enfin, le système doit pouvoir fonctionner même avec un nombre d'exemples d'apprentissage très réduit. En effet, même avec un système non-contraignant, il semble difficile de demander à un utilisateur de fournir plus d'une dizaine d'exemples d'apprentissage d'un même objet à un robot personnel.

3.3 État de l'art des interfaces pour l'apprentissage d'un lexique de mots et de sons nouveaux à un robot

Nous avons présenté à la section 2.3 une revue générale des interfaces humain-robot. Nous allons ici présenter plus particulièrement les avantages et faiblesses des interfaces utilisées, dans la littérature dans le cadre de l'acquisition du langage chez les robots.

3.3.1 Interactions directes

Comme expliqué précédemment, l'objectif principal de certains des travaux présentés ci-dessus est de développer des systèmes compatibles avec les modèles du développement humain. Ils essaient donc de transposer aussi précisément que possible les interactions humaines. Les modes de communications fréquemment utilisés pour ce type d'interaction sont :

- la reconnaissance de geste et notamment de pointage (e.g. [Scassellati 1996] [Hafner 2004][Haasch 2004])
- le suivi de regard (e.g. [Scassellati 1996][Breazeal 2002a][Atienza 2003] [Haasch 2004][Staudte 2009])
- la reconnaissance vocale et de l'intonation (e.g. [Roy 2003][Dominey 2007] [Dominey 2009])

En plus de permettre l'étude des modèles humains de l'acquisition du langage, ces interfaces sont aussi intéressantes pour le développement de systèmes pour la robotique personnelle. En effet, ces interfaces ne nécessitent pas d'apprentissage particulier, même pour les utilisateurs non-experts, et sont donc potentiellement très intuitives. Malheureusement, les techniques existantes de reconnaissance des gestes, de suivi de regard ou de reconnaissance de la parole ne sont pas à l'heure actuelle suffisamment robustes en environnement non-contrôlé (problèmes de bruit, d'occlusion ou encore de changement de lumière). Ce problème est particulièrement critique lors de l'utilisation de robots personnels dont la morphologie et l'appareil perceptif n'est généralement pas compatible avec ces modes d'interactions (capteurs bruités et de qualité limitée, angle de vue restreint, petite taille, etc.). De plus, les utilisateurs non-experts, directement concernés par la robotique personnelle, ne connaissent pas les limites intrinsèques de leur robot et sont donc d'autant plus prompts à faire de fausses suppositions quant aux capacités d'interactions du robot. Ces problèmes de robustesse peuvent mener à une interaction difficile et peu satisfaisante pour l'humain.

FIGURE 3.1 – A cause de leur manque de robustesse en milieu ouvert, les interactions directes nécessitent généralement l'utilisation d'une installation expérimentale contrainte et un protocole très strict limitant leur utilisation aux utilisateurs experts connaissant les limites du robot (source [Oudeyer 2006]).

En plus de dégrader l'expérience utilisateur, le manque de robustesse de ce type d'interaction a aussi été identifié comme une limite des différents systèmes d'apprentissage proposés. En particulier, il a été montré que les problèmes d'interface pouvaient entrainer la collecte de mauvais exemples d'apprentissage et ainsi diminuer les performances globales du système d'apprentissage [Kaplan 2005].

3.3.2 Agiter les objets

Une autre approche, fréquemment utilisée pour s'attaquer aux problèmes de pointage et d'attention partagée, est d'agiter directement les objets devant la caméra du robot [Lömker 2002][Wersing 2006]. Cela permet de facilement doter un robot de mécanismes attentionnels, où le robot focalise son attention sur les objets en mouvement. Cette technique permet aussi de facilement délimiter les frontières de l'objet par soustraction de l'arrière plan, i.e. la partie immobile de la scène. La segmentation d'objets est un problème difficile sans connaissance a priori des objets (voir section 6.5.2) et peut donc ici être efficacement contournée.

Bien que très intéressante, avec cette approche les utilisateurs ne peuvent que montrer à leur robot des objets de petites tailles et légers qui peuvent facilement être portés et agités en face du robot. Il n'est ainsi par exemple pas possible de montrer à son robot des objets tels qu'une table, une prise électrique ou encore un tableau accroché au mur. Par ailleurs, pour les personnes âgées ou les personnes handicapées, agiter des objets peut être une tâche fatigante, voire impossible.

3.3.3 Interfaces basées sur des objets médiateurs

Dans cette thèse, nous proposons d'étudier les interfaces basées sur des objets médiateurs, dans le cadre de l'apprentissage social du langage chez les robots (voir une revue de ces interfaces en section 2.3.2). Nous pensons que ces interfaces peuvent aider à résoudre les problèmes d'attention entre un humain et un robot, ainsi que contourner les problèmes de robustesse rencontrés lors des interactions directes, liés aux différences d'espace sensori-moteurs. En effet, pour pouvoir attirer, détecter et manipuler l'attention d'un agent, il est nécessaire d'être capable d'estimer précisément ses capacités et limites perceptives [Kaplan 2004]. Par exemple, pour évaluer quel objet un agent regarde, il est nécessaire de connaître la position et l'angle de vision de son(ses) capteur(s) visuel(s). Tous les humains étant dotés de capacités sensori-motrices similaires, l'homme a la capacité de se « projeter » dans le corps d'un autre humain afin d'estimer ce qu'il perçoit. Dans le cadre de l'interaction humain-robot, la difficulté pour l'humain d'évaluer les capacités réelles du robot rend cette projection complexe. Le robot Nao a, par exemple, deux yeux d'apparence similaire à ceux d'un humain. Pourtant, sa caméra a un angle de vue beaucoup plus restreint que le champ de vision humain. Si l'interaction/interface ne permet pas à l'utilisateur d'estimer facilement le champ de vision d'un robot, il devra apprendre à l'évaluer avant de pouvoir interagir de manière robuste avec lui.

Nous pensons que l'utilisation d'objets médiateurs peut aider l'utilisateur à interpréter les capacités du robot et à canaliser l'interaction pour la rendre plus robuste. Cette idée a déjà été appliquée à la communication entre un humain et certains singes bonobos, où des livres de symboles ont été utilisés pour faciliter les situations d'attention partagée. Cette interface médiatrice permettait notamment de faciliter le problème du pointage d'un objet. L'humain pouvait indiquer directement l'objet sur un clavier de symboles présenté au singe [Rumbaugh 1996]. Nous proposons

ici d'introduire des interfaces médiatrices humain-robot afin d'aider l'humain à se projeter dans le corps de son robot, et ainsi rendre l'interaction plus intuitive, effectivement utilisable et robuste. De plus, contrairement à la méthode où l'utilisateur agite les objets face au robot, ces interfaces ne sont pas contraintes par la taille ou le poids des objets à enseigner, et pourront donc être utilisées avec tous les types d'objets.

En plus d'aider l'utilisateur à atteindre des situations d'attention partagée avec un robot, nous pensons que ce type d'interface peut également permettre à l'humain d'aider le robot à se construire des représentations des mots qu'il apprend, à les reconnaître et à les regrouper en catégories. Le robot pourrait ainsi, à travers l'interface, tirer profit des capacités de l'humain, et améliorer son apprentissage.

Ces interfaces présentent, selon nous, un intérêt potentiel très fort pour le développement de la robotique personnelle, et plus particulièrement dans le cadre de l'enseignement de mots nouveaux associés à des objets nouveaux qui nous intéresse ici. Bien qu'apparemment plus complexe et plus contraignante à utiliser que la transposition des interactions humaines, nous pensons que ce type d'interface, en favorisant une interaction robuste, efficace mais aussi distrayante, pourrait, au final, être perçu comme plus facile à utiliser, plus efficace et moins contraignant par les utilisateurs dans le cadre de l'apprentissage social chez les robots personnels.

3.4 Restriction aux instances d'objets visuels

Dans sa généralité, le problème de l'acquisition sociale des premiers mots du langage chez les robots soulève de très nombreux challenges décrits au début de ce chapitre. Parmi ces défis, se pose en particulier le problème de savoir comment le robot peut inférer la signification d'un mot nouveau à partir de différents exemples. Ce défi est communément appelé le problème « Gavagaï » [Quine 1960]. Quine décrit dans son ouvrage, l'exemple d'un explorateur qui entendrait un natif prononcer le mot « Gavagaï » en désignant un lapin. Pour l'explorateur, ce mot pourrait signifier « lapin », mais aussi « blanc » ou encore « allons chasser ! ». Dans cette thèse, nous nous restreignons à un sous-problème de l'apprentissage du langage, afin d'éviter ces problèmes de conceptualisation. Nous cherchons ici à développer un système qui permette à un utilisateur d'associer des mots nouveaux à des instances d'objets visuels : i.e. les mots (des étiquettes comme au chapitre 6 ou des mots vocaux comme au chapitre 9) seront associés à une représentation visuelle d'un objet. Par exemple le mot « balle » sera associé à une balle d'aspect visuel particulier et non pas avec toutes les balles. Nous nous limiterons à des mots seuls et ne nous intéresserons pas non plus à l'acquisition de construction grammaticale.

Ce que nous appelons ici « objet visuel » correspond à une région d'une image ayant des propriétés visuelles caractéristiques. Cette définition très générale inclut donc en particulier les objets physiques visuels du quotidien (e.g. un jouet, des clés, une balle) mais inclue également des objets plus abstraits tels qu'une prise électrique, un dessin sur un mur, une cage d'escalier ou encore une porte ouverte. Il nous semble

en effet très important de pouvoir enseigner des mots nouveaux associés à tous ces types d'objets visuels à un robot. Par exemple, l'utilisateur pourrait souhaiter montrer une cage d'escalier à un robot afin de lui indiquer de ne pas s'en approcher.

En nous limitant à l'association de mots nouveaux avec des instances d'objets visuels nouveaux, nous évitons les problèmes de conceptualisation et ne devons traiter qu'un problème de reconnaissance visuelle d'objets. Cette limite nous permet ainsi de nous focaliser sur les questions d'interactions et d'interfaces centrales du travail de cette thèse.

3.5 Classification d'objets visuels

L'association de mots nouveaux avec des instances d'objets visuels est une problématique bien connue et très étudiée dans les domaines de la perception visuelle et de l'apprentissage machine [Sivic 2003][Lazebnik 2005][Bouchard 2005]. Ce problème peut être considéré comme un problème d'apprentissage supervisé, où une base de données d'images étiquetées (c'est-à-dire que pour chaque image un identifiant est associé, qui correspond à la classe d'objets à laquelle elle appartient) est utilisée pour entrainer un système d'apprentissage. Ce système sera ensuite utilisé pour prédire l'appartenance d'une nouvelle image à une classe d'objets. Le but étant alors d'essayer de maximiser les performances en généralisation de ce classifieur.

Différentes approches ont été proposées pour s'attaquer à ce problème. L'approche dominante à l'heure actuelle est l'approche dite par « sac-de-mots visuels » (*visual bags-of-words*) [Sivic 2003][Csurka 2004]. Cette technique, issue de la classification automatique de texte, consiste à extraire d'une image un ensemble de points caractéristiques (e.g. les coins, les contours) et à calculer une description locale de l'image pour chacun de ces points. Parmi les descripteurs les plus couramment utilisés, nous pouvons citer les histogrammes de couleurs, les histogramme d'orientation du gradient [Schiele 1996][Mikolajczyk 2003][Lowe 2004][Bay 2008] ou encore la description du contour. Les descripteurs extraits sont ensuite quantifiés à l'aide d'un dictionnaire et sont alors appelés des mots. L'ensemble des mots extraits d'une image seront utilisés pour la représenter. Dans ces approches, la notion de structure est aussi souvent supprimée. Il est possible de combiner plusieurs types de descripteurs, afin d'améliorer la robustesse du système. Différentes méthodes statistiques telles que l'utilisation de poids favorisant les mots les moins fréquents [Sivic 2003], les machines à vecteur de support (*Support Vector Machine SVM*) [Pontil 1998] ou encore le *boosting*, peuvent ensuite être utilisées pour entrainer le classifieur. Il existe des variantes de l'approche par sac-de-mots visuels, prenants en compte la structure entre les points extraits [Lazebnik 2009] ainsi que des implémentations incrémentales [Filliat 2007]. Le contexte peut aussi être utilisé pour améliorer la robustesse de la classification. Par exemple, une voiture aura une plus grande chance d'être détectée sur une route que dans un arbre [Schmuedderich 2010].

Une autre approche cherche à reconnaître les objets comme un ensemble de sous-parties, reliées entre elles selon une structure particulière. Cette approche est

surtout utilisée pour classifier des objets visuels spécifiques comme des voitures ou des visages [Bouchard 2005]. Les classes d'objets sont ici modélisées par une représentation de la structure topographique reliant les différentes parties ainsi que par une description de l'apparence de chacune des sous-parties.

Des bases de données d'exemples ont été constituées afin de pouvoir évaluer et comparer les performances en classification de ces différentes approches. La base de données PASCAL VOC (*Pattern Analysis, Statistical Modelling and Computational Learning : Visual Object Classes*) [Everingham 2010] est, par exemple, constituée de 20 classes d'objets (personnes, animaux, véhicule, intérieur) et de plus de 30 000 images. Ces images représentent ces différents objets dans des scènes réalistes et non-segmentées. La moitié de ces images peuvent être utilisées pour l'entrainement et l'autre moitié pour l'évaluation. Comme nous pouvons le voir, ces bases de données sont de taille très importante et l'apprentissage peut s'effectuer sur un très grand nombre d'exemples. Les algorithmes décrits précédemment supposent donc généralement l'existence de nombreux et « bons » exemples d'apprentissage. Pourtant les performances du système d'apprentissage sont directement dépendantes de la qualité et du nombre des exemples utilisés pour l'entrainer. Or, la constitution de ces bases de données d'exemples n'est pas en général étudiée en tant que défi scientifique.

Ici, nous cherchons à fournir à un robot la capacité de reconnaître n'importe quel objet visuel que l'humain souhaitera lui enseigner. Nous ne pouvons donc pas utiliser une base de données pré-construite. Les exemples utilisés pour l'apprentissage devront donc être collectés par le robot via les interactions avec l'humain. De part cette contrainte, le nombre d'exemples sera probablement très limité (de l'ordre de la dizaine d'exemples par objet).

3.6 Les défis de l'interaction humain-robot pour l'enseignement conjoint de mots et d'objets visuels nouveaux

Pour pouvoir résoudre le problème de l'association de mots nouveaux avec des objets visuels nouveaux dans le contexte de la robotique personnelle, il faut fournir la capacité aux utilisateurs de faire collecter à un robot les exemples d'apprentissage qui serviront à construire la base de données d'apprentissage nécessaire aux bonnes performances des méthodes de classification d'objets visuels décrites ci-dessus. En particulier, il faut déjà pouvoir répondre aux questions suivantes :

- Comment permettre à un humain non-expert de guider son robot personnel et lui faire collecter des exemples d'apprentissage d'objets visuels nouveaux ?
- Comment pousser les utilisateurs à fournir des exemples d'apprentissage de la meilleure qualité possible ? Les performances du système d'apprentissage sont en effet directement dépendantes de la qualité des exemples qui lui sont fournis. Cette question sera d'autant plus critique dans un contexte d'utilisation où l'utilisateur devra lui-même collecter les exemples d'apprentissage et

en fournira donc un nombre très limité (probablement moins de dix exemples par objet, là où les méthodes standards de classification d'objets visuels en utilisent plusieurs centaines voire plusieurs milliers).
- Comment parvenir à faire collecter ces exemples d'apprentissage à un humain à travers des interactions intuitives et peu contraignantes ?
- Comment reconnaître les mots/sons associés aux objets si on ne suppose pas l'utilisation d'étiquettes ou de symboles ?

Ces questions soulèvent les défis d'interactions suivants :

Attirer l'attention

Pour qu'un humain puisse montrer des objets visuels nouveaux à un robot, il faut tout d'abord qu'il ait la capacité d'attirer l'attention du robot. Cette question d'apparence très simple peut déjà poser des problèmes de robustesse du fait des différences d'appareil sensori-moteur existantes entre l'humain et le robot. Comment attirer l'attention d'un robot qui a un appareil sensori-moteur potentiellement très différent du notre ? Comment attirer son attention quand il est occupé à ses propres activités ? Comment un humain peut-il savoir lorsqu'un robot le voit ou l'entend ? Comment attirer son attention quand l'utilisateur est en dehors de son champ de vision ? Est-il préférable d'attirer l'attention du robot vers soi ou vers l'interaction elle-même ?

Parvenir à attirer l'attention d'un robot n'est pas suffisant, il faut aussi que cette interaction soit robuste et efficace mais aussi intuitive, afin de ne pas dégrader l'expérience utilisateur et donc ne pas le décourager. Comment fournir à l'humain un système d'attraction de l'attention du robot qui soit à la fois robuste, efficace et intuitif ?

Pointer

Une fois parvenue à attirer l'attention d'un robot, l'humain doit pouvoir indiquer/pointer les objets visuels qu'il veut lui montrer. Comment désigner un objet à un robot, et plus particulièrement comment peut-il y parvenir quand l'objet est hors du champ de vision du robot ?

Le pointage doit être robuste et précis afin d'éviter les fausses détections qui pourraient amener à l'apprentissage d'un mauvais exemple (voir la figure 3.2). Comment permettre à un humain non-expert de désigner de manière précise et robuste un objet visuel nouveau à un robot ? Il faut aussi que cette interaction soit fluide et intuitive. Après avoir correctement interprété la direction du pointage, comment le robot peut-il séparer l'objet visuel de son arrière-plan, sans fond neutre et sans modèle apriori des objets (problème de la segmentation : voir section 6.5.2) ?

Attention partagée

Pour permettre à l'enseignement d'objets visuels nouveaux à un robot d'être robuste et efficace, il est nécessaire de mettre en place des mécanismes d'attention

FIGURE 3.2 – Pour permettre à des utilisateurs de désigner un objet particulier dans un environnement complexe, il est nécessaire de fournir un mécanisme de pointage robuste et précis, afin d'éviter que l'interaction soit contraignante et/ou qu'elle entraine la collecte de mauvais exemples d'apprentissage.

partagée (présentés à la section 1.3.2). En particulier, l'humain et le robot doivent être capable de détecter et manipuler l'attention de l'autre. Ils doivent aussi pouvoir se coordonner. Ici, nous allons principalement étudier comment permettre à un robot de savoir précisément sur quoi se porte l'attention de l'humain, et comment permettre à l'humain de savoir précisément sur quoi se porte l'attention du robot : pour cela, il faut permettre à l'humain, et spécialement s'il est non-expert, d'appréhender correctement l'appareil sensori-moteur du robot.

Nommer

Une fois l'attention de l'humain et du robot focalisée sur le nouvel objet visuel, il faut que l'humain puisse nommer cet objet. Plus précisément, comment permettre à un utilisateur d'introduire un symbole à associer à cet objet ? Quelle forme de symbole doit être utilisée pour permettre au robot de le percevoir, l'enregistrer et le reconnaître par la suite ? Dans le cas de mots vocaux comment reconnaître quand deux mots correspondent à un même objet ? Comment reconnaître des mots vocaux identiques prononcés par des locuteurs différents ? Quelle représentation faut-il utiliser pour pouvoir représenter les mots vocaux de manière robuste ?

Découverte de catégories et association

Les exemples d'apprentissage doivent pouvoir être conjointement groupés dans les domaines visuels et acoustiques. Ce problème est particulièrement difficile si on ne suppose pas l'utilisation de symboles ou de catégories pré-existantes dans aucune de ces deux modalités. Comment reconnaître quand deux exemples se rapportent à un même objet ? Dans le cas où il n'existe pas de comparaison sûre des exemples d'apprentissage, comment permettre à un humain de participer à la clusterisation des exemples d'apprentissage et ainsi lever les doutes existants ?

Une fois ces catégories créées, il faut pouvoir construire des associations entre les catégories visuelles et auditives. Comment représenter, stocker et réutiliser ces associations par la suite ?

Rechercher

Enfin, après une phase d'apprentissage, comment permettre au robot de réutiliser ses nouvelles connaissances ? En particulier, nous souhaitons permettre à l'humain de demander à un robot de rechercher (ou simplement montrer) un objet visuel appris auparavant, et il faut donc notamment pouvoir reconnaître un mot déjà appris. Comment y parvenir à travers une interaction qui soit simple et intuitive ?

Nous pouvons aussi nous demander si les commandes de recherche peuvent également être utilisées pour obtenir de l'information supplémentaire de la part de l'utilisateur. En particulier, comment la recherche d'objets visuels peut-elle être utilisée pour obtenir de nouvelles associations fournies par l'humain de manière transparente ?

3.7 Approches techniques envisagées

Dans cette thèse, nous proposons d'utiliser des interfaces basées sur des objets médiateurs, comme celles décrites en section 3.3.3, pour s'attaquer aux problèmes d'interaction présentés ci-dessus. Plus précisément, nous pensons que ces objets peuvent faciliter l'attraction, l'évaluation et la manipulation de l'attention d'un robot par un humain non-expert, ainsi que l'aide à la perception et à la construction de représentations internes adéquates, lors de l'apprentissage conjoint d'objet visuels et de mots acoustiques. Nous proposons, par exemple, d'utiliser un terminal mobile communicant, sur lequel nous affichons le flux vidéo issu de la caméra d'un robot, permettant ainsi à l'utilisateur de savoir précisément ce que le robot voit. Le pointage d'objet visuel peut ensuite être effectué via l'écran tactile. Nous souhaitons aussi étudier l'utilisation d'un pointeur laser pour désigner des objets à un robot. En particulier, nous souhaitons étudier différents types de retour fournis à l'utilisateur de ce que perçoit un robot. Nous proposons aussi d'utiliser une commande spécifique pour demander au robot de collecter un exemple d'apprentissage (prendre une photo de l'objet et enregistrer le mot à associer) et ainsi faciliter les situations d'attention partagée.

Nous envisageons également l'utilisation d'une approche standard par sac-de-mots visuels (voir section 3.5) pour représenter et reconnaître les objets visuels. Une implémentation incrémentale nous permettrait d'enseigner n'importe quel type d'objet visuel à un robot, et de mettre à jour la représentation interne du robot à chaque nouvel exemple. Nous souhaitons aussi développer notre propre système de reconnaissance de mots vocaux. En effet, les systèmes « *off-the-shelf* » existants éprouvent généralement des difficultés lors de la reconnaissance de mots isolés dans des conditions non-contraintes avec du bruit environnant. Nous proposons de développer un système permettant à l'humain d'utiliser n'importe quelle langue et même n'importe quel son (par exemple des onomatopées). Pour cela, nous utiliserons directement des descripteurs de la parole afin de représenter et comparer les mots nouveaux associés aux objets visuels. Nous nous intéresserons particulièrement au couplage de cette mesure de similarité à l'interface, permettant à l'utilisateur de participer au processus de reconnaissance vocale et d'en améliorer la robustesse.

Ce système complet — interface, perception et apprentissage — permettra, d'une part d'avoir un système effectivement utilisable par des humains non-experts dans des conditions réelles, et d'autre part d'évaluer et de comparer qualitativement et quantitativement le rôle des différentes interfaces sur la qualité des exemples collectés, et donc sur les performances du système dans son ensemble.

Chapitre 4

Démarche « centrée utilisateur »

Sommaire

Résumé du chapitre

Dans ce chapitre, nous allons présenter la démarche « centrée utilisateur » suivie tout au long du travail de cette thèse. Cette approche permet de s'assurer de l'utilisabilité réelle d'une interface. Pour cela, nous présenterons, tout d'abord, une analyse du contexte d'utilisation (robotique personnelle, utilisateurs non-experts et environnement du quotidien). Ensuite, nous décrirons le procédé de développement et d'évaluation utilisé ici. Nous décrirons en particulier les méthodologies et critères d'évaluation intéressants dans le cadre de notre travail sur l'enseignement conjoint de mots nouveaux associés à des objets visuels à un robot.

4.1 Cycle de design itératif « centré utilisateur »

Comme indiqué précédemment, l'objectif de cette thèse est d'étudier le rôle de l'interface dans l'enseignement conjoint de mots associés à des objets visuels nouveaux à un robot. Plus précisément, nous allons étudier ce rôle dans un système permettant à des utilisateurs d'interagir avec un robot personnel, de lui montrer des objets du quotidien et finalement d'enseigner des mots nouveaux pour ces objets. Nous cherchons à développer un système qui soit effectivement utilisable par

des humains non-experts dans des conditions « réalistes » hors laboratoire, c'est-à-dire que notre système doit pouvoir être utilisé sans entrainement particulier et dans n'importe quel environnement domestique. Afin de nous assurer de remplir ces conditions, nous avons adopté une démarche « centrée utilisateur ».

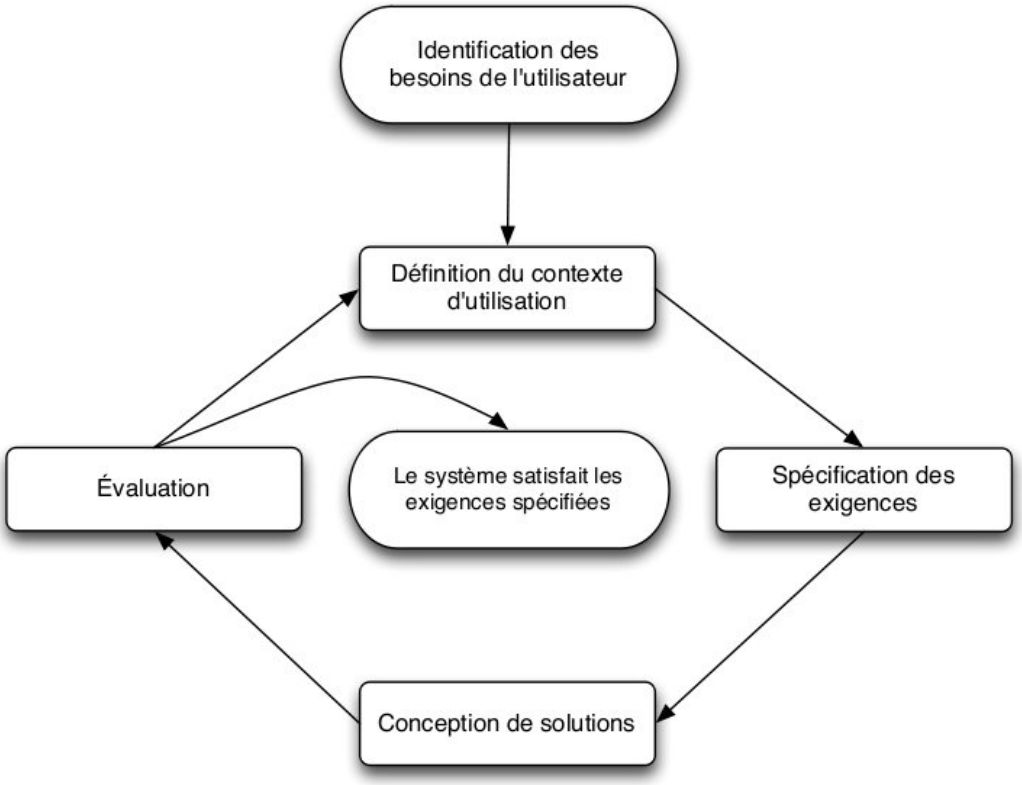


FIGURE 4.1 – Le modèle de conception centré utilisateur décrit par la norme ISO 13407. Chaque produit développé est testé auprès d'utilisateurs spécifiques correspondant aux conditions réalistes d'utilisation.

L'idée de base d'une démarche « centrée utilisateur » a été décrite par la norme ISO 13407. Ce modèle correspond particulièrement à nos besoins, il met l'utilisateur au centre d'un cycle de développement s'assurant ainsi de l'utilisabilité réelle du système. Une représentation graphique de ce modèle peut être vue sur la figure 4.1. Ce modèle peut être résumé en trois grandes étapes :

- L'analyse du contexte
- Le développement
- L'évaluation

Ce cycle peut être réitéré jusqu'à ce que l'évaluation donne des résultats satisfaisants.

Chacune de ces étapes doit être adaptée à notre contexte particulier, comme nous allons le détailler dans les sections suivantes.

4.2 Analyse du contexte

Enseigner des mots associés à des objets visuels nouveaux à un robot social est une tâche très particulière. En effet, cette tâche n'est pas à l'heure actuelle effectuée par des utilisateurs dans leur vie quotidienne. Cependant, comme montré dans le chapitre 1, de nombreuses analyses semblent indiquer que les robots personnels vont massivement intégrer notre quotidien dans le futur et cette interaction pourrait donc devenir fréquente. Cependant, à l'heure actuelle nous ne pouvons que faire des hypothèses sur les futures conditions d'utilisations. Dans cette section, nous allons essayer de répondre aux questions suivantes :

- Quel utilisateur pour quel usage ?
- Dans quel contexte ? Où ? Quand ?

4.2.1 Utilisateurs non-experts

Il est important d'identifier quelles seront les personnes qui utiliseront notre système et dans quel but. Nous nous plaçons ici dans un contexte de robotique personnelle (voir le chapitre 1) où c'est bien « monsieur tout le monde » qui aura à interagir avec des robots au quotidien. La plupart des utilisateurs ne seront donc pas experts, c'est-à-dire qu'ils ne devraient pas avoir besoin d'entrainement spécifique ni de formation particulière pour pouvoir interagir avec leur robot. De la même manière que la plupart des utilisateurs ne souhaitent pas avoir besoin de lire un manuel d'instructions avant de pouvoir utiliser leur nouveau smartphone, ils ne souhaiteront probablement pas avoir besoin d'instructions complexes avant de pouvoir interagir avec un robot. Les interfaces humain-robot développées devront donc être faciles à prendre en main, intuitives mais également non-contraignantes voir distrayantes afin de maintenir les utilisateurs motivés.

Il semble également raisonnable de supposer que les utilisateurs ne connaîtront ni ne comprendront pas nécessairement la technologie utilisée à l'intérieur des robots. Ils feront donc des suppositions (vraies ou fausses) quant aux capacités du robot : e.g. si un robot est doté de ce qui ressemble à des yeux humains, les utilisateurs supposeront vraisemblablement qu'il aura un champ de vision comparable à celui d'un humain. Comme expliqué à la section 2.2, ces attentes des utilisateurs peuvent avoir un impact très important sur l'interaction et l'expérience utilisateur.

4.2.2 Robot personnel et social

Dans les expériences décrites dans la suite de cette thèse, nous avons utilisé des robots qui représentent bien à nos yeux la robotique personnelle et sociale actuelle et celle du futur proche : l'Aibo [1] et le NAO [2]. Ces robots ont un aspect de jouet et une forme zoomorphique ou anthropomorphique qui permet des interactions faciles avec le grand public. Ils sont aussi munis d'un appareil sensori-moteur relativement

1. `http://support.sony-europe.com/aibo/`
2. `http://www.aldebaran-robotics.com/`

complexe avec un grand nombre de degrés de liberté et des capteurs variés tels qu'une caméra ou un microphone. Cependant, ces capteurs sont souvent limités : e.g. les caméra présentes sur ces robots ont un angle de vue restreint et un taux de rafraîchissement faible. Enfin, les prochaines générations de ce type de robots pourraient donner naissance à des robots relativement bon marché et donc susceptibles d'être utilisés massivement dans le futur.

Nous n'avons volontairement pas utilisé de système permettant d'améliorer artificiellement les capacités de ces robots telles que des caméras grand angle ou des caméras accrochées au plafond. En effet, il est très important pour nous que notre système puisse fonctionner avec des solutions facilement envisageables et à bas coût à relativement court-terme. Les solutions basées sur des robots plus performants mais plus chers ou les solutions basées sur la domotique ne nous semblent pas facilement applicable à la robotique personnelle dans un futur proche.

Par eux mêmes ces robots sont statiques et inertes. Nous avons donc ajouté de petits comportements du robot amusants pour l'utilisateur mais lui permettant également de mieux comprendre ce que le robot faisait ou essayait d'exprimer. Par exemple, lorsqu'aucune commande n'était envoyée au robot pendant un laps de temps trop important, le robot baillait ou se grattait la tête (voir les exemples sur la figure 4.2).

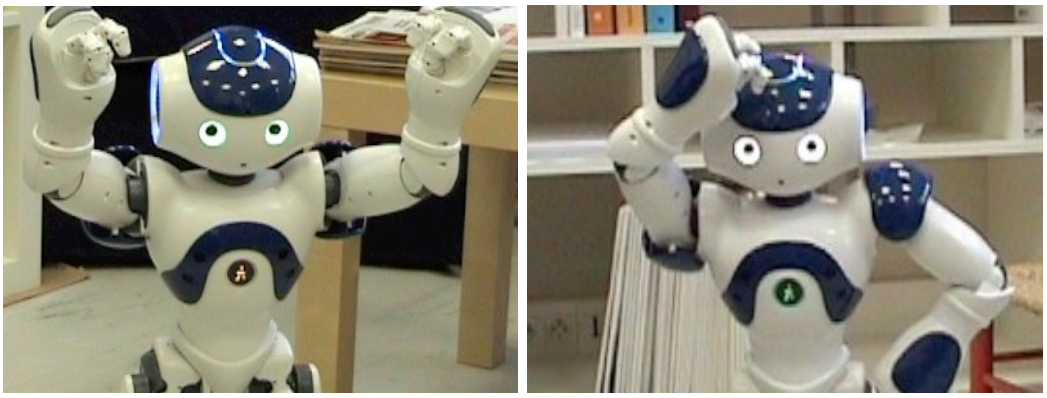


FIGURE 4.2 – Des comportements tels que « content » sur la gauche ou « se gratter la tête » sur la droite ont été ajoutés afin de rendre le robot plus vivant et d'aider les utilisateurs à mieux comprendre son comportement.

4.2.3 Environnement quotidien

Notre contexte de robotique personnelle impose que notre système puisse fonctionner dans un environnement du quotidien où il est très difficile d'imposer des contraintes ou restrictions. Les environnements domestiques sont notamment :

- a priori inconnus
- changeants
- non-contraints

– incontrôlés
– très variés
– chargés en objets divers
– avec un fond sonore ambiant

Nos algorithmes devront donc pouvoir fonctionner dans de telles conditions. De même, nous ne pouvons faire que peu de suppositions sur le type d'objets que les utilisateurs voudront montrer à leur robot. Aux côtés des objets du quotidien (e.g. un journal, un jouet ou un téléphone), il pourrait aussi être intéressant de montrer des objets très différents à un robot (e.g. une porte, des escaliers, une personne, un autre robot, etc.).

4.2.4 Co-location

Enfin, lors de ce type d'interaction (enseignant / élève) l'humain sera probablement toujours présent à côté de son élève robotique afin de pouvoir interagir directement et physiquement avec son environnement immédiat. Nous nous sommes donc focalisés dans cette thèse sur des interactions dîtes en « co-location », c'est-à-dire où le robot et l'humain sont physiquement présents l'un à côté de l'autre. Cependant, une des interfaces, celle basée sur l'iPhone présentée en section 6.3.1, sera également utilisable en télé-opération. Nous parlerons de cette possibilité d'interaction en détails dans la partie discussion et perspectives de cette thèse (voir section 10.5.2).

4.3 Développement

Le système étant entièrement nouveau, l'inclusion de l'utilisateur dans la boucle de développement est capitale. Afin de bien séparer les différentes questions de recherche abordées dans cette thèse (voir chapitre 3.6), nous avons choisi de les aborder incrémentalement. Ceci nous a permis de mieux identifier les difficultés inhérentes à chacune de ces problématiques et d'évaluer chacune des composantes indépendamment.

Comme il était nécessaire d'avoir rapidement un système complet afin de pouvoir valider notre approche et réaliser des premiers tests, une implémentation simplifiée des parties non-encore étudiées a parfois été utilisée. Plus de détails seront donnés dans les chapitres suivants.

4.4 Évaluation

4.4.1 Protocole expérimental

4.4.1.1 Rendre la tâche réaliste

Comme expliqué précédemment, demander à des utilisateurs, non familiers des problématiques liées à la robotique, d'enseigner des mots nouveaux associés à des objets visuels nouveaux à un robot est une tâche artificielle. Les premières expériences

pilotes que nous avons effectuées ont révélé que les utilisateurs se démotivaient rapidement si la tâche demandée n'était pas justifiée concrètement. Nous avons donc utilisé une mise en scène permettant de rendre cette tâche réaliste et sensée mais nous avons également essayé de rendre nos expériences amusantes pour les participants afin qu'ils restent concentrés et motivés sur ce qui leur était demandé tout au long des tests.

4.4.1.2 Reproductibilité

Un grand nombre de participants était nécessaire afin d'obtenir des statistiques significatives sur l'utilisabilité de nos interfaces et il était donc nécessaire de réaliser des expériences à grande échelle. Pour cela, il a fallu mettre en place des protocoles expérimentaux simples et reproductibles permettant de répéter la même expérience un grand nombre de fois et dans les même conditions.

4.4.2 Critères d'évaluation

Il est toujours difficile de trouver les métriques permettant d'évaluer réellement la pertinence, la qualité et l'utilisabilité d'une interface ou bien encore le ressenti global de l'expérience. En nous basant sur la littérature existante en interaction homme-machine, et surtout en interaction homme-robot, nous avons extrait deux grands domaines d'évaluation pertinents pour notre étude [Yanco 2004b][Walters 2005] :

4.4.2.1 Utilisabilité réelle et perçue

L'utilisabilité telle que définit par Nielsen s'intéresse à l'évaluation d'une interface pour réaliser une tâche [Nielsen 1994]. Cette définition regroupe des notions telles que l'efficience, l'erreur ou la facilité d'apprentissage tout à fait pertinentes dans notre contexte.

Il est important de bien séparer l'utilisabilité réelle de l'utilisabilité perçue. La première est évaluée à travers des mesures quantitatives comme le temps mis pour réaliser une tâche, le nombre d'erreurs ou encore la réussite. La seconde provient de mesures qualitatives où c'est l'utilisateur lui même qui décrit son ressenti vis à vis de l'utilisation de l'interface. Elle peut être mesurée à l'aide de questionnaires ou d'entretiens. Ces deux notions sont importantes pour nous. En effet, il est bien sûr important que nos interfaces permettent réellement à un utilisateur d'enseigner des mots nouveaux associés à des objets visuels à un robot, mais il est tout aussi important qu'il y parvienne, de son point de vue, facilement et efficacement.

Une des mesures essentielles à ce travail, sur l'étude du rôle de l'interface dans l'enseignement conjoint de mots associés à des objets visuels nouveaux, est la qualité des exemples collectés par le robot. En effet, ces exemples serviront d'apprentissage au robot et auront donc une influence majeure sur les performances du système dans son ensemble. Pour évaluer la qualité des exemples acquis par le robot, c'est-à-dire les représentations visuelles des objets ainsi que les mots associés, ils seront enregistrés

et analysés qualitativement. Ils seront également utilisés comme base d'entrainement d'un système d'apprentissage, dont les performances en généralisation seront évaluées quantitativement à travers des tâches de classification.

4.4.2.2 Expérience utilisateur

Parvenir à utiliser efficacement et facilement nos interfaces n'est pas suffisant. Pour nous, il est également indispensable que les utilisateurs trouvent leur utilisation non-contraignante, ou même distrayante et amusante. En effet, il est probable que les utilisateurs délaisseront rapidement ces interfaces si leur utilisation est ennuyeuse. Weiss et al. ont défini des critères permettant d'évaluer l'expérience utilisateur lors d'une interaction avec un robot [Weiss 2009]. L'émotion, le ressenti durant l'expérience ou encore la perception de la collaboration permettent de caractériser l'expérience utilisateur. Ce type de données peut être collecté à travers des questionnaires.

Le tableau 4.1 récapitule les principaux critères d'évaluation utilisés dans la suite de ce rapport de thèse ainsi que les méthodes d'évaluation utilisées :

	Critère	Méthode d'évaluation
Utilisabilité	Efficience	Collecte d'exemples Mesures qualitatives Questionnaires
	Facilité d'apprentissage	
Expérience utilisateur	Émotion	Questionnaires
	Interaction sociale	

TABLE 4.1 – Les principaux critères d'évaluation utilisés pour caractériser, évaluer et comparer les interfaces développées dans le reste de cette thèse.

4.5 Les différentes étapes expérimentales

Dans les chapitres suivants, nous décrirons les étapes clés du processus itératif de développement/évaluation suivi tout au long du travail de cette thèse :

- Nous nous sommes tout d'abord intéressés au développement d'interfaces humain-robot pour la navigation (chapitre 5). Différentes interfaces, basées sur des objets médiateurs, ont été évaluées et comparées lors d'une étude utilisateurs, où des participants devaient faire traverser un parcours d'obstacle à un robot.
- Puis, nous avons étudié la question de l'enseignement d'objets visuels à un robot afin qu'il les reconnaisse. Les questions de pointage et d'attention partagée nous ont particulièrement intéressées (chapitre 6 et 7). Nous avons développé des interfaces permettant à un utilisateur de faire collecter des exemples d'apprentissage d'objets visuels à un robot. Nous avons utilisé ici des étiquettes

symboliques associés à ces images. Ces symboles ont été obtenus en simulant l'utilisation d'un système de reconnaissance vocale. Les interfaces ont été comparées lors d'une étude utilisateurs à grande échelle hors laboratoire.

– Nous avons ensuite proposé une approche possible pour la collecte semi-automatique d'exemples d'apprentissage d'objets visuels par un robot (chapitre 8). Cette approche a été évaluée à travers une expérience pilote réalisée dans notre laboratoire.
– Enfin, nous avons étudié la question de l'enseignement conjoint de mots vocaux nouveaux associés à des objets visuels nouveaux à un robot. Nous avons proposé un système de reconnaissance vocale, basé sur l'intégration d'une mesure de similarité entre les mots acoustiques et d'une interface bien conçue. Nous avons étudié le rôle de l'interface dans l'aide à la reconnaissance vocale ainsi que dans le regroupement et l'association de ces différents exemples d'apprentissage (chapitre 9). Nous avons simulé l'utilisation de cette interface, en utilisant des exemples collectés lors d'une expérience hors laboratoire précédente.

Chapitre 5

Interfaces pour piloter un robot

Sommaire

Résumé du chapitre

Nous présenterons, dans ce chapitre, différentes interfaces, basées sur des objets médiateurs, tels qu'un iPhone ou une Wiimote, et présentant des métaphores d'interaction variées. Elles ont été conçues pour étudier le problème de la navigation d'un robot dans un environnement du quotidien, mais avec l'objectif à plus long terme, de les utiliser pour l'enseignement conjoint de mots nouveaux associés à des objets visuels. Nous présenterons une étude utilisateur de ces interfaces, où les participants devaient guider un robot à travers deux parcours d'obstacles. Nous montrerons notamment que l'utilisabilité réelle et perçue de nos interfaces peuvent s'opposer. Une des interfaces, celle basée sur des gestes réalisés directement sur l'écran tactile d'un iPhone, sera en effet évaluée comme la moins efficace et perçue comme la plus efficiente par les utilisateurs.

5.1 Objectifs

Ce chapitre décrit la première itération complète de notre cycle de développement présenté au chapitre précédent. Lors de cette première étape, nous nous sommes

intéressés à la question suivante : Comment permettre à un utilisateur non-expert de diriger un robot à travers un environnement du quotidien de manière intuitive, efficace et non-contraignante ?

Pour enseigner des mots associés à des objets visuels nouveaux à son robot il faudra en effet commencer par guider le robot à travers un environnement domestique jusqu'à l'objet que l'on souhaite lui montrer. Cette question, simple en apparence, soulève déjà des problématiques complexes d'attraction de l'attention ainsi que de navigation qui seront développées dans le reste de ce chapitre. Ces résultats ont été publiés dans [Rouanet 2009a].

5.1.1 Attirer l'attention

Avant de pouvoir interagir avec le robot et le diriger effectivement il faut déjà être capable d'attirer son attention lorsque celui-ci est occupé à ses propres activités. Cette question, triviale du point de vue de l'interaction humaine, pose déjà des problèmes importants d'interaction humain-robot. La transposition des interactions humaines pour résoudre ce problème (e.g. suivi de regard, la reconnaissance des gestes ou de la parole) pose des problèmes de robustesse en milieu non-contraint (cf. la section 3.3.1 pour plus de détails) et ne permet donc pas une interaction intuitive et efficace.

Nous allons donc essayer de répondre aux questions suivantes :

- Comment un humain peut-il de manière robuste et intuitive attirer l'attention d'un robot vers lui-même ou bien vers l'interaction lorsque le robot est occupé à ses propres activités ?
- Comment parvenir à attirer l'attention du robot même lorsque l'humain n'est pas présent dans le champ de vision du robot ?

5.1.2 Navigation

Une fois parvenu à attirer l'attention du robot, l'utilisateur peut commencer à guider le robot jusqu'à un objet précis à travers un environnement potentiellement inconnu, changeant et encombré. En particulier, nous allons nous intéresser aux questions suivantes :

- Comment un utilisateur peut-il diriger efficacement et intuitivement un robot dans un environnement inconnu et/ou changeant ?
- Comment guider un robot à travers un environnement domestique encombré jusqu'à un objet particulier ?
- Comment déplacer facilement un robot doté d'un squelette complexe sans surcharger la charge de travail de l'utilisateur ?

Si on cherche ici à diriger un robot, le but à plus long terme est d'amener le robot à regarder un objet précis afin de le nommer. Ceci va influencer les critères d'évaluation de réussite de cet objectif : e.g. la trajectoire suivie pour atteindre le but n'aura pas une grande importance. Nos critères d'évaluation seront présentés et discutés dans la partie 5.3.2 de ce chapitre.

5.2 Interfaces développées

Afin de répondre aux questions décrites ci-dessus nous avons développé trois interfaces différentes basées sur des objets « médiateurs » (voir la figure 5.1). Comme expliqué dans la section 3.3, contrairement aux interactions directes qui souffrent de problème de robustesse, ce type d'interface permet aux utilisateurs non-experts d'attirer l'attention d'un robot personnel ainsi que de le guider de manière intuitive et robuste dans un environnement du quotidien. Le lecteur pourra se rapporter à la section 2.3.2 pour une revue des différentes interfaces basées sur des objets médiateurs utilisées en interaction humain-robot.

FIGURE 5.1 – L'utilisation d'interfaces basées sur des objets médiateurs permet à un utilisateur non-expert d'intuitivement et efficacement attirer l'attention de son robot ainsi que de le guider à travers un environnement du quotidien.

Nous avons essayé de développer trois interfaces très différentes afin d'explorer des types d'interaction variés et d'étudier l'impact des différentes métaphores sur la perception qu'en ont les utilisateurs. La métaphore d'interaction est la manière dont l'interface représente les actions nécessaires pour réaliser une tâche, ici faire naviguer un robot. Nous proposons, dans ce chapitre, différentes métaphores d'interaction à l'utilisateur : nous pouvons par exemple lui donner l'impression de piloter directement le robot, ou bien simplement lui faire indiquer une destination. Ces métaphores permettent de cacher la complexité du système robotique derrière des commandes de haut niveau. Elles se reposent généralement sur des interactions bien connues des utilisateurs. Cependant, en cachant une partie de la complexité

de la tâche nous influençons le modèle mental que l'utilisateur se fait du robot et de l'interaction qu'il a avec lui. Ce problème est particulièrement critique avec des utilisateurs non-experts (voir les sections 2.2 et 4.2.1 pour plus de détails).

Bien que différentes, les trois interfaces décrites ci-dessous présentent un ensemble de points communs :

- Elles sont portables afin de permettre à l'utilisateur de toujours avoir l'interface avec lui, de pouvoir suivre le robot (co-location) et d'interagir directement avec l'environnement s'il le souhaite.
- Elles sont basées sur des objets de la vie courante et reposent sur des techniques d'interaction « classiques » afin de permettre aux utilisateurs une prise en main rapide.
- Elles ne nécessitent pas d'équipement complexe ou cher.
- Elles doivent être non-contraignantes, voir amusantes, pour permettre des interactions sur le long terme sans lassitude.

Elles fournissent également les mêmes possibilités d'interaction :

- Faire avancer / reculer le robot
- Faire tourner le robot sur lui même
- Orienter sa tête

5.2.1 iPhone Go-To (IGT)

La première interface est basée sur un iPhone. Le flux vidéo de la caméra du robot est affiché sur l'écran de l'appareil. Cet écran tactile est aussi utilisé comme un « tableau interactif » : i.e. les utilisateurs peuvent dessiner, directement sur le flux vidéo, un ensemble de gestes correspondant à différentes commandes.

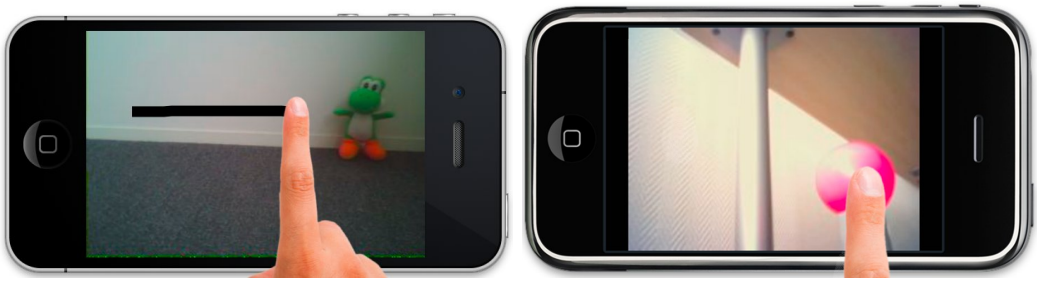


FIGURE 5.2 – Avec l'interface iPhone Go-To les utilisateurs peuvent définir des « trajectoires » directement sur le flux vidéo de la caméra du robot affiché sur l'écran de l'appareil afin de déplacer le robot ou de le faire regarder un endroit précis. Sur l'image de gauche, l'utilisateur effectue un trait vers la droite afin de faire tourner le robot sur lui même dans cette direction. Sur l'image de droite, en « tapant » sur un point de l'écran, l'utilisateur demande au robot d'orienter sa tête afin de centrer sa vision vers ce point.

Les utilisateurs peuvent faire avancer / reculer le robot simplement en dessinant un trait vertical sur l'écran de l'iPhone. Ils peuvent également le faire tourner sur

lui-même à l'aide de traits horizontaux. Des heuristiques très simples sont utilisées pour reconnaître ces différents gestes (e.g. utilisation de la distance par rapport à la droite passant par le premier et dernier point du trait). Ces traits peuvent être vus comme des sortes de trajectoires que le robot va suivre.

Pour orienter la tête du robot vers un point précis, il suffit de « taper » sur ce point à l'écran. Cette méthode permet de montrer très facilement un endroit particulier au robot. Cependant, on ne peut montrer qu'un point qui est déjà dans le champ de vision du robot afin de pouvoir « taper » dessus. Dans le cas contraire, il faut soit tourner le robot, soit orienter la tête du robot vers l'objet, de proche en proche, jusqu'à ce que l'objet soit dans son champ de vision.

Interagir directement sur l'écran permet aux utilisateurs d'envoyer des commandes plus « concrètes » et donc potentiellement plus faciles à prendre en main. Cependant, ce mode d'interaction entraine également quelques occlusions du flux vidéo affiché à l'écran.

Afficher le retour visuel de ce que voit le robot sur l'écran de l'iPhone permet de savoir très précisément et très facilement ce qui est dans le champ de vision du robot. Par contre, cela force l'utilisateur à séparer son attention entre l'attention directe où il regarde le robot et l'attention indirecte où il regarde ce que le robot voit à travers l'écran.

5.2.2 iPhone flèches directionnelles (IFD)

La deuxième interface est également basée sur un iPhone. Avec cette interface aussi, le retour visuel de ce que perçoit le robot est affiché sur l'écran de l'appareil. Par dessus ce flux vidéo sont dessinées quatre flèches directionnelles semi-transparentes (voir la figure 5.3). Ces flèches permettent à l'utilisateur de déplacer le robot (avant, arrière, tourner à gauche, tourner à droite). La métaphore utilisée ici est celle dite du « véhicule volant », connue notamment pour être la métaphore utilisée dans les jeux-vidéo à la première personne. Lorsque l'utilisateur appuie sur la flèche du haut, le robot avance. Dès que l'utilisateur relâche la flèche le robot s'arrête. Cette interface est donc très facile à comprendre et à prendre en main. Par contre, cette métaphore nécessite que l'utilisateur appuie constamment sur une flèche lorsqu'il veut déplacer le robot. Ceci peut être contraignant et génère également de nombreuses occlusions visuelles de l'écran.

La tête du robot peut être orientée en utilisant deux sliders représentant les valeurs de roulis et tangage de la tête du robot.

5.2.3 Wiimote TUI (WTUI)

Enfin, la troisième interface est basée sur le contrôleur Wiimote de la console Wii [1]. Cet appareil est doté d'un accéléromètre trois axes et permet donc de mesurer les mouvements effectués par la main de l'utilisateur. Le contrôleur Wiimote est utilisé ici comme une interface tangible (voir la section 2.3.2.3 pour plus de détails).

1. `http://www.nintendo.com/wii`

FIGURE 5.3 – Cette interface (IFD) permet à l'utilisateur de déplacer le robot à l'aide des flèches directionnelles dessinées sur l'écran ainsi que d'orienter la tête du robot à l'aide des deux sliders.

Les mouvements appliqués à l'appareil sont transformés en commandes envoyées au robot. Par exemple, lorsque l'utilisateur penche la Wiimote vers l'avant le robot avance, lorsqu'il tourne la Wiimote vers la gauche, le robot tourne également vers la gauche. De même, il est possible d'orienter la tête du robot en orientant le contrôleur. Pour séparer les mouvements du robot et les mouvements de sa tête, l'utilisateur doit presser un bouton en même temps qu'il oriente la Wiimote pour bouger la tête du robot.

Le contrôleur Wiimote offre une granularité de mouvements qu'il n'est pas possible de reproduire avec le robot. Les valeurs réelles fournies par la Wiimote sont discrétisées en 5 valeurs possibles : (avant, arrière, gauche, droite et zone morte). Cette transformation peut donc engendrer une certaine frustration pour les utilisateurs qui espèrent que le robot soit capable de reproduire l'ensemble des mouvements qu'ils font avec l'interface.

Ce type d'interface permet d'avoir une interaction très intuitive, ne nécessitant que très peu d'apprentissage. De plus, elle permet à l'utilisateur de toujours focaliser son attention directement sur le robot car il n'a pas besoin de regarder ce qu'il fait avec ses mains. Par contre, cette interface ne fournit aucun retour visuel sur ce que le robot perçoit.

5.3 Expérimentation

5.3.1 Protocole expérimental

Afin d'évaluer ces trois interfaces lors d'une interaction plausible, nous avons conçu une expérimentation basée sur un parcours d'obstacles. Pour cela, nous avons recréé un salon avec une table des chaises, un canapé, etc. (l'environnement recréé

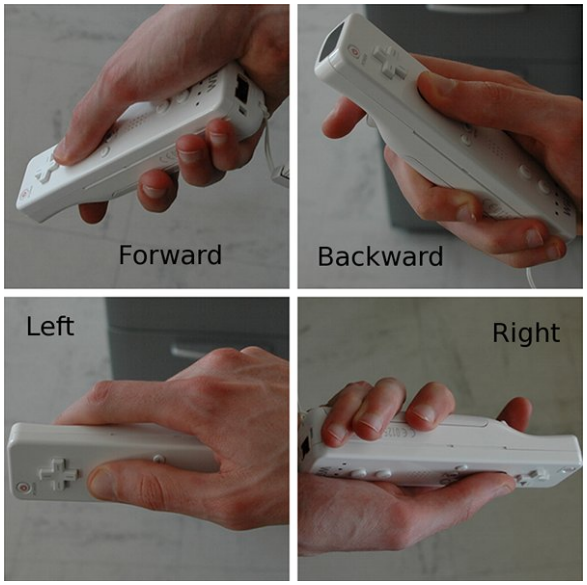


FIGURE 5.4 – L'interface basée sur la Wiimote permet aux utilisateurs, en orientant la Wiimote de déplacer le robot ainsi que d'orienter sa tête dans une direction particulière.

peut être vu sur la figure 5.5). Le robot devait traverser ce salon depuis la ligne de départ jusqu'à parvenir à montrer une balle rose (automatiquement détectée par le robot) qui représentait la fin du parcours. Un ensemble de portes, représentées par les lignes au sol, devaient être franchies par le robot afin de forcer les participants à lui faire prendre des virages serrés. Deux différents parcours ont été définis : un simple et un plus difficile avec des virages plus serrés. La tâche du parcours d'obstacle a été choisie car elle regroupe nos deux objectifs principaux :

- naviguer à travers un environnement domestique réaliste
- guider le robot jusqu'à un objet précis

Nous avons utilisé le robot Aibo pour cette expérience car il se prête bien à un parcours d'obstacles et représente également bien un robot personnel et social (une justification plus précise du choix du robot est donnée en section 4.2.2). De plus, c'est un robot à quatre pattes qui a donc une marche complexe ce qui force l'utilisation de commandes de haut niveaux. Cette marche était relativement lente. Le robot glissait parfois sur le sol rendant sa navigation plus difficile. Il est également important de noter qu'il y avait une latence très importante entre l'envoi d'une commande et son exécution effective ($\sim 500ms$). Cette latence bien que gênante pour l'expérience utilisateur était identique pour toutes les interfaces et n'était donc pas un facteur discriminant. De plus, un retour visuel immédiat, sous forme de leds, était visible directement sur le robot permettant à l'utilisateur de savoir s'il avait bien envoyé une commande. Comme expliqué dans la section 4.2.2, des comportements basiques (aboiements, remuer la queue...) ont été ajoutés afin de rendre l'interaction plus

FIGURE 5.5 – Un parcours d'obstacle a été conçu à travers un salon reproduisant un environnement domestique afin de comparer les trois interfaces développées.

vivante.

30 personnes (20 hommes et 10 femmes) ont participé à notre expérience. Ils ont été recrutés sur le campus de l'université Bordeaux 1. La plupart (20) faisait partie du personnel administratif de l'université et n'avait aucune connaissance particulière en robotique. Ils étaient âgés de 19 à 50 ans ($M = 28, STD = 8$). Chacun des participants a testé deux interfaces, effectuant à chaque fois le parcours facile et le parcours difficile. L'ordre d'utilisation des interfaces était tiré aléatoirement afin d'éviter l'effet d'habituation. Les participants pouvaient se déplacer librement dans la pièce pendant toute la durée de l'expérience.

5.3.2 Mesures

5.3.2.1 Quantitatives

Durant les expériences, des mesures quantitatives ont été relevées afin de comparer ces trois interfaces. Le temps de parcours total (pour le parcours facile et pour le parcours difficile) ainsi que le temps où le robot était réellement en mouvement étaient mesurés. Le nombre de commandes envoyé était également enregistré.

5.3.2.2 Qualitatives

Une fois les deux parcours réalisés, les participants devaient remplir un questionnaire visant à évaluer l'utilisabilité perçue des différentes interfaces. Les participants devaient évaluer un ensemble d'affirmations sur une échelle de Likert en 5 points. Cette échelle est une mesure très répandue dans les tests de psychologie, où les participants doivent exprimer leur degré d'accord vis-à-vis d'un ensemble d'affir-

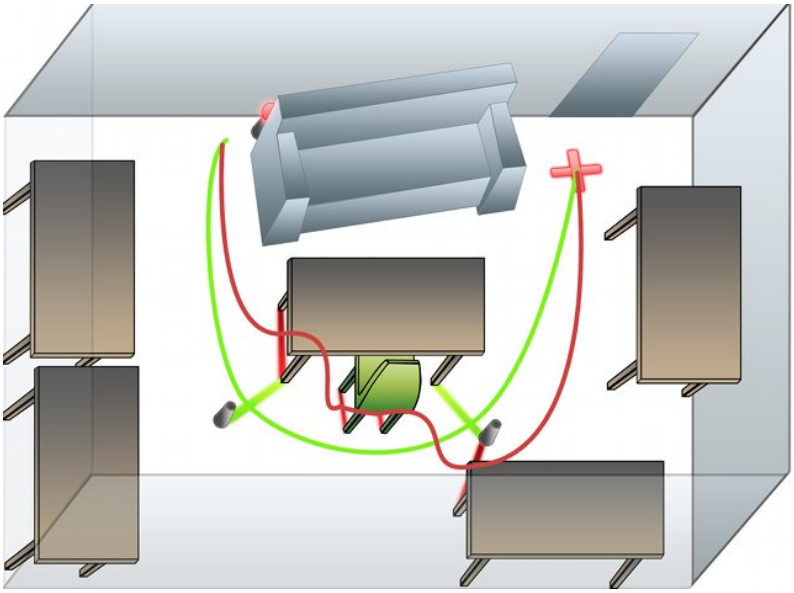


FIGURE 5.6 – Ce plan reproduit la configuration exacte de l'environnement de test. Deux parcours ont été définis : un simple (la ligne verte) et un plus difficile (la ligne rouge).

mations. Les choix possibles sont généralement de ce type : tout à fait d'accord, d'accord, ni d'accord ni pas d'accord, pas d'accord, pas du tout d'accord. Lors de cette expérience, les affirmations du test étaient les suivantes :

- Q1 : Il était facile d'apprendre à utiliser cette interface.
- Q2 : Il était facile de se souvenir des différentes commandes de cette interface.
- Q3 : Il était facile de déplacer le robot avec cette interface.
- Q4 : Il était facile d'orienter la tête du robot avec cette interface.
- Q5 : Utiliser cette interface était distrayant.
- Q6 : La latence entre l'envoi d'une commande et son exécution était dérangeante.
- Q7 : Dans l'ensemble, cette interface était satisfaisante.

Les participants devaient également choisir leur interface préférée.

5.3.3 Résultats

Comme on peut le voir sur le tableau 5.1, récapitulant les différentes mesures quantitatives relevées durant les expériences, il n'y a pas de différence significative entre les différentes interfaces pour le parcours facile. Par contre, pour le parcours difficile, le temps de parcours ainsi que le nombre de commandes envoyées avec l'interface iPhone Go-To sont légèrement supérieurs à ceux obtenus avec les deux autres interfaces. Cette différence peut sans doute être expliquée par le fait qu'avec cette interface, les utilisateurs focalisaient plus leur attention sur l'interface elle même que sur le robot lui-même. Cette hypothèse ne peut cependant pas être directement sou-

tenue par une mesure quantitative mais seulement par des observations qualitatives lors des expériences.

Parcours	Facile			Difficile		
Interface	IGT	IFD	WTUI	IGT	IFD	WTUI
Temps (min)	2.8 (1.3)	2.5 (0.7)	2.7 (0.4)	4.9 (2.2)	4.3 (2.9)	4.0 (1.1)
En mouvement	84 %	75 %	78 %	94 %	78 %	91 %
Nb d'actions	19 (17)	16 (7)	18 (3)	38 (13)	27 (15)	31 (9)

TABLE 5.1 – Moyenne (écart type) du temps de parcours total, du pourcentage de temps en mouvement et du nombre d'actions envoyées pour chacune des interfaces et pour chacun des parcours.

Les utilisateurs ont jugé les trois interfaces très faciles à apprendre. Ils ont également indiqué que déplacer le robot avec les interfaces iPhone (IGT et IFD) était plus facile qu'avec l'interface basée sur la Wiimote. Orienter la tête du robot a été jugé très facile avec l'interface iPhone Go-To et facile avec l'interface Wiimote. Cela a été jugé plus difficile avec l'interface iPhone flèches directionnelles, cependant ce faible score peut être expliqué par un problème technique qui rendait l'utilisation des sliders difficile. Les trois interfaces ont été jugées distrayantes, avec un léger avantage pour l'interface iPhone Go-To. Il est également intéressant de noter que bien qu'identique pour toutes les interfaces, la latence a été jugée plus négativement avec l'interface Wiimote. Enfin, les participants ont jugé les trois interfaces satisfaisantes, avec un net avantage pour l'interface iPhone Go-To. L'ensemble de ces résultats est résumé par la figure 5.7.

Les participants ont choisi à 58% l'interface iPhone Go-To comme leur interface favorite (figure 5.8).

5.3.4 Discussion

Comme montré dans la section précédente, il est intéressant de noter que l'interface iPhone Go-To a la plus faible utilisabilité réelle et la plus forte utilisabilité perçue. Plus précisément, bien que les participants ayant utilisé cette interface aient mis en moyenne 25% de temps supplémentaire pour réaliser le parcours difficile qu'avec l'interface Wiimote, ils ont quand même jugé qu'il était plus facile de déplacer le robot avec cette interface. Dans le cas de l'utilisation de l'interface Go-To, le pourcentage de temps plus faible où le robot était en mouvement, indique probablement que les utilisateurs ont passé plus de temps à chercher quelle commande envoyer au robot. Cependant, là encore, les utilisateurs ont jugé qu'il était plus facile d'apprendre à utiliser cette interface. L'interface Go-To a aussi été jugée plus satisfaisante et a été préférée par les participants.

Cette opposition des mesures qualitatives et quantitatives souligne l'impact de l'interface et de la métaphore utilisée sur le ressenti de l'expérience par les utilisateurs. Nos mesures quantitatives indiquent clairement que l'interface Go-To est

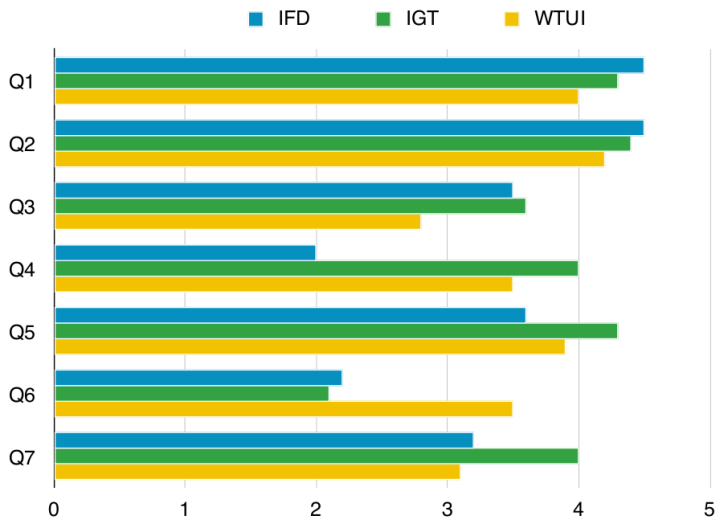


FIGURE 5.7 – Cette figure présente la moyenne des réponses apportées aux questionnaires d'utilisabilité. Les trois interfaces ont été jugés comme très facile à apprendre et les interfaces iPhone ont été perçues comme plus facile pour diriger le robot.
Q1 : Il était facile d'apprendre à utiliser cette interface.
Q2 : Il était facile de se souvenir des différentes commandes de cette interface.
Q3 : Il était facile de déplacer le robot avec cette interface.
Q4 : Il était facile d'orienter la tête du robot avec cette interface.
Q5 : Utiliser cette interface était distrayant.
Q6 : La latence entre l'envoi d'une commande et son exécution était dérangeante.
Q7 : Dans l'ensemble, cette interface était satisfaisante.

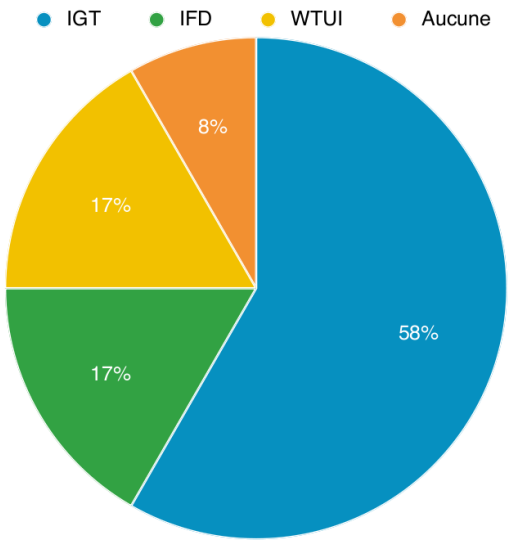


FIGURE 5.8 – Les interfaces préférées par les utilisateurs

moins efficace pour faire naviguer un robot à travers un parcours d'obstacles que les autres interfaces proposées. Les participants ont pourtant jugé que cette interface était globalement plus satisfaisante, mais également, selon leur perception, plus efficace pour déplacer un robot.

Des entretiens post-expérimentation nous ont permis d'identifier quelques raisons qui peuvent expliquer l'intérêt des utilisateurs pour l'interface Go-To. Les utilisateurs ont souligné avoir particulièrement apprécié :

- la capacité de pouvoir surveiller le flux vidéo de la caméra du robot, lorsqu'il se déplace, sans occlusions visuelles
- pouvoir facilement enchaîner les mouvements
- la grande facilité d'utilisation

Nous pouvons formuler plusieurs hypothèses afin d'essayer d'expliquer ces résultats. Tout d'abord, il semble que les interfaces WTUI et IFD ont été jugées négativement, car elles ont été considérées comme inadaptées à cette expérience. Plus précisément, ces interfaces laissaient supposer aux participants un contrôle très bas niveau et très réactif du robot. Or, ici les utilisateurs envoyaient des actions de haut niveau et avec une latence importante. Ces interfaces conviendraient sans doute mieux à l'utilisation d'un robot à roues. Cette différence entre les capacités espérées du robot et ses possibilités réelles a probablement engendrée une déception, dégradant l'expérience utilisateur, et donc l'évaluation subjective de ces interfaces. Ensuite, la possibilité de regarder la vidéo de ce que perçoit le robot sur l'écran de l'interface IGT a eu un impact très positif sur l'expérience utilisateur. Les participants prenaient en effet beaucoup de plaisir à regarder cette vidéo, et étaient donc moins ennuyés par la latence importante.

5.4 Conclusion

Dans ce chapitre nous avons présenté trois interfaces permettant à un utilisateur non-expert de guider un robot à travers un environnement domestique dans l'optique de lui montrer un objet particulier. Ces trois interfaces sont basées sur des objets médiateurs, tels qu'un iPhone ou un contrôleur Wiimote, qui permettent une interaction robuste et intuitive.

Nous avons également présenté les résultats d'une étude utilisateur montrant que les trois interfaces permettent effectivement de réaliser cette tâche. Nous avons aussi montré qu'il existait une importante différence entre l'utilisabilité réelle et l'utilisabilité perçue pour ces interfaces. En particulier, l'interface iPhone Go-To est jugée comme la plus satisfaisante et est également l'interface préférée des utilisateurs bien qu'étant légèrement moins efficace que les autres interfaces.

Dans la suite de notre étude, nous allons continuer à développer plusieurs interfaces car il n'existe pas une interface faisant l'unanimité chez les participants. Cependant, l'accent sera mis sur l'interface iPhone Go-To qui est à la fois efficace et offre une bonne expérience utilisateur. De plus, l'écran tactile offre une grande variété d'interactions possibles comme nous le verrons.

Chapitre 6

Interfaces pour l'enseignement d'objets visuels nouveaux à un robot : utilisation d'étiquettes (*tags*)

Sommaire

Résumé du chapitre

Dans ce chapitre, nous présenterons un système développé pour étudier le rôle de l'interface dans l'apprentissage à un robot de la reconnaissance d'objets visuels

nouveaux associés à des mots. Nous décrirons, en particulier, les différentes interfaces développées, basées sur des objets médiateurs, tels qu'un iPhone, une Wiimote ou encore un pointeur laser. Nous décrirons aussi notre système de perception et d'apprentissage, basé sur une approche par *sac-de-mots* visuels. Ce système complet nous permettra d'évaluer réellement les performances d'apprentissage obtenues en fonction de l'interface utilisée par les participants, et donc d'évaluer l'impact de ces interfaces sur la qualité des exemples d'objets visuels collectés par un robot. Nous décrirons également une autre interface, basée sur des gestes naturels d'un humain, utilisée comme point de comparaison avec nos interfaces médiatrices. Enfin, nous présenterons quelques expériences préliminaires, visant à explorer les performances atteignables par notre système d'apprentissage. Nous évaluerons également l'impact de l'encerclement des objets visuels par les utilisateurs sur les performances de reconnaissance visuelle.

6.1 Objectifs

6.1.1 Apprentissage d'objets visuels nouveaux

Dans ce chapitre, nous allons présenter la deuxième itération du cycle de développement et plus particulièrement un système permettant d'enseigner à un robot comment reconnaître des objets visuels nouveaux associés à des mots nouveaux. Après avoir étudié comment un utilisateur pouvait guider un robot dans un environnement du quotidien, nous nous intéressons ici à l'étape suivante : c'est-à-dire comment un humain peut montrer des objets à son robot et en collecter des exemples d'apprentissage pour s'en construire une représentation visuelle. Comme expliqué dans les chapitres précédents, cette interaction présente un intérêt majeur pour permettre aux utilisateurs d'aider leur robot à découvrir son environnement immédiat.

6.1.2 Restriction aux étiquettes (*tags*)

L'enseignement conjoint d'objets visuels nouveaux associés à des mots vocaux nouveaux à un robot est un problème complexe qui soulève plusieurs défis lorsqu'on n'utilise pas d'identifiant ou de symbole pour représenter une de ces deux modalités (voir la liste des défis en section 3.6). Le robot doit, d'une part, être doté de mécanismes de perception et de reconnaissance visuelle, lui permettant de catégoriser les exemples d'objets visuels. D'autre part, il doit aussi pouvoir percevoir et reconnaître la parole afin de créer des catégories des différents mots. Enfin, il faut pouvoir construire des associations entre ces catégories visuelles et auditives.

Dans ce chapitre, nous proposons, dans un premier temps, de nous attaquer à un sous-problème de ce défi général, en utilisant des symboles (identifiants uniques) pour représenter les mots associés aux objets visuels. Plus précisément, nous allons utiliser des étiquettes, sous forme de chaînes de caractères, comme mots de notre système d'apprentissage. L'utilisation de ces symboles permet de contourner

le problème de la construction des catégories de mots ainsi que de l'association entre les catégories auditives et visuelles. En effet, la comparaison des chaînes de caractères est triviale et sûre, et nous permet donc de regrouper directement les mots équivalents. Nous pouvons ainsi nous focaliser sur la perception et la catégorisation d'objets visuels et plus particulièrement sur les défis d'interaction liés. Les symboles associés seront soit directement entrés par les utilisateurs à l'aide d'un clavier, soit ajouté par un humain expert, simulant ainsi un système de reconnaissance de la parole aussi performant que l'humain. Nous décrirons plus en détails la manière d'entrer les étiquettes dans la section 6.2.3. Nous discuterons également dans cette section de notre choix de ne pas utiliser un système de reconnaissance de la parole standard.

6.1.3 Défis d'interaction et d'interface

Le problème de la catégorisation des objets visuels est un problème connu et très étudié. Nous avons présenté une revue de la littérature existante sur le sujet dans la section 3.5. Comme indiqué précédemment, ce problème a principalement été attaqué du point de vue de la perception visuelle ainsi que de l'apprentissage machine. La plupart des algorithmes développés supposent l'accès à de grandes bases de données d'images pouvant être utilisées pour entrainer ces systèmes. Or, les performances de ces algorithmes sont très dépendantes des exemples fournis en entrée du système, c'est-à-dire dans notre cas des photos des objets prises par le robot.

Selon nous, il est donc indispensable de s'attaquer à ce problème en amont de ces algorithmes de classification. En particulier, dans un contexte d'interactions humain-robot quotidiennes, il faut déjà pouvoir répondre aux questions suivantes :

- Comment permettre à un utilisateur de faire collecter à son robot des « bons » exemples d'apprentissage qui constitueront une base de données d'entrainement permettant d'améliorer les performances du système dans son ensemble ?
- Comment permettre à un utilisateur non-expert de collecter ces exemples à travers des interactions intuitives, peu nombreuses et non-contraignantes ?

Ces questions soulèvent des problématiques de pointage et d'attention partagée décrits dans la section 3.6. Ces questions sont encore plus critiques dans notre contexte de robotique personnelle où les utilisateurs ne souhaiteront probablement collecter que peu d'exemples. Leur qualité en sera d'autant plus cruciale.

Dans le reste de ce chapitre, nous allons présenter un système complet développé, dans le contexte de la restriction aux étiquettes, (perception, apprentissage machine et interfaces) et plus particulièrement les différentes interfaces proposées. Ce système a été implémenté sur le robot Nao. L'évaluation et la comparaison de ces interfaces, à travers une étude utilisateur, seront présentées en détails dans le chapitre suivant. Les résultats décrits dans ce chapitre ont été publiés dans [Rouanet 2009b][Rouanet 2009c][Rouanet 2010a][Rouanet 2010b].

FIGURE 6.1 – Les performances des algorithmes standards de classification/reconnaissance d'objets visuels sont très dépendantes des exemples d'apprentissage fournis en entrée du système (ici les photos prises par le robot). Nous cherchons donc à développer des interfaces permettant aux utilisateurs de collecter des bons exemples d'apprentissage qui amélioreront les performances du système dans son ensemble.

6.2 Perception et reconnaissance visuelle : approche incrémentale par *sacs-de-mots-visuels*

Afin de pouvoir comparer quantitativement l'impact des différentes interfaces sur la qualité des exemples d'apprentissage et donc sur les performances en reconnaissance d'objets visuels de notre système, nous avons utilisé comme référence une méthode standard de l'état de l'art pour l'apprentissage de la reconnaissance visuelle : l'approche *sacs-de-mots-visuels* [Sivic 2003][Csurka 2004]. Elle a aussi été largement utilisée en robotique et plus particulièrement pour la navigation [Wang 2005][Angeli 2008].

Les objets visuels sont ici représentés par un ensemble de caractéristiques distinctives visuelles extraites d'une image, appelées ici *mots*, dont la structure n'est pas conservée. Ces points d'intérêt sont globalement invariants et permettent donc de trouver des correspondances entre deux images. Ces *mots* sont ensuite catégorisés à l'aide d'un *dictionnaire* (via une approche par quantification vectorielle). Une méthode de vote basée sur la fréquence de ces mots dans les différentes catégories est utilisée pour permettre la classification des différents objets visuels. Ces différentes étapes seront décrites en détails dans les sections suivantes.

Les *sacs-de-mots-visuels* présentent un ensemble d'avantages pour notre application. Tout d'abord, cette approche est fréquemment utilisée dans les applications de reconnaissance d'objets visuels et permet donc une comparaison facile avec d'autres travaux. Nous utilisons ici une implémentation incrémentale, présentée dans [Filliat 2007] et [Filliat 2008], qui permet de traiter de nouveaux exemples d'apprentissage sans avoir besoin de retraiter toutes les données précédemment utilisées. Ceci est particulièrement intéressant pour des applications temps réel comme celle présentée ici. De plus, l'utilisation de caractéristiques visuelles locales de l'image rend la reconnaissance robuste aux occlusions partielles d'objets et leur catégorisation permet une grande robustesse au bruit (e.g. bruit sur l'image, changement d'illumination).

6.2.1 Extraction et description de caractéristiques visuelles

Dans cette section, nous allons décrire comment des caractéristiques distinctives visuelles (*image features*) sont extraites d'une image afin de permettre l'identification de correspondances entre des images et ainsi pouvoir procéder à une comparaison fiable d'objets visuels. Ce processus peut être résumé en trois étapes :

1. Des points d'intérêts sont sélectionnés à des endroits particuliers de l'image, tels que des coins ou des contours. La détection de ces points d'intérêt doit permettre de trouver les mêmes points sous différents points de vue. Certaines méthodes utilisent une répartition dense des points d'intérêts, couvrant ainsi uniformément toute l'image.

2. Une description locale de l'image, généralement représentée par un vecteur, est ensuite calculée pour chacun de ces points d'intérêt. Ce descripteur doit à

la fois être distinctif mais également robuste au bruit et aux transformations géométriques.

3. Les descripteurs doivent ensuite pouvoir être comparés. Une distance entre les vecteurs est généralement utilisée (e.g. distance euclidienne).

Il existe une très grande variété de caractéristiques visuelles utilisées, chacune ayant des propriétés particulières. Parmi l'ensemble des caractéristiques visuelles existantes, nous pouvons, par exemple, citer :

- Les points d'intérêts SIFT (*Scale-invariant feature transform*) correspondent aux maxima de l'image de différences de gaussiennes qui permettent d'identifier des points invariants aux changements d'échelle et d'orientation. La description utilisée représente l'histogramme de gradient orienté autour du point détecté [Lowe 2004].
- Les histogrammes de couleurs locaux décomposent l'image en fenêtre adjacente de taille constante. Pour chacune des fenêtres, l'histogramme représente la distribution des couleurs (ou de l'intensité) dans cette zone.

Mikolajczyk et Schimd ont présenté une comparaison des caractéristiques visuelles les plus utilisées dans [Mikolajczyk 2003].

Ici, nous utilisons les caractéristiques visuelles SURF (*Speeded Up Robust Features* [Bay 2008]). SURF propose à la fois une méthode de détection de points d'intérêts ainsi que de description. Cette approche permet d'atteindre des performances similaires aux descripteurs SIFT tout en réduisant sensiblement le temps de calcul nécessaire et permet donc une utilisation temps réel. Tout comme avec l'approche SIFT, les points d'intérêts détectés par SURF sont invariants aux changements d'échelle et de rotation. SURF est aussi très largement utilisé dans le domaine de la reconnaissance/catégorisation d'objets et permet donc une comparaison facile avec d'autres travaux.

6.2.1.1 Détection de points d'intérêts SURF

La détection des points d'intérêts SURF est basée sur la matrice hessienne. Soit un point $\mathbf{x} = (x, y)$ de l'image I, la matrice hessienne $H(\mathbf{x}, \sigma)$ en ce point $\mathbf{x}$ et à l'échelle σ est défini comme :

$$H(\mathbf{x}, \sigma) = \begin{bmatrix} L_{xx}(\mathbf{x}, \sigma) & L_{xy}(\mathbf{x}, \sigma) \\ L_{xy}(\mathbf{x}, \sigma) & L_{yy}(\mathbf{x}, \sigma) \end{bmatrix}$$

où $L_{xx}(\mathbf{x}, \sigma)$ représente la convolution de la dérivée gaussienne de second ordre $\frac{\partial^2}{\partial x^2} g(\sigma)$ avec l'image I au point $\mathbf{x}$. Le calcul de la dérivée gaussienne de second ordre est approximé grâce aux images intégrales pour des raisons de performances [Bay 2008].

Les maxima du déterminant de la matrice hessienne sont utilisés pour localiser les points d'intérêts.

6.2.1.2 Description de points d'intérêts SURF

Pour chacun des points d'intérêt détectés, un descripteur est calculé à partir de la distribution de réponses d'ondelettes de Haar dans le voisinage du point d'intérêt. Plus précisément, une orientation reproductible est identifiée pour chacun des points d'intérêt. Une région carrée est construite autour du point d'intérêt suivant cette orientation. Elle est sous-divisée en 4x4 sous-régions carrées. Pour chacune un descripteur v est calculé comme suit :

$$v = (\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|)$$

où d_x (resp. d_y) représente la réponse d'ondelette de Haar dans la direction horizontale (resp. verticale). Le descripteur du point d'intérêt, de dimension 64, représente le vecteur v pour chacune des 4x4 sous-régions. La comparaison de descripteurs SURF peut être réalisée grâce au calcul de la distance euclidienne entre les vecteurs.

6.2.2 Catégorisation des caractéristiques visuelles : construction d'un dictionnaire

Comme les caractéristiques visuelles SURF que nous utilisons sont sensibles au bruit et sont décrites dans des espaces de grandes dimensions, elles ne sont pas directement utilisées comme *mots* de notre système de *sac-de-mots-visuels* mais elles sont d'abord catégorisées par quantification vectorielle. Le dictionnaire (ou *codebook*) a été construit par regroupement (*clustering*) d'un grand nombre de caractéristiques visuelles extraites dans des images représentatives de notre environnement. La procédure suivante a été utilisée pour construire le dictionnaire utilisé dans toutes les expériences décrites dans la suite de cette thèse :

1. Nous avons enregistré une séquence vidéo ($\sim 5min$) issue de la caméra du Nao alors que le robot se déplaçait à travers le salon de notre laboratoire. Approximativement 1000 images ont été ainsi collectés ($\sim 3fps$). Nous nous sommes assurés qu'aucun des objets visibles lors de la création du dictionnaire n'ont été utilisés lors des expériences.
2. Les caractéristiques visuelles SURF ont été extraites pour chacune de ces images ($\sim$ 100 caractéristiques SURF par image).
3. L'ensemble de ces caractéristiques ont été regroupées grâce à l'algorithme K-Means [1]. La taille du dictionnaire a été fixée à 2^{12} par expérimentation (un K trop grand rend les *mots* insuffisamment distinctifs et un K trop petit rend l'appariement difficile).

Les caractéristiques visuelles sont alors projetées en *mots* par plus proche voisin.

6.2.3 Saisie des étiquettes correspondant aux mots

Comme nous l'avons expliqué dans la section 6.1.2, dans le système décrit dans ce chapitre, les mots associés aux objets visuels sont des symboles représentés sous

1. `http://en.wikipedia.org/wiki/K-means_clustering`

forme de chaînes de caractères. L'utilisation d'identifiants uniques permet de directement regrouper les représentations visuelles d'un même objet.

Nous avons essayé différentes approches pour saisir ces symboles. Nous avons dans un premier temps proposé l'utilisation d'un clavier virtuel affiché sur l'écran d'un TMC. La figure 6.2 représente un prototype de cette interface. Cependant, comme nous le décrirons dans la section 6.3, d'autres types d'interfaces ont été utilisés et certaines ne permettaient pas l'utilisation de clavier virtuel ou réel.

FIGURE 6.2 – Les étiquettes associées aux objets visuels peuvent être directement saisies par l'utilisateur à l'aide d'un clavier virtuel.

Nous avons donc choisi d'utiliser une autre approche pour la saisie des étiquettes. Les images étaient manuellement annotées par un humain expert. Nous pouvions ainsi simuler l'utilisation d'un algorithme de reconnaissance aux performances optimales, ou tout du moins aussi élevées que celle d'un humain.

Nous avons fait le choix de ne pas utiliser un système de reconnaissance automatique de la parole standard. En effet, ces systèmes se basent généralement sur les séquences de mots pour améliorer la robustesse de la reconnaissance. Or, nous cherchons dans cette thèse à utiliser des mots isolés. De plus, nous ne souhaitons pas nous limiter à une seule langue, ni même aux sons linguistiques. Les utilisateurs pourraient en effet vouloir associer des onomatopées aux objets visuels.

Nous nous sommes attaqués à cette problématique dans le chapitre 9, où nous avons proposé une mesure de similarité entre des mots acoustiques. Cette mesure a été couplée à une interface bien conçue qui permettait à l'utilisateur de participer au processus de reconnaissance vocale et ainsi d'en améliorer la robustesse.

6.2.4 Classification d'objets visuels

Dans l'approche *sacs-de-mots-visuels*, les images utilisées sont représentées par un histogramme de *mots* non-ordonnés, c'est-à-dire sans conserver aucune structure géométrique. Il existe des versions de cette méthode utilisant ce type d'informations mais nous ne les avons pas utilisées ici.

Pour notre application, il est intéressant que le classifieur puisse être entrainé de manière incrémentale, c'est-à-dire qu'il soit possible de traiter de nouveaux exemples et donc de reconnaître de nouveaux objets sans avoir besoin de retraiter l'ensemble des données déjà vues. Pour permettre cela, nous utilisons une méthode où entrainer le système signifie mettre à jour un modèle statistique des différents objets et où l'évaluation correspond à évaluer la probabilité de chaque objet pour une image donnée.

L'image I est représentée par l'histogramme des occurrences des différents mots du dictionnaire :

$$H_I = \{O_{w_i}, \forall w_i \in D\}$$

où O_{w_i} est le nombre d'occurrences du mot w_i appartenant au dictionnaire D.

Une catégorie C_i, c'est-à-dire un objet visuel, est alors représentée par la somme des histogrammes des images appartenant à cette catégorie.

$$H_{C_i} = \sum H_I, \forall I \in C_i$$

L'apprentissage d'un nouvel exemple d'apprentissage signifie simplement la mise à jour de cet histogramme.

Afin de reconnaître un objet dans une nouvelle image nous utilisons une méthode de vote basée sur la fréquence des différents mots pour chacun des objets. Plus précisément, chacun des mots w vote pour tous les objets o tels que $O_{wo} \neq 0$ où O_{wo} représente le nombre d'occurrences du mot w pour l'objet o. Chaque vote a un poids déterminé par la méthode *tf-idf* (*term frequency-inverted document frequency*) qui permet de pénaliser les mots les plus communs [Sivic 2003]. L'objet reconnu est l'objet ayant obtenu le score le plus élevé. L'algorithme 1 décrit la procédure plus en détails.

Une mesure de la qualité du vote peut être définie simplement par la différence normalisée entre le meilleur et le second meilleur score :

$$quality = \frac{score\_winner - score\_second}{\sum score}$$

6.3 Interfaces développées

Nous allons maintenant décrire les quatre interfaces développées pour permettre à des utilisateurs non-experts d'enseigner à leur robot des objets visuels nouveaux associés à des mots, exprimés sous forme d'étiquettes, nouveaux. Ces interfaces ont été conçues dans l'optique d'améliorer la qualité des exemples d'apprentissage que les utilisateurs fournissent au système d'apprentissage. Nous avons également essayé de rendre ces interfaces intuitives et agréables à utiliser.

Nous avons exploré l'utilisation de différents objets médiateurs ainsi que différents types de retour, fournis à l'utilisateur, de ce que le robot perçoit. Trois de nos interfaces sont effectivement basées sur des objets médiateurs tels qu'un iPhone, un contrôleur Wiimote ou encore un pointeur laser. Nous avons choisi des objets bien connus dont la prise en main devrait être relativement facile. La dernière interface

Algorithm 1 RECOGNIZE_OBJECT(new_image)

{D is the codebook}
{O_{wo} is the number of occurences of the word w for the object o.}

$keypoints \leftarrow extract\_SURF(new\_image)$
$words \leftarrow categorize(keypoints, D)$
for each w **in** $words$ **do**
  for each o **in** $known\_objects$ **do**
    if $O_{wo} \neq 0$ **then**
      $tf \leftarrow \frac{O_{wo}}{\sum 0_{w_i o} \forall w_i}$
      $idf \leftarrow \log \frac{|D|}{|\{d : w \in d\}|}$
      $score_o \leftarrow score_o + tf * idf$
    end if
  end for
end for
$recognized\_object \leftarrow \arg\max_o score_o$

a quant à elle été ajoutée pour permettre de comparer celles basées sur des objets médiateurs avec une interface a priori plus naturelle et directe, basée sur des gestes. Elle se révèlera au final moins efficace et moins facile à utiliser pour les humains.

Afin de pouvoir être comparée équitablement, chacune de ces interfaces fournissaient exactement les mêmes fonctionnalités aux utilisateurs :

- diriger / guider le robot
- attirer son attention dans une direction ou vers un objet particulier
- définir une zone à l'intérieur de l'image correspondant à l'objet (seules les interfaces iPhone et laser fournissent cette possibilité, pour les deux autres, l'image entière est utilisée comme « zone objet »)

Le déclenchement de la capture d'un exemple se faisait par interaction directe avec le robot dans un souci de comparaison avec l'interface gestuelle. L'étiquette était automatiquement ajouté lors de la capture (plus de détails seront données au chapitre suivant).

Il est également important de noter que toutes ces interfaces sont basées sur les même capacités sensori-motrices du robot Nao. Nous avons volontairement choisi de ne pas améliorer ses capacités en utilisant des périphériques externes tels que des caméras au plafond ou une caméra grand angle bien que cela permettrait d'accroître l'utilisabilité de nos interfaces, car nous souhaitons les évaluer dans un contexte de robotique personnelle réaliste dans un futur proche, c'est-à-dire avec un robot ayant des capacités réalistes (voir la section 4.2.2 pour plus de détails).

Chacune des interfaces développées va être présentée en détail dans les sections suivantes en insistant sur leurs différences. Certaines de ces interfaces (l'interface iPhone et l'interface Wiimote) sont des extensions plus ou moins modifiées des in-

terfaces de guidage décrites au chapitre précédent. Des liens vidéos décrivants les différentes interfaces développées sont disponibles en annexe A.

6.3.1 Interface iPhone

La première interface est basée sur un iPhone et reprend les principes de base de l'interface iPhone Go-To présentée en section 5.2.1. L'écran de l'appareil est utilisé pour afficher de manière continue le flux vidéo issu de la caméra du robot comme on peut le voir sur la figure 6.3. C'est particulièrement important ici car cela permet à l'utilisateur de savoir très exactement ce que perçoit le robot et ainsi permettre de réellement atteindre des situations d'attention partagée. Comme expliqué précédemment, avoir un retour visuel de ce que perçoit le robot est particulièrement important pour les utilisateurs non-experts, qui sont plus enclins à faire de fausses suppositions sur les capacités du robot, comme par exemple, sur son champ de vision. Cependant l'attention des utilisateurs est séparée entre attention directe et indirecte pouvant surcharger leur charge cognitive de travail.

FIGURE 6.3 – L'interface iPhone avec le flux vidéo permettant aux utilisateurs de surveiller ce que voir le robot. Les utilisateurs peuvent dessiner des « trajectoires » directement sur l'écran tactile de l'appareil interprétées comme des commandes.

Les utilisateurs interagissent avec le robot à travers des gestes naturels effectués directement sur le retour visuel. Grâce au très large succès de l'iPhone on peut tirer parti d'une interface bien connue permettant aux utilisateurs de rapidement la prendre en main. Nous avons conservé les traits horizontaux et verticaux, présentés au chapitre précédent, pour diriger le robot. Le robot se déplace sans s'arrêter jusqu'à ce que l'utilisateur retouche l'écran. Ces gestes envoient des commandes de marche au robot (e.g. « marche tout droit » ou « tourne sur toi-même »). Nous avons également conservé la capacité de simplement « taper » un endroit particulier

(a) Reculer (b) Tourner à gauche (c) Pointer (d) Encercler

FIGURE 6.4 – L'interface iPhone fournit les commandes suivantes aux utilisateurs. Ils peuvent faire avancer et reculer le robot à l'aide de traits verticaux (figure 6.4a) et le faire tourner sur lui-même à l'aide de traits horizontaux (figure 6.4b). Ils peuvent aussi lui indiquer de regarder un endroit précis directement en tapant sur l'écran (figure 6.4c), le robot va alors orienter sa tête pour centrer sa vision en ce point. Enfin, ils peuvent encercler une partie de l'écran (figure 6.4d) pour lui indiquer que c'est un objet visuel qu'il doit apprendre à reconnaître.

de l'écran pour demander au robot de regarder à cet endroit. Le robot oriente alors sa tête pour centrer sa vision sur le point indiqué par l'utilisateur. Les différentes commandes sont récapitulées dans le schéma de la figure 6.4.

Nous avons poussé ce concept d'esquisse encore plus loin. Ainsi, lorsque l'utilisateur veut montrer un objet au robot afin de lui enseigner un mot nouveau associé à cet objet, il doit bien sûr commencer par s'assurer que l'objet est bien présent dans le champ de vision du robot en regardant l'écran de l'iPhone, ensuite il encercle l'objet directement sur l'écran tactile. Encercler est vraiment un geste très naturel et correspondant très bien à la « sélection » d'un objet grâce à une interface tactile (comme sur la figure 6.5). Cette métaphore a souvent été utilisée, notamment pour permettre à un utilisateur de « sélectionner » un objet. Schmalstieg et al. l'ont, par exemple, utilisée pour permettre de sélectionner un objet dans un monde virtuel 3D [Schmalstieg 1999]. Hachet et al. utilisent des cercles 2D pour positionner de manière précise une caméra dans un espace en trois dimensions [Hachet 2008]. De même que pour les traits, nous utilisons des heuristiques très simples pour détecter ces gestes d'encerclement (le geste effectué par l'utilisateur doit être suffisamment long et revenir à un point proche de son point de départ).

L'utilisateur, en encerclant les objets, fournit également, de manière complètement transparente, une information très intéressante puisqu'il délimite la zone de l'image où l'objet est présent. La partie intérieure du cercle effectué par l'utilisateur est extraite en utilisant des algorithmes de traitement d'image classiques (Bresenham [2] et Flood Fill [3]). Comme nous le détaillerons ci-dessous cette information est très précieuse car séparer l'objet de l'arrière plan (segmenter) reste un problème très complexe en environnement non-contraint.

Le geste d'encerclement a donc deux fonctions distinctes très importantes :

– Il permet de signaler au robot que l'utilisateur souhaite lui montrer un nouvel objet visuel qu'il doit apprendre à reconnaître.

2. `http://en.wikipedia.org/wiki/Bresenham's_line_algorithm`
3. `http://en.wikipedia.org/wiki/Flood_fill`

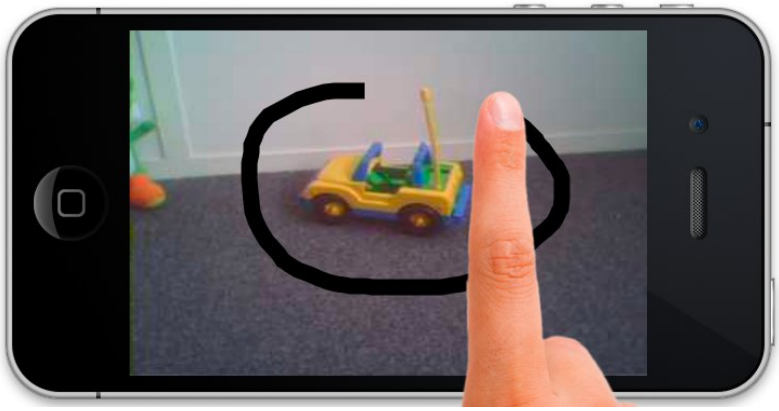


FIGURE 6.5 – L'utilisateur encercle l'objet pour demander au robot de le prendre en photo mais également afin de le segmenter.

– Il permet aussi de faire segmenter l'objet à l'utilisateur.

6.3.2 Interface Wiimote

Cette deuxième interface est basée sur le contrôleur Wiimote. Contrairement à la version présentée au chapitre précédent, ici, pour déplacer le robot les utilisateurs se servent de la croix directionnelle. Pour faire avancer le robot, il suffit de presser la flèche du haut et pour arrêter le robot il suffit de la relâcher. Pour permettre aux utilisateurs d'orienter la tête du robot, nous avons conservé l'interaction présentée à la section 5.2.3 où les mouvements appliqués à la Wiimote sont reproduits par la tête du robot. Nous avons fait ce choix afin de mieux séparer ces deux types de mouvements.

FIGURE 6.6 – Les utilisateurs peuvent demander au robot de se déplacer (marcher avant/arrière et tourner sur lui-même) à l'aide de la croix directionnelle visible sur l'image de gauche. Ils peuvent aussi lui orienter la tête en inclinant le contrôleur Wiimote comme sur l'image de droite.

L'intérêt majeur de cette interface est de permettre aux utilisateurs de constamment focaliser leur attention directement sur le robot car ils n'ont pas besoin de regarder l'interface. Par contre, cette interface ne fournit aucun retour de ce que le robot perçoit. Les utilisateurs doivent donc « deviner » si un objet est bien présent dans le champ de vision du robot. Cette absence de retour nous fournit un point de comparaison avec les autres interfaces très intéressant.

6.3.3 Interface Wiimote-Laser

La troisième interface est également basée sur le contrôleur Wiimote mais couplé avec un pointeur laser. La Wiimote est utilisée pour diriger le robot, de la même manière que pour l'interface précédente. Le pointeur laser est, quant à lui, utilisé pour attirer l'attention du robot de manière plus précise, c'est-à-dire pour orienter la tête du robot (voir la figure 6.7).

FIGURE 6.7 – Les utilisateurs peuvent déplacer le robot à l'aide de la Wiimote. Ils peuvent aussi attirer l'attention du robot vers un objet en le désignant avec le pointeur laser.

Avec cette interface, le robot suit constamment du regard le point laser et essaie de toujours le maintenir au centre de son champ de vision. L'utilisation d'un pointeur laser pour attirer l'attention sur un point précis est une pratique courante, lors des présentations orales par exemple, et permet donc d'intuitivement attirer l'attention d'un robot sur un objet particulier. Les utilisateurs peuvent orienter la tête du robot, de proche en proche, dans la bonne direction ou ils peuvent directement pointer sur un objet si celui-ci est déjà présent dans le champ de vision du robot.

Le point laser est automatiquement détecté dans les images issues de la caméra du robot. Comme la plupart des robots personnels actuels, Nao a une caméra avec

une faible résolution (640x480). À cette limite, viennent s'ajouter des conditions d'éclairage changeantes. Nous avons utilisé un laser vert très puissant [4]. Le vert est une couleur probablement moins présente dans un environnement quotidien et est donc plus saillante. Nous avons utilisé un algorithme de détection du laser basé sur la reconnaissance de région elliptique verte et très lumineuse.

Contrairement à Kemp et al. qui utilisent un robot doté d'une caméra omni-directionnelle [Kemp 2008] ou bien Ishii et al. qui utilisent des caméras fixées au plafond agrandissant artificiellement le champ de vision du robot [Ishii 2009], dans nos expériences le robot a un champ de vision très restreint. Or, pour attirer l'attention du robot avec le pointeur laser, les utilisateurs doivent déjà être capables d'estimer correctement son champ de vision (le pointeur laser ne pouvant être détecté qu'à l'intérieur du champ de vision du robot). Comme expliqué précédemment, cette tâche peut être complexe, particulièrement pour les utilisateurs non experts, à cause des différences d'appareils visuels entre un humain et un robot. Pour aider les utilisateurs à comprendre si le robot détecte le pointeur laser et donc à mieux appréhender son champ de vision, nous avons ajouté plusieurs retours utilisateurs à cette interface :

- Tout d'abord, lorsque le robot suit le laser, les utilisateurs peuvent voir la tête du robot bouger et ont donc un premier retour direct visuel.
- Nous avons aussi ajouté un retour haptique. Plus précisément, le contrôleur Wiimote vibre lorsque le robot détecte le laser.

Grâce à la combinaison de ces deux retours les utilisateurs peuvent donc s'assurer que le robot détecte le laser et donc qu'un objet est bien présent dans son champ de vision.

Le retour de ce que perçoit le robot présenté par cette interface n'est cependant pas aussi complet que celui affiché sur l'interface iPhone par exemple. En effet, les utilisateurs ne peuvent pas directement voir ce que le robot perçoit. Ils ne peuvent donc pas réellement être sûrs que le robot détecte bien l'objet, mais ils peuvent seulement s'assurer que le robot voit la partie de l'objet sur laquelle le point laser est présent. En effet, à cause des limites physiques du robot, il se peut qu'il ne puisse pas centrer sa vision sur le point laser et donc qu'une partie de l'objet soit en dehors de son champ de vision. Par contre, le laser fournit un retour visuel direct aux utilisateurs qui peuvent s'assurer qu'ils pointent bien précisément là où ils le souhaitent et peuvent donc ajuster leur geste si besoin. Ceci est particulièrement intéressant dans un environnement où il peut y avoir un grand nombre d'objets et où une légère erreur de l'humain avec le pointeur laser peut revenir à désigner un mauvais objet et donc à collecter un mauvais exemple d'apprentissage.

Une fois que les utilisateurs sont parvenus à attirer l'attention du robot sur un objet qu'ils veulent nommer, ils encerclent l'objet directement avec le point laser. De même que pour l'interface iPhone, ce geste permet à la fois de signaler au système qu'on veut réellement introduire un nouvel exemple d'apprentissage mais permet également de segmenter grossièrement l'objet. Cependant, ici l'encerclement

4. `http://www.apinex.com/ret2/gpcf01an.html`

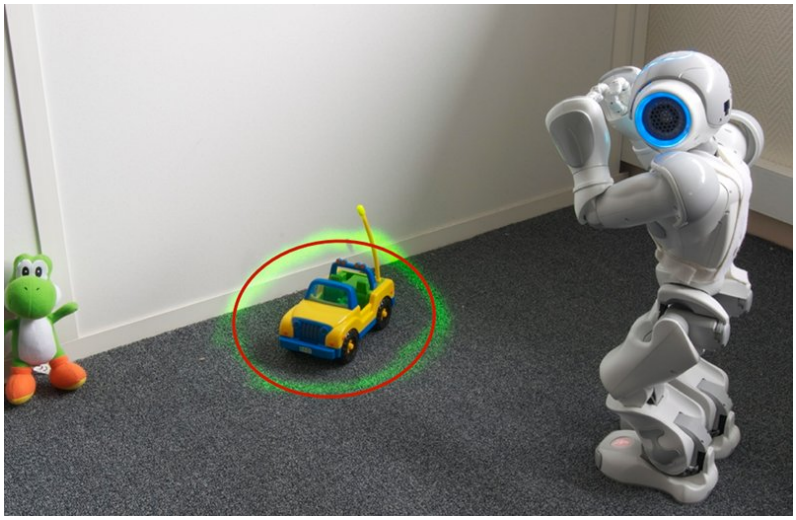


FIGURE 6.8 – Les utilisateurs peuvent entourer avec le pointeur laser l'objet qu'ils souhaitent montrer à leur robot. L'ellipse moyenne (en rouge) est utilisée pour délimiter la zone de l'image où l'objet est présent.

ne se fait pas directement sur l'image mais dans l'environnement lui-même, entrainant quelques problèmes techniques. Tout d'abord, le taux de rafraichissement de la caméra étant faible ($\sim 3fps$), les utilisateurs doivent entourer l'objet plusieurs fois pour nous permettre d'obtenir suffisamment de points pour reconstruire l'esquisse. L'ensemble des points détectés sont enregistrés et utilisés pour calculer l'ellipse moyenne, les mouvements d'encerclement étant généralement elliptiques, à l'aide de la formule suivante :

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0 \; avec \; A \neq 1$$

$$\Rightarrow Bxy + Cy^2 + Dx + Ey + F = -x^2$$

Écrit sous forme matricielle :

$$\alpha X = \beta \; avec$$

$$\alpha = \begin{pmatrix} x_1 * y_1 & y_1^2 & x_1 & y_1 & 1 \\ ... & ... & ... & ... & ... \\ x_n * y_n & y_n^2 & x_n & y_n & 1 \end{pmatrix} \beta = \begin{pmatrix} -x_1^2 \\ ... \\ -x_n^2 \end{pmatrix}$$

Comme ce système d'équation est surdéterminé on cherche à trouver la valeur de X qui minimise l'équation quadratique suivante (moindre carrés) :

$$\arg\min_X = \left\| \beta - \alpha X \right\|^2$$

ce qui revient à résoudre l'équation :

$$\hat{X} = (\alpha^T \alpha)^{-1} \alpha^T \beta$$

Une fois l'ellipse extraite, elle est utilisée pour déterminer la zone de l'image où l'objet était présent. Cependant, la projection des points 3D dans le plan image (2D) peut, plus particulièrement lorsque le fond n'est pas planaire, couper l'objet (voir les exemples sur la figure 6.9). Au lieu d'améliorer la qualité des exemples d'apprentissage nous verrons plus tard qu'encercler avec le pointeur laser dégrade la qualité des exemples dans certains cas.

FIGURE 6.9 – Encercler avec le pointeur laser entraine des problèmes de projection des points détectés sur le plan de la caméra du robot qui peut couper les objets.

6.3.4 Interface basée sur des gestes naturels

Avec la dernière interface les utilisateurs peuvent guider le robot directement grâce à des gestes des mains ou des bras. Afin que cette interface soit aussi naturelle que possible, nous n'avons pas restreint l'utilisation à certains gestes particuliers. Cependant, la reconnaissance et l'interprétation des gestes en milieu non-contraint pose toujours des problèmes de robustesse à l'heure actuelle, nous avons donc utilisé un protocole de type magicien d'Oz (*Wizard of Oz : WoZ*) où un humain (le magicien) contrôlait le robot afin de répondre correctement aux gestes qu'il percevait. Ce type de protocole est souvent utilisé lors d'expériences utilisateurs en interaction humain-robot [Maulsby 1993][Green 2004][Walters 2005]. En effet, il permet de contourner facilement et efficacement certaines limitations techniques actuelles, telles que la reconnaissance de gestes. Le magicien, généralement un humain expert, contrôle un robot lors d'une expérience utilisateur. Il opère discrètement, en étant par exemple caché dans une pièce adjacente à la salle d'expérimentation, afin que les participants de l'expérience ne réalisent pas que le robot est en fait télé-opéré par un humain (voir le schéma de la figure 6.10). Ce protocole nous a permis de développer une interface qui ne soit pas limitée par les performances de l'algorithme de reconnaissance de gestes ici remplacé par l'humain.

Comme expliqué précédemment, nous ne voulons pas augmenter les capacités du robot artificiellement, ainsi l'humain guidant le robot ne pouvait percevoir l'interaction qu'à travers les yeux du robot : i.e. il ne voyait l'interaction qu'à travers le flux

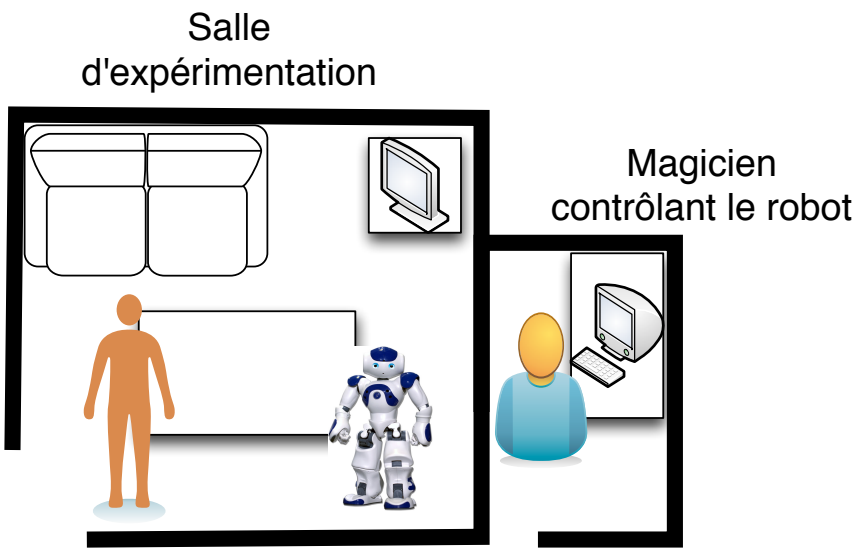


FIGURE 6.10 – Nous avons utilisé un protocole de type magicien d'Oz pour la reconnaissance de gestes. Un humain expert, le magicien, était caché dans une pièce adjacente à la salle d'expérimentation. Il percevait l'interaction à travers les yeux du robot et devait le contrôler en fonction des gestes de l'humain qu'il reconnaissait.

vidéo de la caméra du robot. Cela permet aussi de pouvoir comparer les différentes interfaces de manière plus équitable car elles sont toutes basées sur exactement les mêmes capacités sensori-motrices du robot.

Les magiciens ont été recrutés dans notre laboratoire. Ils connaissaient donc tous bien le robot utilisé et les problématiques liées à son appareil perceptif (angle de vision restreint et petite taille). Des instructions concernant l'utilisation de l'interface leurs étaient fournies. Ils pouvaient s'entrainer à diriger le robot jusqu'à ce qu'ils s'estiment prêts. En plus de leur expliquer comment diriger le robot, nous leur présentions également les objectifs de l'interface. Les magiciens avaient pour consigne d'interpréter les gestes de l'utilisateur et de piloter le robot en conséquence. La feuille d'instructions fournie aux magiciens est disponible en annexe B. Les magiciens étaient remplacés fréquemment (après approximativement 30 minutes d'interaction), afin d'éviter le phénomène d'habituation aux comportements des utilisateurs.

Comme nous le verrons à travers l'analyse des résultats présentés en section 7.4, l'humain contrôlant le robot ne pouvait percevoir l'interaction qu'à travers les yeux du robot et donc avec un angle de vue très réduit, la plupart des gestes effectués par l'utilisateur étaient en fait en dehors du champ de vision du robot ; ils étaient donc tout simplement invisibles pour le magicien. Par conséquent, cette interface a priori naturelle entraine une interaction peu robuste et donc très peu satisfaisante pour l'utilisateur, et ce même en utilisant un algorithme de reconnaissance et

FIGURE 6.11 – Avec ce mode d'interaction, le robot est guidé grâce aux gestes de la main et du bras faits par les utilisateurs. Afin d'avoir une reconnaissance robuste nous avons utilisé un protocole de type WoZ où un humain contrôlait le robot en fonction des gestes reconnus mais il ne pouvait percevoir l'interaction qu'à travers la caméra du robot.

d'interprétation de gestes aussi performant qu'un humain !

Cette interface présente un intérêt central dans notre recherche, car elle permet de comparer une interface a priori naturelle avec nos interfaces basées sur des objets médiateurs. Elle nous permet de plus d'étudier l'utilisabilité de ce type d'interaction si on suppose un algorithme de reconnaissance de gestes aussi performant qu'un humain. Elle diffère cependant fortement des trois autres interfaces. En effet, pour l'humain contrôlant le robot le problème de segmentation des objets est trivial. Ceci constitue un biais important en faveur de cette interface : le magicien a naturellement tendance à centrer la vision du robot sur l'objet (ce qui reste un problème difficile pour le robot). Ce biais renforce les résultats négatifs de cette interface que nous décrirons dans les sections suivantes. Malgré cet avantage certain par rapport aux autres interfaces, elle est très importante pour notre recherche, car elle permet d'étudier les performances qu'il serait possible d'obtenir avec des algorithmes de reconnaissance aussi performants qu'un humain utilisés avec des robots personnels actuels. Cette interface va notamment nous permettre de valider l'hypothèse que l'utilisation d'objets médiateurs permet d'obtenir des interactions plus robustes, plus efficaces et donc au final plus intuitives que des interactions a priori naturelles mais n'étant pas à l'heure actuelle suffisamment robustes dans des conditions d'utilisation réalistes (c'est-à-dire avec un robot ayant des capacités sensori-motrices très limitées), et ce, même avec des algorithmes de reconnaissance aussi performants qu'un humain.

6.4 Évaluation du système de reconnaissance d'objets

Comme expliqué précédemment, l'évaluation et la comparaison de ces quatre interfaces fera l'objet d'un chapitre à part entière afin de décrire précisément le protocole et l'installation expérimentale utilisés pour garantir des résultats statistiquement intéressants. Dans cette section, nous allons présenter les expérimentations pilotes qui nous ont permis de valider notre système de reconnaissance d'objets et d'avoir une première idée des résultats que l'on peut espérer obtenir dans des conditions « idéales ». Nous n'avons donc pas ici essayé de collecter des exemples d'apprentissage avec nos interfaces mais plutôt utilisé des exemples de bonne qualité afin de tester notre système de reconnaissance.

6.4.1 Construction d'une base de données d'exemples

Nous avons voulu tester notre système de reconnaissance avec des images d'objets du quotidien. A notre connaissance, la base de données la plus proche de ce que nous recherchons est la base de données ETH-80 [5]. Cependant, les objets y sont groupés par classe et non par instance (deux images correspondant au tag « balle » peuvent correspondre à deux balles ayant des aspects visuels différents). De plus, les photos ont été prises avec un fond neutre et des angles de vues très similaires. Or, dans notre contexte, le robot devra être capable de reconnaître des objets sous des points de vue très différents et avec des fonds différents.

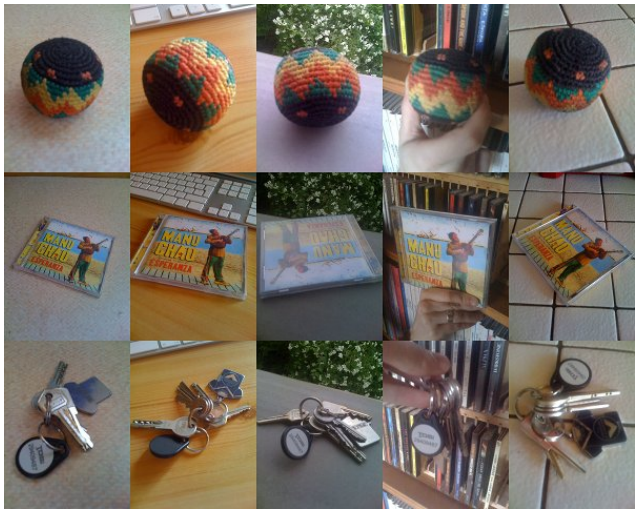


FIGURE 6.12 – Quelques exemples d'apprentissage d'objets du quotidien ont été collectés par un utilisateur expert afin de constituer une base de données d'exemples idéale et ainsi tester les performances atteignables par notre système de reconnaissance. Les objets utilisés étaient des objets de la vie de tous les jours, transportables et très texturés.

5. `http://people.csail.mit.edu/jjl/libpmk/samples/eth.html`

Nous avons donc choisi de construire notre propre base de données d'exemples d'apprentissage. Nous avons sélectionné 20 objets du quotidien différents (une balle, un CD, un journal, des clés...). Ces objets sont tous transportables et très texturés (notre système de reconnaissance d'images étant basé sur la texture uniquement). Pour chacun des objets, 10 photos ont été prises avec cinq arrière-plans différents (2 photos par objet et par arrière-plan). Le nombre d'exemples dans notre base est volontairement faible, afin de coller à notre scénario de robotique personnelle, où il est peu probable que les utilisateurs souhaiteront collecter un grand nombre d'exemples d'apprentissage. Il est donc indispensable que notre système de reconnaissance puisse fonctionner même avec peu d'exemples d'apprentissage. Enfin la résolution des images était réduite pour correspondre à celle de la caméra d'un robot personnel « classique » (640x480). Des exemples d'images de cette base de données peuvent être vus sur la figure 6.12. Les photos ont été prises par un utilisateur expert travaillant dans notre laboratoire.

6.4.2 Évaluation

Le protocole suivant a été utilisé pour évaluer les performances de notre système de reconnaissance :

- N images par objets étaient choisies par tirage aléatoire.
- Le système de reconnaissance était entrainé avec ces images.
- Le système de reconnaissance était testé sur les (10 - N) images non utilisées lors de l'entraînement.
- L'expérience était répétée 20 fois. Les résultats décrits représentent la moyenne des valeurs obtenues lors de ces tests.

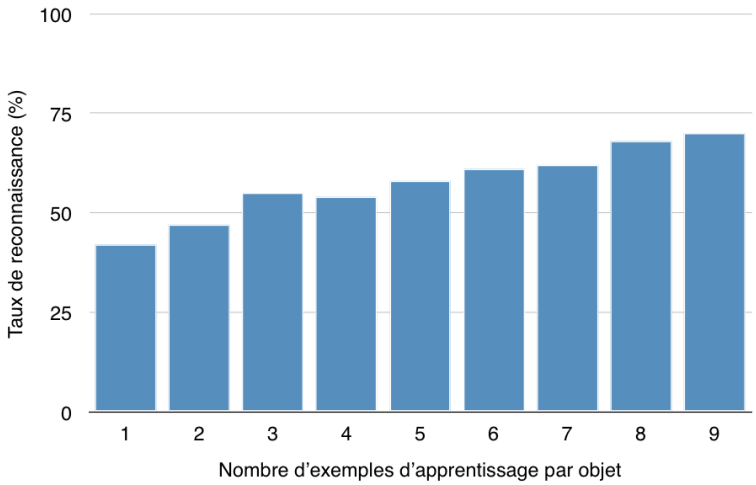


FIGURE 6.13 – Le système de reconnaissance d'objets visuels utilisé dans nos travaux permet de reconnaître des objets même avec peu d'exemples dans la plupart des cas ($\sim$ 75% avec 9 exemples par objet).

Comme on peut le voir sur le graphique 6.13, notre système permet donc d'atteindre un taux de reconnaissance relativement élevé même avec peu d'exemples d'apprentissage ($\sim$ 75% avec 9 exemples par objet). Par contre, il est clair que même avec de très bons exemples collectés par un utilisateur expert il n'est pas possible d'obtenir un taux de reconnaissance de 100%.

6.5 Impact de l'encerclement sur les performances

6.5.1 Double rôle de l'encerclement

Demander aux utilisateurs d'encercler les objets qu'ils veulent montrer au robot a deux rôles très importants :

- Cela permet d'une part de clairement séparer, par deux commandes différentes, le fait de vouloir attirer l'attention du robot sur un objet et le fait de réellement vouloir lui enseigner un mot nouveau pour cet objet. Cette interaction est donc cruciale pour la bonne compréhension de l'interaction. Ce type de sélection, « au lasso », est très intuitive et correspond donc très bien à cette tâche.
- De plus, en encerclant les objets les utilisateurs nous fournissent, de manière complètement transparente, une délimitation grossière de la zone de l'image à l'intérieur de laquelle l'objet est présent. Cette information permet d'obtenir une segmentation, certes approximative mais déjà très utile, de l'image.

6.5.2 Problème de la segmentation

La segmentation reste en effet un problème difficile dans un environnement non-contraint ou inconnu. Plusieurs approches ont été développées ces dernières années afin de s'attaquer à cette question. Cependant toutes ces méthodes souffrent de problème de robustesse. Parmi les différents types d'algorithmes existants nous pouvons par exemple citer :

- L'algorithme *region-growing* essaie de déterminer les régions où la texture ou la couleur sont approximativement homogènes. Ces régions sont agrandies incrémentalement à partir d'une graine [Adams 1994]. Cette classe d'algorithmes ne peut pas traiter des objets complexes composés de plusieurs sous-parties ayant des couleurs ou des textures variées. De plus, la texture d'un objet peut aussi être très similaire à la texture de l'arrière plan (voir les exemples sur la figure 6.14).
- Une autre classe d'algorithmes essaie de délimiter les bords d'un objet grâce à ses mouvements (*motion based segmentation* [Arsenio 2003]). Ce type d'algorithme se limite donc aux objets que l'on peut bouger, agiter. Avec cette approche on ne pourra donc pas segmenter des objets fixes tels qu'une prise électrique par exemple.
- Enfin, une autre famille d'algorithmes utilise des images de profondeur pour déterminer des régions de l'image qui appartiennent à une même surface (range

segmentation [Bab-Hadiashar 2006]). Cette approche ne permet donc évidemment pas de traiter des objets plats (comme des posters).

FIGURE 6.14 – Certains objets ne peuvent pas être segmentés à l'aide des méthodes classiques. Par exemple, l'objet sur la gauche est pratiquement de la même couleur que le fond. L'objet du milieu ne peut pas être bougé et ne peut donc pas être segmenté à l'aide des méthodes se basant sur le mouvement. Enfin, l'objet sur la droite ne pourra être segmenté par les méthodes utilisant les images de profondeur.

Ce problème n'a pas, à l'heure actuelle, de solution universelle existante, du fait même de la définition floue de ce qu'est un objet. Les enfants humains mettent du temps à apprendre le concept « d'objet visuel » [Pylyshyn 2001]. Pour l'humain adulte, qui possède une connaissance a priori et culturelle des objets, c'est un problème trivial. Ainsi en demandant à l'utilisateur de résoudre ce problème pour nous, nous pouvons traiter n'importe quel type d'objet dans n'importe quel environnement.

6.5.3 Évaluation de l'encerclement

Afin d'avoir une première évaluation de l'impact que l'encerclement peut avoir sur les performances du système de reconnaissance, nous avons manuellement encerclé les objets sur les images de la base de données décrite dans la section précédente. L'encerclement a été fait par un utilisateur expert et non à l'aide d'une des interfaces. Ceci introduit bien sûr un biais, mais permet déjà d'avoir une idée quant aux performances gagnées grâce à l'encerclement des objets sur l'image. Le même test que celui présenté à la section précédente a été reproduit afin de comparer les performances obtenues en utilisant soit l'image entière, soit uniquement la portion de l'image entourée par l'utilisateur.

Comme nous pouvons le constater sur le graphique 6.15, encercler les images permet d'obtenir un gain du taux de reconnaissance de l'ordre de 20%. Ce gain est très intéressant car il permet de réduire significativement le nombre d'exemples d'apprentissage nécessaire pour obtenir un taux de reconnaissance similaire. En

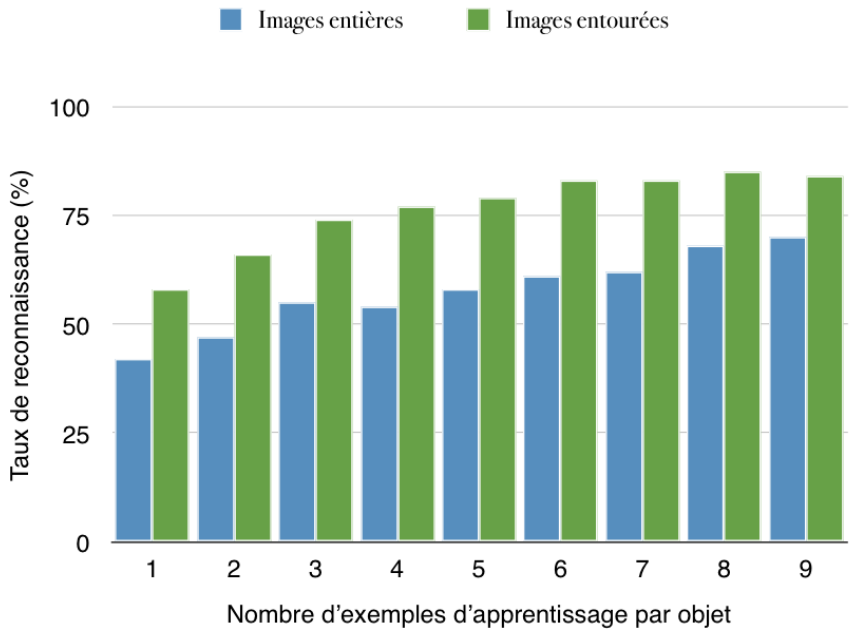


FIGURE 6.15 – L'encerclement permet d'améliorer d'approximativement 20% le taux de reconnaissance obtenu et donc de diminuer le nombre d'exemples d'apprentissage nécessaires pour obtenir un score équivalent.

particulier, trois images encerclées permettent d'obtenir un résultat équivalent à neuf images entières. Le nombre d'exemples nécessaire a été divisé par trois !

Il est bien sûr important de noter que ces résultats ne présentent qu'une première idée de ce qu'il est possible d'obtenir avec notre système. La base de données d'images a été construite dans des conditions idéales, bien que proche de notre contexte réel d'utilisation, par un utilisateur expert et non à l'aide de nos interfaces. De même, les objets ont été encerclés par un utilisateur expert. Afin d'avoir des résultats plus écologiquement valides nous allons décrire en détails dans le chapitre suivant une étude utilisateurs complète et à grande échelle où des humains non-experts ont utilisé les quatre interfaces présentées ici pour faire collecter des exemples d'apprentissage à un robot personnel.

Chapitre 7

Évaluation réaliste à travers un jeu robotique

Sommaire

Résumé du chapitre

Nous allons ici présenter une étude utilisateurs, à grande échelle, et réalisée hors de notre laboratoire, visant à comparer les différentes interfaces proposées, et surtout leur impact sur la qualité des exemples d'apprentissage collectés par un robot. Lors de cette expérience, les participants devaient en effet enseigner des mots nouveaux associés à des objets visuels nouveaux à un robot. Nous discuterons de la conception de cette expérience sous forme de jeu robotique, pour impliquer les utilisateurs et les maintenir motivés, malgré l'abstraction de la tâche qui leur était demandée. Nous présenterons également les résultats obtenus, qui nous ont notamment permis de montrer que l'interface iPhone permettait aux utilisateurs même non-experts de collecter des exemples de très bonne qualité. Cette interface a également été perçue comme facile à prendre en main et facile à utiliser par les utilisateurs.

7.1 Objectifs

Dans le chapitre précédent nous avons présenté quatre interfaces permettant à un utilisateur d'attirer l'attention du robot sur un objet particulier et d'associer un nom à cet objet. Ces interfaces ont été couplées avec un système d'apprentissage permettant au robot de reconnaître visuellement ces objets qu'on lui montrait. Ces interfaces n'avaient jusqu'à présent été évaluées qu'à travers des expériences pilotes avec peu de participants, pour la plupart familiers de nos travaux de recherches. Ces études ont été décrites aux chapitres 5 et 6. Or, comme détaillé dans la section 1.4, notre objectif à moyen terme est de permettre à un large public d'utiliser ces interfaces dans leurs interactions quotidiennes avec leur robot personnel. Il existait donc un besoin très fort de mener de vraies études utilisateurs. Pour pouvoir comparer quatre interfaces différentes et obtenir des statistiques valides, il était nécessaire d'avoir un nombre de participants important. Nous souhaitions aussi avoir des participants non-experts et réaliser une expérience hors du contexte de notre laboratoire.

Nous devions notamment nous assurer que ces interfaces étaient intuitives et efficaces. Nous devions aussi nous assurer que l'utilisation de ces interfaces était non-contraignante voir distrayante afin de maintenir les utilisateurs intéressés et donc permettre des interactions sur le long terme.

Nous avons donc conçu une étude utilisateur à grande échelle et dans des conditions représentant aussi fidèlement que possible une utilisation réaliste possible de nos interfaces. Plus de 100 personnes ont participé à cette étude qui s'est déroulée à Cap Sciences [1], le musée des sciences de Bordeaux, de juin à novembre 2010. Dans cette étude, nous cherchions à atteindre les objectifs suivants :

- Collecter des données à travers une interaction qui soit la plus réaliste possible afin de pouvoir tester notre système dans des conditions écologiquement valides : nous avons demandé à des participants non-experts de montrer des objets à un robot, possédant des capteur limités, dans un environnement du quotidien non-contrôlé.
- Comparer les différentes interfaces développées et leur impact, notamment celui des différents retours, de ce que le robot perçoit, sur la qualité des exemples d'apprentissage. Cet impact a été évalué à l'aide de tests de performance en classification / reconnaissance.
- Évaluer l'utilisabilité perçue et l'expérience utilisateur suivant les différentes interfaces.

Il était crucial pour nous de pouvoir réaliser une étude en dehors du cadre de notre laboratoire afin de pouvoir être confronté à un public correspondant mieux aux potentiels futurs utilisateurs de « descendants » de système du type de celui présenté au chapitre précédent. Cependant, comme nous avons pu le constater à travers les études utilisateurs pilotes, demander à des participants, non familiers avec la robotique, de nommer des objets à un robot reste pour eux une tache très

1. `http://www.cap-sciences.net/`

abstraite et artificielle et il existe donc un réel besoin de la justifier. Nous avons donc conçu notre expérience sous forme de jeu robotique permettant d'impliquer et de motiver les participants. Cette expérience a été décrite dans [Rouanet 2011].

7.2 Jeu robotique

Une des manières de pouvoir s'attaquer à tous les problèmes mentionnés ci-dessus est de concevoir notre étude utilisateurs comme un jeu robotique. En effet, les jeux sont un moyen très efficace de capter l'attention des utilisateurs ainsi que de les engager à participer. Par exemple, les jeux sérieux (ou *serious game*) sont fréquemment utilisés à des fins éducatives, d'entrainement à une tâche spécifique difficile ou impossible à reproduire en situation réelle [Michael 2005]. D'autre part, nous pensons que de la même façon que les jeux vidéos ont permis à des utilisateurs novices de résoudre des taches complexes et inhabituelles facilement, grâce à l'utilisation de mécanismes tels que des tutoriels ou des briefings, nous pourrions créer une expérience robotique, construite comme un jeu, qui aiderait les utilisateurs à mieux comprendre et à se souvenir de toutes les étapes nécessaires à l'enseignement de mots nouveaux associés à des objets visuels à un robot. Le scénario du jeu peut aussi permettre de justifier cette tâche et pousser les utilisateurs à vouloir la réaliser aussi bien que possible. Enfin, présenter l'expérience comme un jeu permet d'attirer un public large et varié où les participants sont moins stressés.

FIGURE 7.1 – Nous avons conçu un jeu robotique, afin d'évaluer le rôle des différentes interfaces développées, dans des conditions réalistes (hors-laboratoire et avec des utilisateurs non-experts). Le jeu permet de motiver et d'impliquer les participants, ainsi que de proposer un protocole précis et facilement reproductible.

Nous avons mis au point un scénario qui satisfaisait tous ces critères. L'histoire suivante était décrite aux participants : « Un robot arrive d'une autre planète et a été envoyé sur Terre afin de mieux comprendre ce qui semble être une coutume humaine courante : jouer au football. En effet, de leur planète lointaine, les robots ont reçu des informations partielles de cette pratique et veulent enquêter plus en détails sur ce phénomène. Un robot a donc été envoyé sur Terre, dans la chambre d'un fan de football, afin de collecter d'autres indices. Malheureusement, le robot a été abimé durant son voyage et ne va pas pouvoir compléter sa mission seul. Vous allez donc devoir l'aider à y parvenir ! ». L'utilisateur devait donc aider le robot à collecter ces indices, c'est-à-dire faire prendre au robot des photos de différent objets ayant un lien avec le football à l'aide d'une des interfaces. Afin de rendre l'histoire plus vivante et plus intéressante chaque fois qu'un exemple était collecté, le robot faisait un commentaire, souvent humoristique, sur la prétendue utilité de l'objet en question.

7.2.1 Environnement de jeu

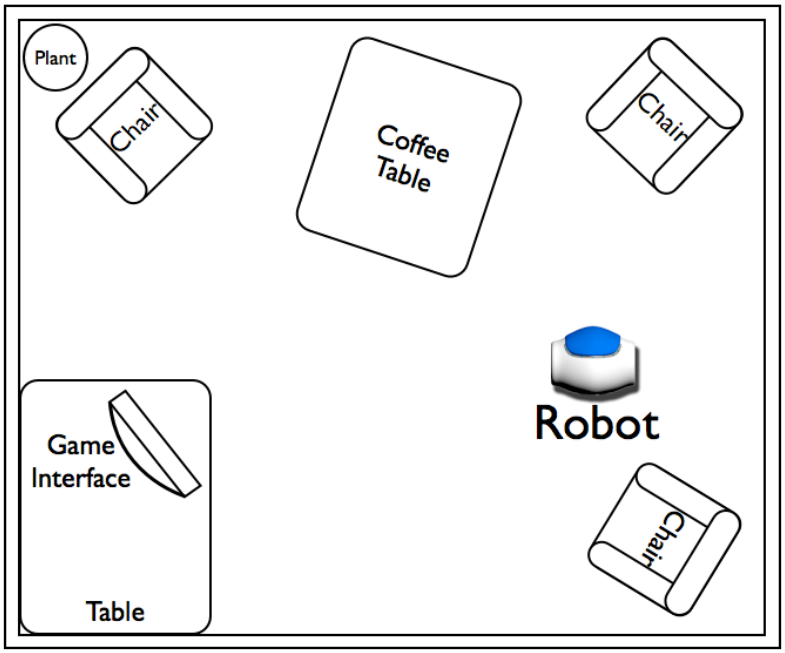


FIGURE 7.2 – Un salon d'une dizaine de mètre carrés a été reproduit. Nous avons utilisé des meubles ainsi qu'un grand nombre d'objets de décoration pour le rendre vivant. La plupart des objets avaient un lien direct avec le football afin de bien coller à notre scénario.

Pour cette expérience, nous avons recréé un salon d'une dizaine de mètres carrés situé à côté du coin café du musée. La figure 7.2 présente un plan de ce salon. Nous avons ajouté des meubles (tables, chaises) ainsi que beaucoup d'autres petits objets

du quotidien (journal, peluches, jouets, posters...) afin que la pièce ait réellement l'air habitée. Parmi tous ces objets, 12 sont directement liés au football et sont les objets à montrer au robot (voir la figure 7.3). Ces objets ont été choisis parce qu'ils correspondent bien à notre histoire mais également parce qu'ils sont suffisamment gros et texturés pour pouvoir être reconnu efficacement par notre système de reconnaissance visuelle si les exemples d'apprentissage sont de bonne qualité. Un lien vers une vidéo présentant l'installation expérimentale est disponible en annexe A.

L'environnement de jeu devait permettre de :

- Reproduire un environnement du quotidien pour permettre aux participants de faire l'expérience dans un environnement non-stressant et sans avoir le sentiment d'être observés.
- Réaliser l'expérience dans un environnement plausible afin que les utilisateurs aient à faire évoluer le robot dans une zone complexe et qu'ils collectent des exemples dans des conditions réalistes (éclairage changeant, arrière plan complexe, nombreux objets...).
- Immerger les utilisateurs dans l'histoire.

FIGURE 7.3 – Nous avons sélectionné 12 objets pour leur lien direct avec le football et car ils sont suffisamment gros et texturés pour être facilement reconnus. Chacun des participants devait montrer quatre objets parmi ces 12 (tirés aléatoirement) aux robots.

L'agencement général de la pièce est resté inchangé durant toutes les expériences afin d'avoir un environnement de test constant. Par contre, les petits objets étaient fréquemment déplacés (approximativement toutes les cinq expériences) et ce afin de simuler un environnement « vivant », modifié au cours du temps par ses habitants. Nous souhaitions en effet que les objets visuels montrés au robot puissent être reconnus en dépit de leur position dans la pièce. Plus précisément, cela signifie que

le robot doit être capable de reconnaître un objet visuel présent devant des arrières plans différents.

7.2.2 Interface du jeu

Comme expliqué plus haut, la conception de notre expérience s'inspire clairement des mécanismes utilisées dans les jeux vidéos. Afin de pouvoir afficher des informations relatives à l'expérience aux participants, un grand écran était disposé à l'intérieur du salon d'expérimentation. Cet écran servait entre autre à afficher les vidéos présentant l'histoire aux joueurs. Cette interface servait également à la présentation du tutoriel, où chacune des étapes était successivement décrite à travers une vidéo présentant les étapes à suivre pour réaliser cette tâche précise. Les tutoriels étaient propres à l'interface utilisée par les participants. Ils étaient cependant tous construits de manière identique. Les étapes décrites étaient les suivantes :

- avancer / reculer
- tourner à gauche / à droite
- orienter la tête du robot vers un endroit particulier
- faire collecter un exemple d'apprentissage au robot

FIGURE 7.4 – L'histoire du jeu était racontée à travers des vidéos affichées sur l'interface du jeu. Cet écran servait aussi à afficher les instructions des tutoriels pas-à-pas.

Le tutoriel servait également d'introduction à l'histoire et permettait aux utilisateurs de se familiariser avec le robot. La période d'apprentissage et le jeu étaient, de plus, regroupés ainsi en une seule histoire. Utiliser des tutoriels vidéos pour expliquer aux participants le fonctionnement des interfaces permet également de contrôler très exactement les informations communiquées aux participants et de s'assurer que tous les participants reçoivent exactement les mêmes consignes. Le tutoriel durait en moyenne 5 minutes.

7.2.3 Robot

Nous avons utilisé le robot Nao pour cette expérience. Comme nous l'avons expliqué dans le chapitre 4, ce robot représente bien selon nous le présent de la

robotique personnelle et sociale. Il possède, selon nous, des capacités sensori-motrices qui correspondent à celles dont seront certainement dotés les futurs robots utilisés quotidiennement par les humains. Son utilisation permet donc de reproduire des conditions potentiellement proches des applications futures de nos interfaces et de notre système complet. De plus son aspect de jouet et sa forme humanoïde collent très bien avec notre scénario de jeu. L'utilisation d'un seul robot, le Nao, pour nos expériences représente toutefois une limite de notre étude. Plus précisément, les capacités particulières du Nao, comme son appareil perceptif ou sa taille, ont nécessairement influencé les résultats présentés ci-dessous. Nous discuterons plus en détail de cette limite dans la section 10.4.

Comme expliqué dans la section 4.2.2, des comportements ont été ajoutés pour rendre le robot plus vivant. Ces comportements, tels que bailler, se gratter la tête ou encore regarder autour de lui, étaient déclenchés automatiquement lorsque le robot restait inactif trop longtemps ou lorsque les utilisateurs envoyaient une mauvaise commande. Nous avons également utilisé les couleurs des yeux du robot et des sons organiques pour exprimer différentes émotions permettant à l'utilisateur de mieux comprendre l'interaction.

7.3 Étude utilisateurs

7.3.1 Recrutement

Les expériences se sont déroulées de juin à novembre 2010 à Cap Sciences. 107 personnes ont participé à ce test.

La plupart des participants (74) étaient des visiteurs directement recrutés à l'intérieur du musée. Ces participants n'avaient en général que très peu de connaissance en robotique et jamais interagi avec ce type de robot auparavant. Cependant, ils venaient visiter un musée des sciences et étaient donc probablement assez ouverts aux nouvelles technologies. Le reste des participants a été recruté sur le campus de l'Université Bordeaux 1. Ces participants avaient une formation technologique et étaient également plus habitués à utiliser ce type d'interface. Par contre, ils ne possédaient pas de connaissance particulière en robotique.

Sur les 107 participants, 77 étaient des hommes et 30 des femmes. Les participants étaient âgés de 10 à 76 ans (M = 26.3, STD = 14.8).

7.3.2 Protocole expérimental

Chacun des participants ne testait qu'une seule des quatre interfaces possibles. 32 personnes ont testé l'interface *iPhone*, 27 l'interface *Wiimote*, 33 l'interface *Laser* et enfin 15 l'interface basée sur les *Gestes*.

Les participants devaient suivre le protocole expérimental suivant :

1. Remplir un formulaire de consentement (voir annexe C)
2. Remplir un pré-questionnaire visant à évaluer leur profil ainsi que leur vision a priori de la robotique

3. Faire l'expérience
   - Tutoriel
     (a) Réveiller le robot en lui touchant la tête
     (b) Le faire avancer
     (c) Le faire tourner à gauche puis à droite
     (d) Orienter sa tête vers la gauche, le haut, le bas et la droite
     (e) Le faire regarder votre visage (ou un ballon dans le cas de l'interface avec le pointeur laser)
     (f) Lui faire prendre une photo (collecter un exemple d'apprentissage)
     (g) Prononcer votre nom pour l'associer à votre visage
   - Mission
     (a) Attirer l'attention du robot vers un des 12 objets possibles (le choix des objets était tiré aléatoirement)
     (b) Faire prendre au robot une photo de cet objet
     (c) Prononcer le nom de l'objet

     Les étapes (a) à (c) étaient répétées quatre fois.
4. Remplir des post-questionnaires sur l'utilisabilité et l'expérience utilisateur.

L'expérience, questionnaires inclus, durait approximativement entre 20 et 30 minutes par participant.

7.3.3 Mesures

Durant les expériences, différentes mesures étaient enregistrées afin de pouvoir évaluer et comparer nos interfaces. Les mesures quantitatives étaient les suivantes :

- Les photos prises par le robot étaient sauvegardées puis utilisées comme entrée de notre système d'apprentissage lors de tests de classification / reconnaissance. Les images étaient automatiquement étiquetées (l'étiquette de l'objet était associé aux images). En effet, c'est l'interface de jeu qui indiquait aux participants quels objets visuels ils devaient faire prendre en photo au robot. Nous pouvions donc associer directement l'étiquette correspondante.
- Le son prononcé par les utilisateurs était également enregistré. Cependant, cette information n'a pas été utilisée dans les tests présentés dans ce chapitre mais le sera dans le chapitre 9.
- Le temps mis pour réaliser le tutoriel, pour collecter chacun des quatre exemples et enfin la durée totale de l'expérience.

Des mesures qualitatives ont aussi été collectées à travers des questionnaires. Les assertions principales portant sur l'utilisabilité et l'expérience utilisateur étaient les suivantes (voir annexe D) :

- Il était facile d'apprendre à utiliser cette interface.
- Il était facile de déplacer le robot.
- Il était facile de faire regarder un objet au robot.

– Interagir avec le robot était facile.
– Le robot mettait du temps à réagir.
– L'interface était plaisante à utiliser.

Nous avons également ajouté quelques questions portant sur le jeu lui même :

– Terminer le jeu était facile.
– Le jeu était distrayant.
– J'ai eu l'impression de faire équipe avec le robot.
– Je m'imagine jouer à d'autres types de jeux robotiques à l'avenir.

Les utilisateurs devaient exprimer leur accord vis-à-vis de ces affirmations sur une échelle de Likert à 5 niveaux.

7.4 Résultats

7.4.1 Analyse qualitative des images

Nous avons, tout d'abord, triés les images collectées, c'est-à-dire les exemples d'apprentissage correspondant aux objets visuels, en trois catégories (voir figure 7.5) : 1) les images où l'objet était entièrement visible, 2) les images où l'objet était seulement partiellement visible, 3) et enfin les images où l'objet n'était pas du tout présent. Les objets étaient définis comme « partiellement visible » dès qu'une partie de l'objet était absente de l'image. La figure 7.6 présente la répartition des images en trois catégories. Nous avons réalisé une analyse de variance à un facteur (« one-way ANOVA ») sur la condition « entièrement visible » et trouvé une différence significative entre les quatre interfaces ($F_{3,103} = 13.7, p < 0.001$). En particulier, nous pouvons remarquer pour les conditions où nous ne fournissions pas de retour sur ce que le robot perçoit aux utilisateurs (les interfaces *Wiimote* et *Gestes*), l'objet était entièrement visible dans seulement 50% des cas ! Le test post-hoc Tukey a montré que fournir un retour aux utilisateurs améliore significativement ce résultat (80% pour l'interface *laser* et 85% pour l'interface *iPhone*). De plus, nous pouvons constater que l'interface *iPhone* et en particulier son retour visuel évite, la plupart du temps (seulement 2% des cas), aux utilisateurs de collecter des mauvais exemples (quand l'objet n'est pas du tout visible).

FIGURE 7.5 – Les exemples d'apprentissage collectés durant l'expérience ont été triés en trois catégories. L'image de gauche représente un exemple où l'objet, la canette, est entièrement visible. Elle n'est que partiellement visible sur l'image du milieu et pas du tout visible sur l'image de droite.

Nous avons également divisé les images en deux sous-ensembles :
- les gros objets (les deux posters, les deux maillots et le ballon)
- les petits objets (tous les autres)

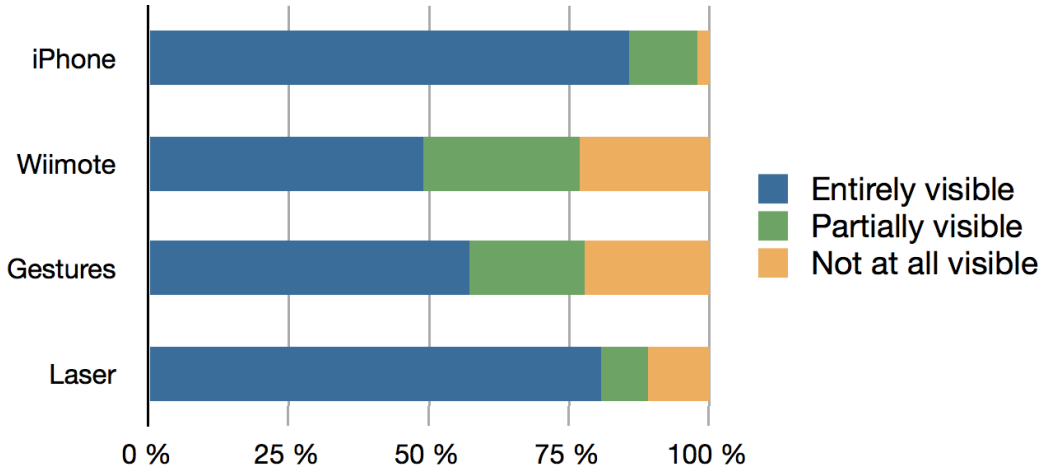


FIGURE 7.6 – Cette figure présente la répartition des images collectées en trois catégories : l'objet est 1) entièrement, 2) partiellement ou 3) pas du tout visible sur les images. On peut observer que sans retour (interfaces *Wiimote* ou *Gestes*) l'objet n'est entièrement visible que dans seulement 50% des exemples. Fournir un retour permet d'améliorer significativement ce résultat (80% pour l'interface *Laser* et 85% pour *l'iPhone*).

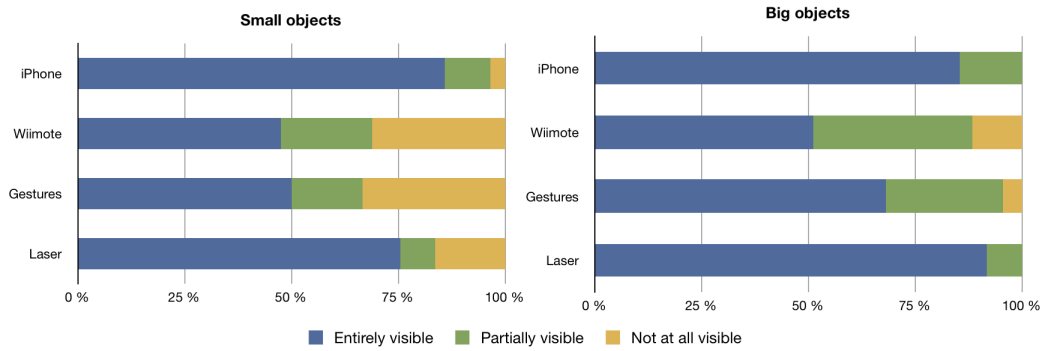


FIGURE 7.7 – Cette figure présente des résultats similaires à la figure 7.6, mais ici les images collectées ont été séparées en deux sous-ensemble : les gros et petits objets. Nous pouvons remarquer que les différences entre les quatre interfaces sont encore davantage accentuées pour les petits objets : avec l'interface *Wiimote* ou les *Gestes* les participants ont fourni de mauvais exemples d'apprentissage (condition « pas du tout visible ») dans approximativement un tiers des cas.

Comme nous pouvons le voir sur la figure 7.7, les différences entre les interfaces sont encore davantage accentuées pour les petits objets. En particulier, nous pouvons constater que le manque de retour de ce que le robot perçoit amène les utilisateurs à collecter approximativement un tiers où l'objet n'est pas du tout visible. Alors que le retour fourni par l'interface *laser* permet d'améliorer ce résultat (le taux d'erreur est seulement de 20%), seule l'interface *iPhone* permet d'empêcher les utilisateurs de collecter de mauvais exemples d'apprentissage des petits objets. Enfin, nous pouvons également observer que toutes les interfaces permettent de collecter des exemples plutôt bons (où l'objet est entièrement visible) des gros objets. Par contre, alors que les objets sont pratiquement systématiquement entièrement visibles avec les interfaces *iPhone* et *Laser* (dans plus de 85% des cas), ils étaient seulement partiellement visibles dans un tiers des cas avec les conditions *Wiimote* ou *Gestes*.

7.4.2 Analyse quantitative des images

Nous avons également utilisé les images collectées durant les expériences comme entrée de notre système d'apprentissage afin de pouvoir réellement mesurer la qualité des exemples et leur impact quantitatif sur les performances générales du système. Comme expliqué précédemment, notre système de reconnaissance est basé sur la technique des sacs de mots visuels. Pour les tests décrits ci-dessous, nous avons construit un dictionnaire en enregistrant une séquence vidéo de cinq minutes (approximativement 1000 images) capturée par la caméra du Nao alors que celui-ci se déplaçait dans le salon de notre laboratoire. Nous nous sommes assurés qu'aucun des objets ou mobiliers visibles durant cet enregistrement n'était utilisé dans l'environnement expérimental. Le dictionnaire construit était ainsi indépendant de notre installation expérimentale.

Nous avons utilisé le protocole suivant pour réaliser nos tests :

- N images par objet étaient tirées aléatoirement parmi les images collectées par les utilisateurs avec une interface particulière (nous regroupions donc des images collectées par plusieurs utilisateurs différents). En effet, collecter un seul exemple de cinq objets différents prenait déjà 20 à 30 minutes par participant, il n'était donc pas possible de leur demander de collecter plusieurs exemples des douze objets. Ceci introduit un biais que nous avons contrebalancé en sélectionnant aléatoirement les images utilisées et en répétant nos tests un grand nombre de fois. Comme montré dans les résultats ci-dessous la variabilité entre utilisateurs (indiqué par l'écart type) est relativement faible. En particulier, dans la plupart des cas la variabilité entre interfaces est largement supérieure à la variabilité entre utilisateurs d'une même interface.
- Nous entrainions notre système avec ces images.
- Nous testions notre apprentissage sur une base de tests (voir ci-dessous). L'apprentissage effectué avec les images d'entrainement était évalué lors d'une tâche de classification de nouvelles images.
- Le test était répété 50 fois en choisissant un ensemble d'images d'entrainement

différent à chaque fois.
– Le résultat final représente la moyenne et l'écart type du taux de reconnaissance obtenu pour l'ensemble des tests.

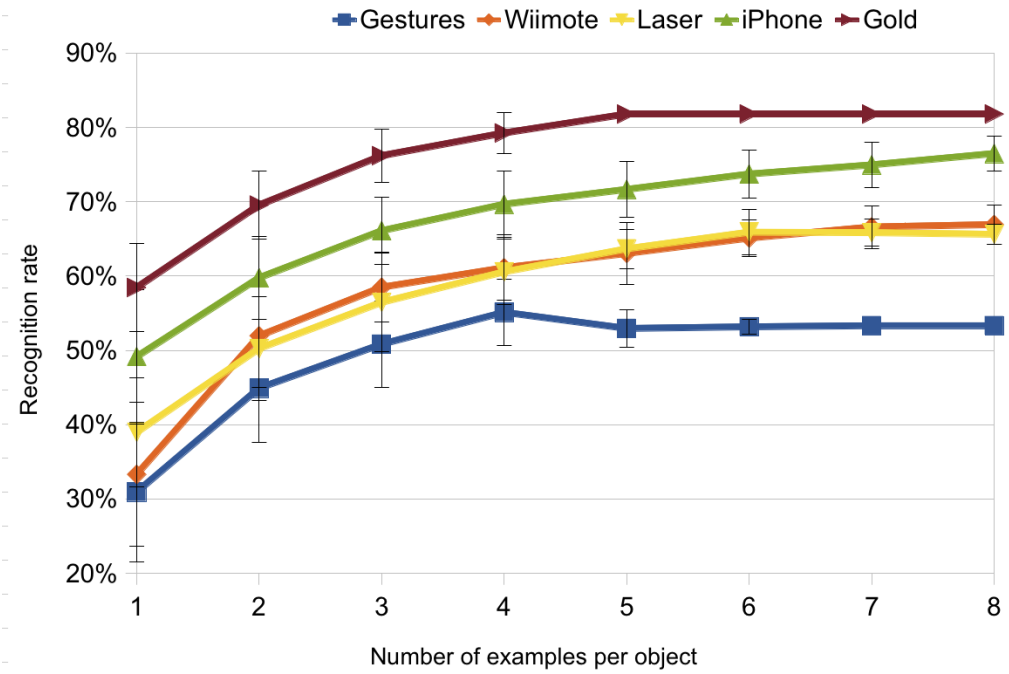


FIGURE 7.8 – **Taux de reconnaissance pour les 12 objets :** Cette figure montre l'impact de l'interface sur la qualité des exemples d'apprentissage et donc sur les performances en généralisation de notre système de reconnaissance. Nous pouvons notamment constater que l'interface *iPhone* permet aux utilisateurs de collecter des exemples d'une qualité significativement plus élevée que les autres interfaces. De plus, cette interface permet à des utilisateurs non-experts de fournir au système d'apprentissage des exemples d'une qualité quasiment similaire à la qualité de nos exemples « gold » collectés par un utilisateur expert.

La base de données de test a été construite par un utilisateur expert qui a collecté 10 exemples d'apprentissage de chacun des 12 objets. Pour cela, il a utilisé l'interface utilisée par le magicien de la condition *Gestes*. Les images ont été prises dans les mêmes conditions expérimentales que lors des tests utilisateurs. Elles représentent pour nous les exemples « optimaux » que l'on peut espérer obtenir comme entrée de notre système d'apprentissage. Ces données ont été séparées en deux sous-groupes : la première moitié a été utilisée comme des exemples d'apprentissage optimaux (appelé « gold » dans les différentes figures) alors que la deuxième moitié a été utilisée comme base de données de test. Ces exemples « gold » ont été utilisés de la même manière que des exemples collectés avec une interface. Ils nous ont servis de point de comparaison permettant de définir le taux de reconnaissance maximal

atteignable par notre système de reconnaissance avec nos conditions expérimentales.

Comme montré sur la figure 7.8, nous pouvons remarquer tout d'abord que les exemples collectés à l'aide de l'interface *iPhone* permettent au système d'apprentissage d'obtenir un taux de reconnaissance significativement plus élevé qu'avec les trois autres interfaces. Nous pouvons en particulier noter qu'en moyenne trois exemples d'apprentissage collectés avec cette interface permettent d'obtenir un aussi bon taux de reconnaissance que huit exemples collectés avec une des autres interfaces. De plus, l'interface *iPhone* semble permettre aux utilisateurs même non-experts de fournir au système d'apprentissage des exemples d'une qualité quasiment similaire à la qualité des exemples « gold » collectés par un utilisateur expert. Nous pouvons également remarquer que même avec très peu de « très bons » exemples (comme trois ou quatre) nous pouvons obtenir un taux de reconnaissance pour 12 objets différents assez élevé (environ 70% avec l'interface *iPhone*). Il est aussi intéressant de noter que le score le plus faible a été obtenu avec les exemples collectés avec l'interface *Gestes*. Ce résultat peut sans doute s'expliquer par l'utilisabilité très faible de cette interface (plus de détails seront donnés sur ce problème dans la section suivante).

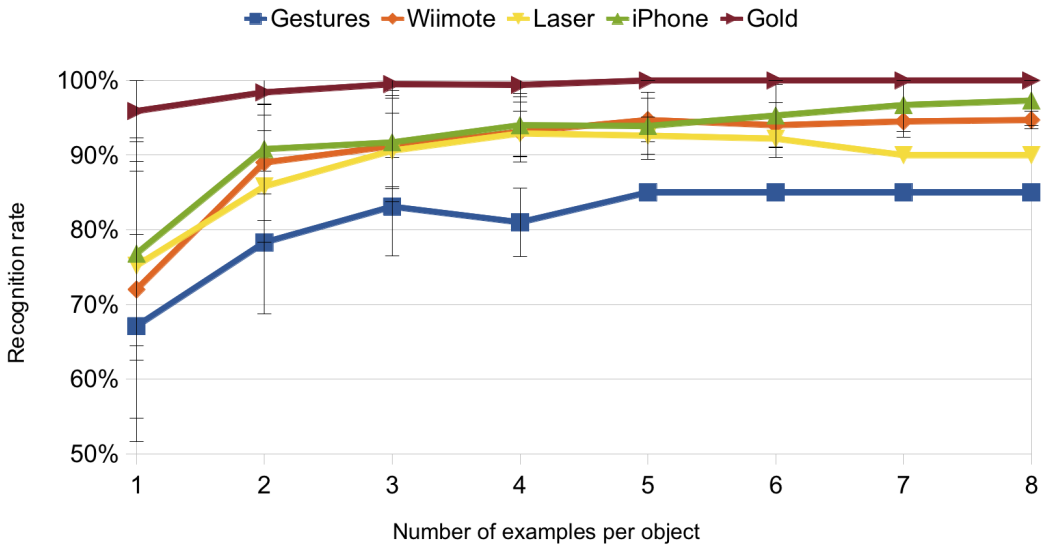


FIGURE 7.9 – **Taux de reconnaissance pour les cinq « gros » objets :** Nous pouvons remarquer que les trois interfaces basées sur des objets médiateurs permettent d'obtenir des exemples d'une qualité similaire. Pour les objets de taille importante l'interface ne semble pas avoir un impact majeur sur le taux de reconnaissance.

Comme dans la section précédente, nous avons également séparé ici les 12 objets en deux sous-catégories : les gros et les petits objets. Comme nous pouvons le voir sur la figure 7.9, le taux de reconnaissance pour ces gros objets est très élevé (aux alentours de 90%) pour les interfaces basées sur les objets médiateurs (*iPhone*, *Wiimote* et *Laser*). Il n'y a pas de différence significative entre ces interfaces. Par

contre, nous pouvons voir sur la figure 7.10 que pour les petits objets nous obtenons des résultats significativement plus élevés avec l'interface *iPhone* qu'avec les trois autres interfaces. Ainsi, alors que l'interface ne semble pas jouer un rôle prépondérant dans la qualité des exemples d'apprentissage et donc dans la reconnaissance des gros objets, les interfaces telles que l'interface *iPhone* semblent permettre aux utilisateurs non-experts d'obtenir un taux de reconnaissance significativement plus élevé pour les petits objets et plus particulièrement lorsque notre système d'apprentissage est entrainé avec très peu d'exemples. Or, ces conditions - peu d'exemples et des objets relativement petits - sont probablement le type d'utilisations auxquelles nous pouvons nous attendre, dans un contexte d'interaction humain-robot domestique, et ces résultats sont donc particulièrement intéressants. Enfin, nous pouvons noter que les résultats quantitatifs sont bien cohérents avec les résultats qualitatifs présentés précédemment.

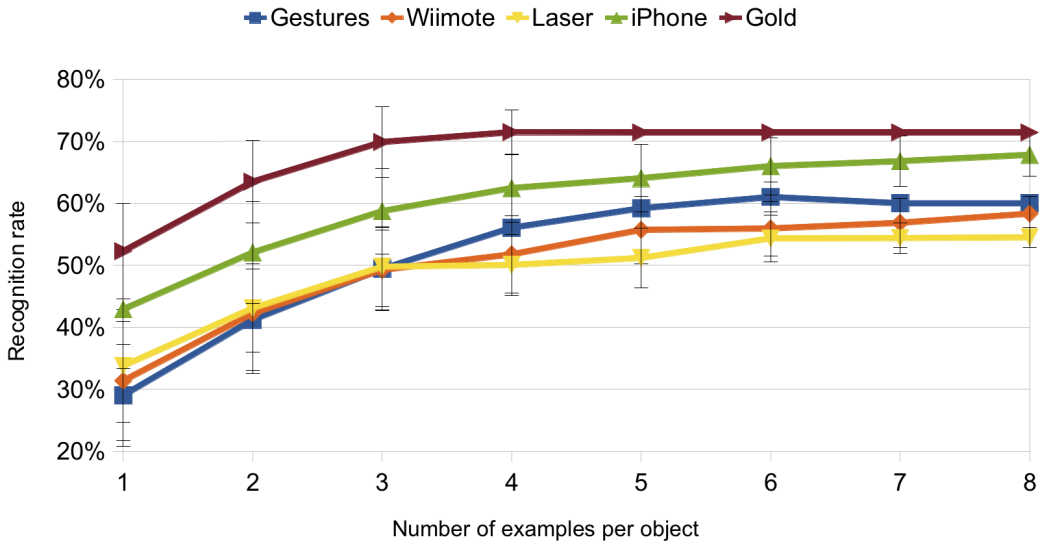


FIGURE 7.10 – **Taux de reconnaissance pour les sept « petits » objets :** Nous pouvons remarquer que l'interface *iPhone* permet aux utilisateurs de collecter des exemples d'apprentissage d'une qualité supérieure à ceux collectés avec les trois autres interfaces. Elle permet donc d'obtenir un taux de reconnaissance plus élevé (plus particulièrement avec peu d'exemples d'apprentissage). Les trois autres interfaces ont obtenu des scores approximativement similaires.

Dans les tests décrits ci-dessus, l'image entière était utilisée comme exemple. Nous ne tirions pas profit de la zone entourée par les utilisateurs avec les interfaces *iPhone* ou *Laser*. Nous avons également étudié à quel point encercler a un impact sur les performances du système dans son ensemble. Comme nous pouvons le voir sur la figure 7.11, encercler avec *l'iPhone* permet d'améliorer le taux de reconnaissance, plus particulièrement lorsque le système est entrainé avec peu d'exemples d'apprentissage. Nous pouvons remarquer que le taux de reconnaissance est entre

5 et 10% plus élevé lorsque seule la zone entourée de l'image est utilisée. Il faut également noter que l'environnement de test était composé de quelques objets mais n'était probablement pas aussi complexe que peut l'être un environnement du quotidien. Dans de telles conditions, encercler aurait probablement un impact encore plus important.

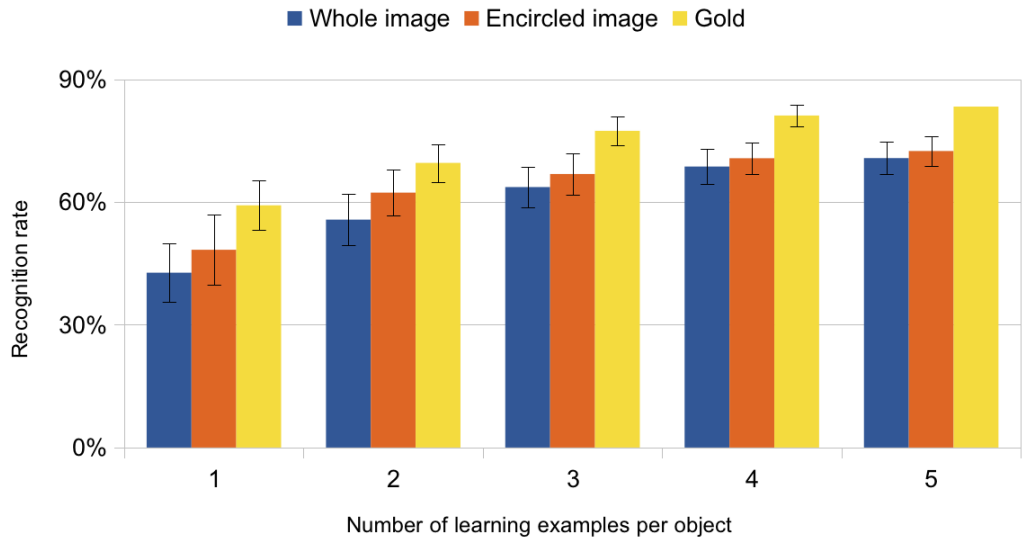


FIGURE 7.11 – Cette figure présente l'impact de l'encerclement avec l'interface *iPhone* sur le taux de reconnaissance. Comme nous pouvons le voir ce geste naturel permet de significativement améliorer le taux de reconnaissance, plus particulièrement lorsque le système est entrainé avec peu d'exemples d'apprentissage.

Nous avons également étudié les performances obtenues en utilisant uniquement la partie entourée des images collectées avec le pointeur laser. Cependant, dans ce cas, nous n'avons pas trouvé de différence significative entre les conditions images entières et images entourées. En regardant de plus près les images collectées avec cette interface nous pouvons constater que l'encerclement est correct dans la plupart des cas et devrait donc théoriquement améliorer les résultats. Cependant, le tracé du point de laser « coupe » fréquemment l'objet, perdant ainsi une partie de l'information (voir les exemples de la figure 6.9 présentée au chapitre précédent).

Les résultats présentés ci-dessus permettent de constater que l'interface a effectivement un impact important sur les performances de reconnaissance obtenues lorsque des utilisateurs non-experts montrent des objets visuels à leur robot personnel. Nous pouvons notamment remarquer que l'interface a un impact très fort sur la qualité des exemples d'apprentissage que l'utilisateur fait collecter à son robot. En effet, nous avons, dans un premier temps, montré que sans fournir de retour sur ce que perçoit le robot à l'utilisateur seulement 50% des exemples collectés avec ce type d'interface peuvent être considérés comme « bons », c'est-à-dire où l'objet était entièrement visible. Nous avons également montré qu'encercler, en particulier

avec l'interface *iPhone*, permet d'améliorer le taux de reconnaissance significativement. Enfin, nous avons montré que trois exemples de très bonne qualité, comme ceux collectés avec l'interface *iPhone*, permettent d'obtenir en moyenne un taux de reconnaissance équivalent à huit exemples collectés avec une autre interface. Ces différents résultats nous semblent particulièrement intéressants et importants dans un contexte d'utilisation où les participants souhaiteront probablement ne donner que très peu d'exemples d'apprentissage à leur robot car cette tâche pourrait très rapidement devenir répétitive.

Nous avons également montré que les interfaces comme l'interface *Laser* ou l'interface *iPhone* conçue spécialement pour cette tâche, permettent à des utilisateurs non-experts de s'assurer que l'objet qu'ils veulent montrer au robot est effectivement dans son champ de vision grâce à différents retours de ce que le robot perçoit. Alors qu'il était bien sûr attendu que fournir de tels retours à l'utilisateur améliorerait l'utilisabilité de nos interfaces et permettrait donc aux utilisateurs de collecter des exemples de meilleure qualité, il est très intéressant de voir que seuls les exemples collectés avec l'interface *iPhone* permettent effectivement d'améliorer significativement les performances du système d'apprentissage. En effet, le score obtenu avec les exemples de l'interface *laser* est à peu près équivalent au score obtenu avec les autres interfaces. Nous pouvons donc constater que le type de retour fourni aux utilisateurs influence fortement la qualité réelle des exemples d'apprentissage. Plus précisément, alors que l'interface *laser* permet aux utilisateurs de savoir si un objet est visible ou non, elle ne donne aucune information sur la façon dont un objet est vu par le robot. Comme nous pouvons le voir sur les exemples de la figure 7.12, beaucoup d'exemples collectés avec cette interface ont été capturés soit loin de l'objet (l'objet est donc très petit sur l'image), soit avec un angle de vue tel que l'objet soit en arrière plan de l'image avec d'autres objets devant.

Ainsi, pour permettre à des utilisateurs non-experts d'apprendre à leur robot personnel à reconnaître des objets visuels avec très peu d'exemples d'apprentissage, nous pensons que l'interface est un paramètre à réellement prendre en compte et qu'elle doit être pensée spécialement pour cette tâche. En effet, les utilisateurs non-experts sont très enclins à faire de fausses suppositions quant à l'appareil visuel d'un robot humanoïde (e.g. supposer qu'il a un champ de vision comparable à celui d'un humain). L'interface doit donc leur permettre de mieux comprendre ce que le robot perçoit mais également les pousser à faire attention aux exemples d'apprentissage qu'ils font collecter à leur robot. Par exemple, l'interface *iPhone*, en présentant directement à l'utilisateur sur l'écran de l'appareil l'exemple d'apprentissage que l'utilisateur doit entourer, force les utilisateurs à surveiller la qualité des exemples. Les humains n'ont pas nécessairement besoin de comprendre les algorithmes d'apprentissage utilisés par le robot pour pouvoir influer sur la qualité des exemples. L'interface peut donc canaliser les interactions de l'utilisateur et ainsi lui permettre, plus ou moins consciemment, d'être un « bon » pédagogue.

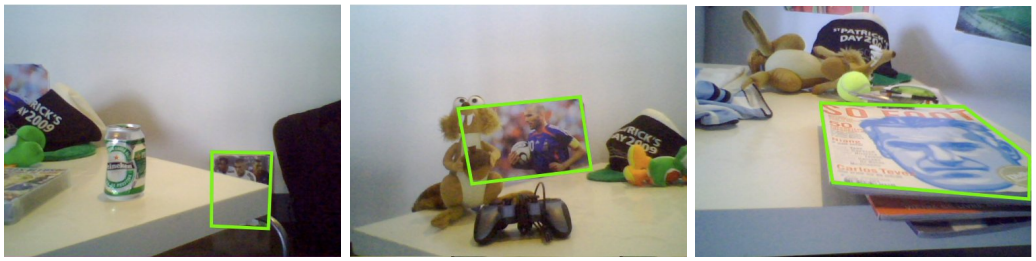


FIGURE 7.12 – Alors que le retour de ce que le robot perçoit fourni par l'interface *laser* permet aux utilisateurs de s'assurer que l'objet qu'ils veulent montrer à leur robot est effectivement visible, cette interface ne donne pas d'indication sur la façon dont l'objet est effectivement perçu par le robot. Par exemple, nous pouvons voir que le jeu vidéo sur l'image de gauche est presque entièrement caché par la table. Sur l'image du milieu nous pouvons voir un premier plan très chargé devant un poster de Zidane (l'objet a montré ici) et finalement sur l'image de droite nous pouvons voir que le magazine a été pris en photo avec un angle de vue quasiment horizontal et est donc très difficilement reconnaissable.

7.4.3 Utilisabilité perçue et expérience utilisateur

La figure 7.13 regroupe les réponses des participants aux questionnaires d'utilisabilité. Nous avons réalisé une analyse de variance à une variable (« one-way ANOVA ») et trouvé des différences statistiques significatives pour les questions Q1 ($F_{3,103} = 6.35, p < 0.001$), Q2 ($F_{3,103} = 2.44, p < 0.05$), Q3 ($F_{3,103} = 6.41, p < 0.001$) et Q6 ($F_{3,103} = 3.38, p < 0.05$) (voir la section 7.3.3 pour la définition de ces questions). Les tests « Tukey post-hoc » ont montré que l'interface *iPhone*, *Wiimote* et *Laser* ont été jugées plus faciles à apprendre et plus pratiques pour déplacer le robot que l'interface basée sur les *Gestes*. Les utilisateurs ont également indiqué qu'il était plus facile de faire regarder un objet au robot avec l'interface *iPhone* et l'interface *Wiimote*. Enfin, ils ont jugé que l'interface *iPhone* était plus agréable à utiliser que l'interface *laser*.

Il est également important de noter aussi que les résultats obtenus par l'interface *Gestes* montrent que même s'il existait des algorithmes de reconnaissance et d'interprétation de gestes aussi performants qu'un humain, cette interaction resterait très compliquée ! L'apprentissage social et plus spécifiquement les situations d'attention partagée semblent particulièrement complexe avec ce type d'interactions. En particulier, au cours des expériences nous avons observé que certains utilisateurs testant l'interface basée sur les *Gestes* se sont parfois dépêchés de finir l'expérience, car des problèmes d'incompréhension du comportement du robot les décourageaient (bien que dans ce cas, le comportement du robot était contrôlé par un humain qui faisait du mieux qu'il pouvait).

La figure 7.14 présente les résultats des questionnaires concernant l'évaluation du jeu robotique. La seule différence statistiquement significative a été trouvée à

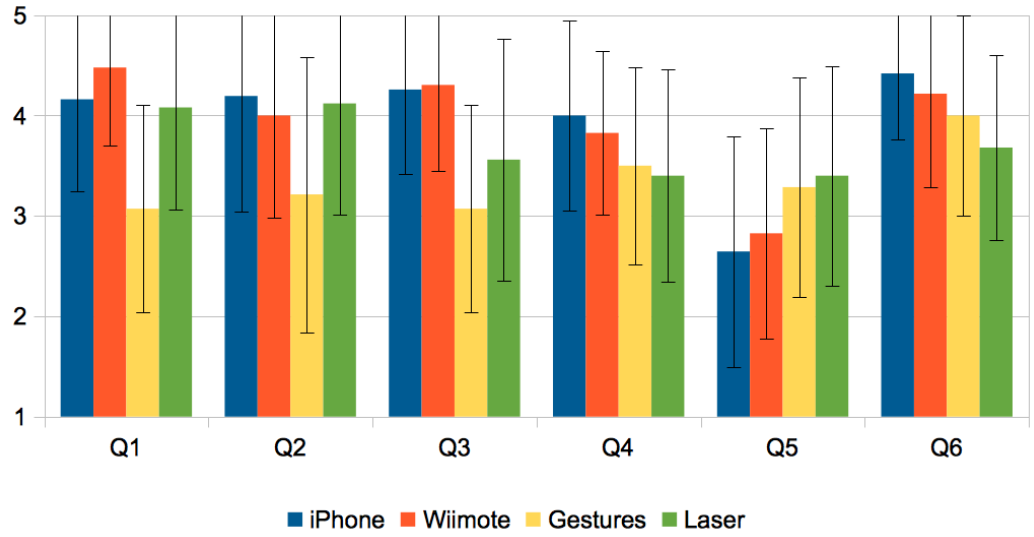


FIGURE 7.13 – **Utilisabilité :** Les participants ont trouvé que l'interface basée sur les *Gestes* était significativement moins intuitive et plus complexe à utiliser que les autres interfaces. Ils ont aussi indiqué que l'interface *iPhone* était plus agréable à utiliser que l'interface *Laser*. L'ordonnée de la figure représente le degré d'accord des participants avec les différentes affirmations (e.g. un score de 5 équivaut à tout à fait d'accord et 0 à pas du tout d'accord).
Q1 : Il était facile d'apprendre à utiliser cette interface.
Q2 : Il était facile de déplacer le robot.
Q3 : Il était facile de faire regarder un objet au robot.
Q4 : Interagir avec le robot était facile.
Q5 : Le robot mettait du temps à réagir.
Q6 : L'interface était plaisante à utiliser.

la question Q1. Nous pouvons voir que les participants ont trouvé le jeu plus facile à finir avec les interfaces basées sur des objets médiateurs (*iPhone*, *Wiimote* et *Laser*) plutôt qu'avec l'interface *Gestes* ($F_{3,103} = 5.17, p < 0.005$). Le jeu a été jugé comme distrayant par les participants indépendamment de l'interface utilisée. Il est également aussi intéressant de noter que l'interface *Gestes* semble augmenter le sentiment de collaboration avec le robot. Les utilisateurs ayant participé au jeu avec cette interface semblent également plus enclin à participer à d'autres jeux robotiques à l'avenir. Cependant ces résultats ne sont pas statistiquement significatifs.

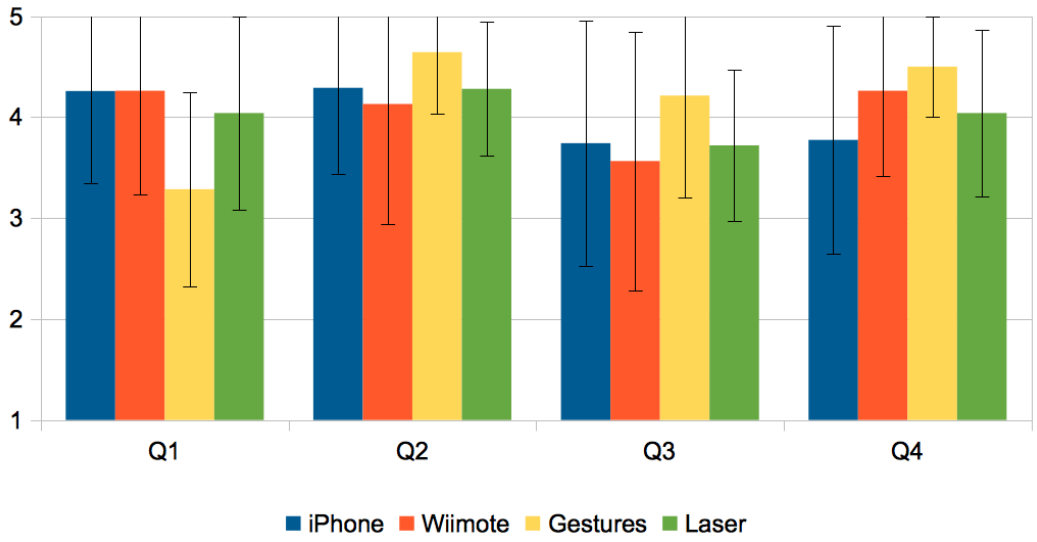


FIGURE 7.14 – **Jeu robotique :** Notre jeu robotique a été jugé comme distrayant par tous les participants. Ils ont trouvé que le jeu était plus complexe à finir avec l'interface gestuelle. Cependant cette interface semble augmenter le sentiment de collaborer avec le robot.
Q1 : Terminer le jeu était facile.
Q2 : Le jeu était distrayant.
Q3 : J'ai eu l'impression de faire équipe avec le robot.
Q4 : Je m'imagine jouer à d'autres types de jeux robotiques à l'avenir.

Il est intéressant de remarquer que bien que l'interface *Gestes* soit jugée comme nettement moins utilisable que les trois autres interfaces, les participants ont quand même jugé que le jeu était aussi amusant avec cette interface qu'avec les autres. Selon nous, ce résultat peut être expliqué par plusieurs facteurs. Tout d'abord, il est important de noter que les participants ne savaient pas s'ils collectaient des bons ou des mauvais exemples d'apprentissage. En particulier, les participants ayant collecté uniquement de très mauvais exemples d'apprentissage pouvaient toujours juger que le jeu était facile et qu'ils l'ont brillamment réussi. Ensuite, alors que les interfaces telles que l'interface *iPhone* ont été spécialement conçues pour permettre à un utilisateur de collecter des bons exemples d'apprentissage, ces interfaces étaient

probablement peu adaptées à un jeu robotique. Les utilisateurs devaient en effet, à la fois surveiller le robot, l'interface de jeu ainsi que l'écran de l'iPhone. De plus, il semble que pour ce type de jeu, l'interface doit être aussi transparente que possible, comme l'interface *Gestes*, afin de permettre aux utilisateurs de se concentrer uniquement sur le jeu lui-même. Enfin, l'interface *Gestes* semble améliorer l'impression de collaborer avec le robot. Nous pensons que ce résultat peut s'expliquer par le fait que les utilisateurs avaient tendance à se tenir plus près du robot avec cette condition. Ils essayaient différents gestes afin de voir comment le robot réagissait et donc pour savoir lesquels étaient les mieux compris. Le biais introduit par l'utilisation d'un protocole magicien d'Oz où le magicien adaptait réellement son comportement à l'utilisateur peut également avoir renforcé ce sentiment de collaboration. Bien que des études supplémentaires devraient être conduites dans cette direction, nos résultats préliminaires semblent indiquer que l'interface *Gestes* pourrait être intéressante dans le développement de jeux robotiques simples. Cependant, l'utilisation de ce type d'interface pose le problème pratique de la conception d'algorithmes de reconnaissance gestuelle dont les performances seraient comparables à celles obtenues par un humain.

7.4.4 Autres mesures

Comme indiqué précédemment, nous avons également chronométré les temps de parcours des utilisateurs. Cependant, en raison de l'extrême variabilité des temps de parcours entre utilisateurs nous n'avons pas trouvé de différence suffisamment significative entre les interfaces pour pouvoir en tirer des conclusions. Nous n'avons pas non plus observé de différences majeures entre les participants recrutés à Cap Sciences et les participants recrutés sur le campus. De même, nous n'avons pas trouvé de corrélation forte entre les différents critères sociologiques (âge, genre) et les performances obtenues.

Chapitre 8

ASMAT : Collecte semi-automatique de nouveaux exemples d'apprentissage

Sommaire

Résumé du chapitre

Nous présentons, dans ce chapitre, une approche de collecte semi-automatique de nouveaux exemples d'apprentissage par un robot. Avec ce système, lorsque l'utilisateur montre un objet visuel à un robot, ce dernier tourne automatiquement autour de l'objet, pour prendre d'autres photos avec des points de vues différents. Nous présenterons une évaluation exploratoire qui montre l'intérêt potentiel de cette méthode. Nous discuterons aussi des limites de cette approche, et notamment du problème de la construction d'un modèle d'un objet visuel utilisé simultanément pour le détecter.

8.1 Collecte semi-automatique d'exemples

Dans les chapitres précédents, nous avons présenté un système permettant à un utilisateur de faire collecter des exemples d'apprentissage à son robot à l'aide d'une interface basée sur des objets médiateurs. Par contre, une des limites de ce système est le fait que pour chaque intervention des utilisateurs un seul exemple est collecté. Cependant, l'aspect visuel d'un objet peut drastiquement changer en fonction de l'angle de vue avec lequel il est perçu, et donc pour pouvoir être reconnu indépendamment du point de vue, différents exemples doivent être collectés [Robbel 2007] (voir la figure 8.2).

Dans ce chapitre, nous présentons une étude préliminaire d'un système que nous avons commencé à explorer. Il permet au robot de collecter automatiquement d'autres exemples d'apprentissage à partir de celui fourni par l'utilisateur. Cette approche a pour but de permettre au robot de se construire un modèle d'un objet plus robuste tout en réduisant le nombre d'interventions de l'utilisateur. Elle a été publiée dans [Rouanet 2009c].

Ce système est basé sur l'interface iPhone présentée au chapitre précédent car c'est celle qui offre le plus de possibilités d'interactions grâce à l'écran tactile. Lorsque les utilisateurs encerclent un objet sur l'écran de l'iPhone, au lieu de simplement collecter un exemple, le robot va tourner autour de l'objet et prendre plusieurs photos de l'objet avec différents points de vues. Cela nécessite que le robot se construise incrémentalement un modèle de l'objet utilisé simultanément pour pouvoir le suivre. Un système similaire a déjà été développé par Dowson et al. [Dowson 2005]. Cependant, leur système est utilisé sur des enregistrements vidéo alors qu'ici nous pouvons directement contrôler la caméra en déplaçant le robot. Nous avons donc appelé cette technique ASMAT (*active simultaneous modelling and tracking*). Notre système d'apprentissage de classification / reconnaissance d'objets visuels étant incrémental, nous avons pu l'utiliser ici.

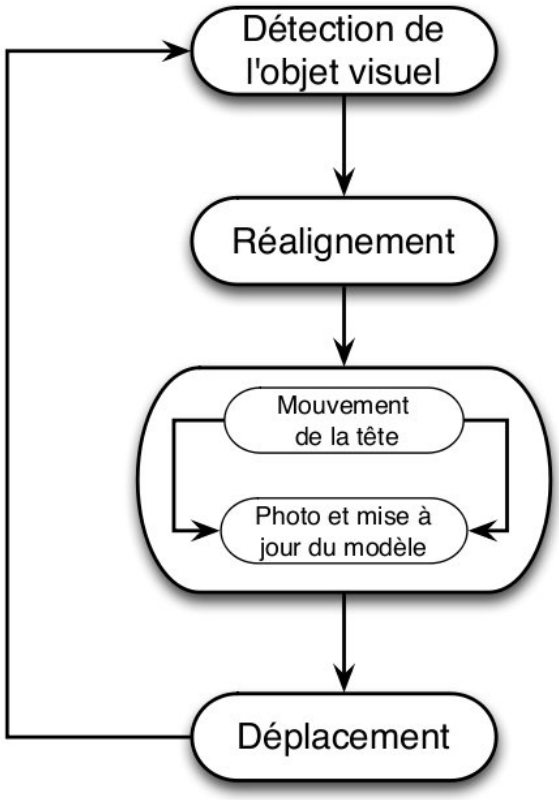


FIGURE 8.1 – Schéma de fonctionnement de l'algorithme ASMAT

De manière plus précise, avec notre approche le robot tourne incrémentalement autour d'un objet (le robot utilisé ici est toujours le Nao). Pour passer d'une position à l'autre, le robot se déplace latéralement et tourne sur lui même afin de conserver l'objet suivi au centre de sa vision. Pour chacune des positions, l'orientation de la tête du robot est légèrement modifiée et une photo de l'objet est prise.

FIGURE 8.2 – Notre système ASMAT (*active simultaneous modelling and tracking*) permet de collecter automatiquement différentes images d'un objet avec différents points de vue en faisant tourner le robot autour de l'objet. Cela permet au robot de se construire automatiquement un modèle de l'objet plus robuste.

Pour chacune de ces images, on recherche tous les points SURF (voir la section 6.2.1) correspondant à notre modèle de l'objet. Un filtre est appliqué pour enlever les points trop éloignés du centre de gravité des points détectés. La boite englobante de ces points définit la zone de l'image où l'objet est présent. Cette partie de l'image est utilisée pour mettre à jour le modèle. Le processus exact est décrit plus en détails par l'algorithme 2. Le schéma de la figure 8.1 résume le fonctionnement globale de l'algorithme.

8.2 Problème de dérive et asservissement supervisé

L'utilisation du modèle construit simultanément pour détecter l'objet entraine un fort risque de dérive exponentielle. En effet, une erreur dans la détection de l'objet entrainera la mise à jour du modèle avec de fausses valeurs et donc une détection encore plus fausse etc. Afin de contourner ce problème difficile de manière très simple, nous affichons sur l'écran de l'iPhone la boite englobante décrivant la zone de l'image représentant l'objet suivi. L'utilisateur peut alors surveiller la procédure et l'interrompre, simplement en touchant l'écran de l'appareil, dès que le système dérive. Il peut ensuite relancer la collecte d'exemples en ré-encerclant l'objet s'il le souhaite.

Algorithm 2 ASMAT(user_encircled_image)

```
keypoints ← extract_keypoints(user_encircled_image)
update_object_model(keypoints)

while not user_stop() and i < N do
  for j in 1 to M do
    move_robot_head()
    keypoints ← extract_keypoints(robot_camera)
    matches ← find_matching_object_model(keypoints)
    elected ← filter_isolate_points(matches)
    bb ← compute_bounding_box_from_points(elected)
    for each kp in keypoints inside bb do
      update_object_model(kp)
    end for
  end for

  walk_step_around_object()
  keypoints ← extract_keypoints(robot_camera)
  matches ← find_matching_object_model(keypoints)
  center ← compute_gravity_center(matches)
  robot_center_sight(center)
  i ← i + 1
end while
```

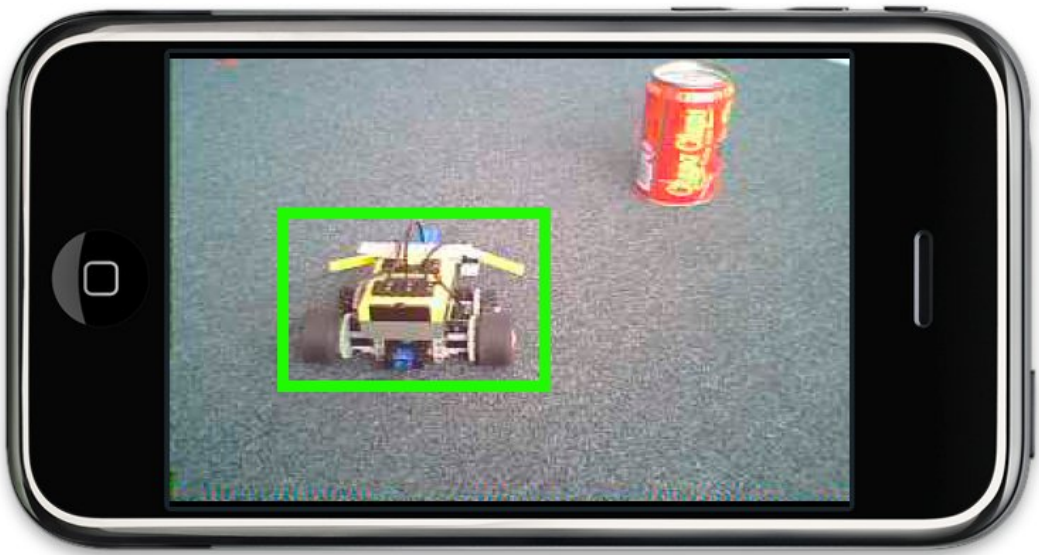


FIGURE 8.3 – Les utilisateurs peuvent visualiser le processus de suivi de l'objet et ainsi interrompre la collecte d'exemples (simplement en touchant l'écran de l'appareil) dès que le système dérive.

8.3 Évaluation

Nous avons réalisé une première étude très exploratoire de notre système afin d'en évaluer le potentiel et en particulier voir si cette approche permet de réduire significativement le nombre d'interventions de l'utilisateur. Nous avons réalisé une expérience très simple, où l'utilisateur devait montrer quatre objets différents au robot à l'aide de l'interface iPhone. Pour chaque exemple, le robot collectait automatiquement 25 autres exemples (le robot se déplaçait cinq fois autour de l'objet, et pour chacune des positions, sa tête était déplacée cinq fois). Nous avons ensuite comparé les performances obtenues en utilisant comme base d'entrainement de notre système d'apprentissage soit uniquement l'image entourée par l'utilisateur, soit l'ensemble des images collectées automatiquement. Les utilisateurs étaient ici des personnes travaillant dans notre laboratoire.

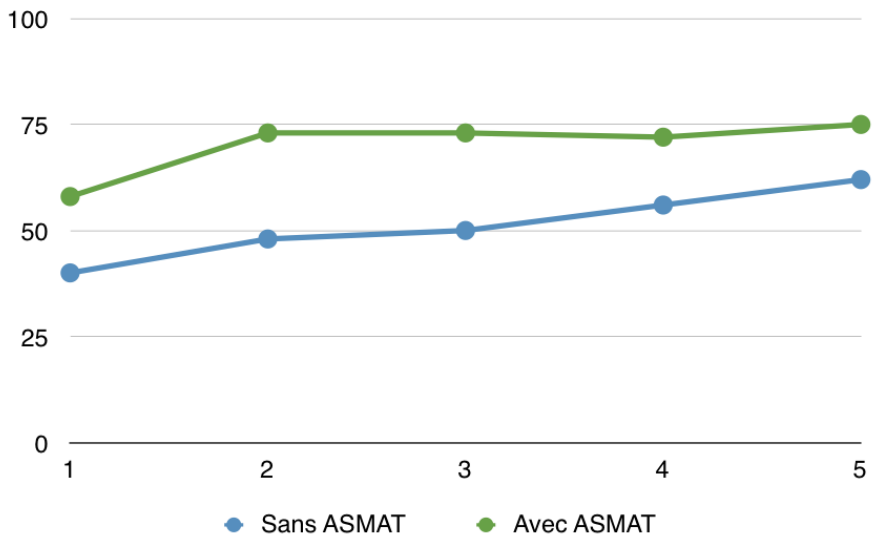


FIGURE 8.4 – Le système ASMAT permet de réduire le nombre d'interventions de l'utilisateur requises pour obtenir un taux de reconnaissance équivalent. En particulier, ici deux interventions avec le système permettent d'obtenir un score plus élevé que cinq interventions sans.

Comme nous pouvons le constater sur la figure 8.4, le système ASMAT permet d'atteindre très rapidement ce qui semble être la valeur maximum ($\sim$ 80%) avec seulement deux interventions de l'utilisateur alors que cinq interventions sans ce système ne permettent pas d'atteindre le maximum.

8.4 Limites de cette approche

L'approche présentée ici, bien qu'intéressante, reste très exploratoire. En particulier, notre système ne fonctionne pas en milieu non-contraint. Avec le Nao, les mouvements n'étaient pas suffisamment précis pour permettre de suivre l'objet de manière fluide. Il arrivait donc fréquemment que l'objet ne soit tout simplement plus dans son champ de vision.

De plus, nous n'avons inclus aucun système de détection d'obstacles et nous supposons donc simplement que le robot puisse faire le tour de l'objet sans problème. Le système ASMAT n'est donc pas à l'heure actuelle directement utilisable lors d'une interaction en milieu non-contrôlé.

Chapitre 9

Interfaces pour l'enseignement conjoint d'objets visuels et de mots acoustiques nouveaux

Sommaire

Résumé du chapitre

Nous présenterons ici une extension de notre système autorisant l'enseignement conjoint à un robot de mots acoustiques nouveaux associés à des objets visuels. Nous décrirons en détails les nouvelles interactions proposées par une interface basée sur un TMC, et en particulier comment l'utilisateur peut entrer un nouveau mot acoustique, lors de la capture d'un exemple d'apprentissage. Nous présenterons également le système de perception et de comparaison de mots acoustiques (basé sur une représentation du son à l'aide de coefficients MFCCs et de la distance DTW) que nous avons conçu. Nous avons couplé et intégré ce système à l'interface pour maximiser

ses performances et son utilisabilité. Nous montrerons, en particulier, comment l'interface peut permettre d'améliorer les performances du système de reconnaissance vocale, en proposant à l'utilisateur de choisir, parmi un ensemble de résultats les plus proches, l'objet correspondant réellement à sa requête. L'utilisation de mots vocaux dont les invariants ne sont pas connus au départ soulève le problème de la catégorisation des mots acoustiques. Nous montrerons comment l'interface peut, à travers des interactions simples et peu nombreuses, permettre à un utilisateur de regrouper incrémentalement les exemples d'apprentissage se rapportant à un même objet. L'effet de ces différentes interactions sur les performances du système d'apprentissage sera évalué à travers des simulations de l'utilisation de l'interface.

9.1 Extension du système existant aux mots acoustiques

Dans les chapitres précédents, nous avons présenté un système permettant à un utilisateur d'apprendre à un robot à reconnaître des objets visuels nouveaux. Dans cette thèse, nous étudions la question plus générale de l'enseignement conjoint de mots et d'objets visuels nouveaux à un robot. Comme nous l'avons expliqué au chapitre 6, nous supposions jusqu'à présent, la possibilité pour l'utilisateur d'associer des mots, sous la forme de symboles/étiquettes, à ces objets visuels. Nous pouvions ainsi contourner les problèmes de perception et reconnaissance des mots, afin de nous focaliser sur les questions liées à la catégorisation/classification d'objets visuels.

Nous avons décrit, en section 6.2.3, deux approches imaginées pour la saisie de ces étiquettes : l'utilisation d'un clavier et la simulation d'un système de reconnaissance vocale basée sur un protocole de type magicien d'Oz. Nous allons décrire, dans les sections ci-dessous, les limites de ces approches, et justifier notre choix de développer notre propre système de reconnaissance de mots acoustiques.

9.1.1 Limites de l'utilisation d'un clavier : utilisabilité et expérience utilisateur

Nous avons proposé d'utiliser un clavier pour permettre aux humains de saisir les mots, sous forme de chaînes de caractères, à associer aux objets visuels qu'ils souhaitent apprendre à un robot. Nous avons présenté un prototype basé sur l'utilisation d'un clavier virtuel affiché sur l'écran d'un TMC. Ce système a été décrit brièvement en section 6.2.3 et publié dans [Rouanet 2009b][Rouanet 2009c]. Représenter les mots sous forme de chaînes de caractères permet de facilement les catégoriser.

Malgré un intérêt certain, cette approche présente aussi des limites importantes. En particulier, comme détaillé au chapitre 1, dans un futur où les robots seront de plus en plus amenés à partager notre quotidien, il est crucial qu'une attention particulière soit portée à l'interaction et à l'interface, afin de permettre à tous les utilisateurs, même non-experts, d'interagir de manière efficace et intuitive avec leur robot personnel. Or, des études pilotes nous ont permis d'identifier que l'utilisation de

clavier (virtuel ou réel) pour entrer les étiquettes au système d'apprentissage posait des problèmes aux utilisateurs. Dans ces expériences, les utilisateurs interagissaient avec un robot à l'aide d'une interface proche de l'interface iPhone décrite en section 6.3.1. Ils devaient montrer des objets visuels à un robot et y associer des noms sous la forme de chaînes de caractères. Les participants remplissaient ensuite des questionnaires d'utilisabilité, portant notamment sur l'utilisation du clavier virtuel pour entrer les noms associés aux objets visuels. Les résultats de cette évaluation ont indiqué que les participants jugeaient l'utilisation du clavier virtuel contraignante. De plus, ils ont indiqué que, selon eux, ce mode d'interaction n'était que peu adapté à la situation [Rouanet 2009b].

Dans cette étude pilote, les utilisateurs ont aussi indiqué préférer utiliser un système vocal. Ce mode d'interaction est, en effet, naturellement utilisé par les humains pour désigner des objets à leurs enfants. Lors de nos expériences, la plupart des utilisateurs ont « parlé » au robot même lorsqu'il était clairement indiqué que le robot ne comprenait pas ce qu'ils disaient ou même ne les entendaient pas [Rouanet 2010a]. L'utilisation de mots vocaux pour nommer les objets visuels semble donc correspondre aux attentes des utilisateurs. Ce mode d'interaction est de plus intuitif, et ne nécessite pas d'apprentissage et/ou d'explication, même pour les utilisateurs non-experts.

9.1.2 Limites des systèmes de reconnaissance vocale

Suivant les résultats de ces expériences pilotes, nous avons étudié la possibilité d'enseigner à un robot des mots acoustiques nouveaux associés à des objets visuels. Dans l'expérience décrite au chapitre 7, les utilisateurs devaient prononcer des mots vocaux lorsqu'ils montraient des objets au robot, afin qu'il apprenne à les reconnaître. Pour pouvoir catégoriser et reconnaître ces sons, nous avons simulé l'utilisation d'un système de reconnaissance vocale, transformant les sons associés aux objets visuels en symboles/identifiants. Dans un souci de robustesse et de facilité d'implémentation, nous avons mis en place un protocole de type magicien d'Oz. Plus précisément, un humain expert associait un identifiant unique aux sons enregistrés par les participants. Nous pouvions ainsi bénéficier d'un système de reconnaissance vocale aussi performant qu'un humain.

La reconnaissance vocale reste en effet un problème difficile. Cette question a fait l'objet de très nombreuses recherches [Jurafsky 2000]. Des descripteurs de la parole, tels que les coefficients MFCCs (*Mel-frequency cepstrum coefficients*) [Zheng 2001] ou RASTA-PLP (*RelAtive SpecTrAl - Perceptual Linear Predictive* [Hermansky 1992]) ont été proposés. Ils permettent de représenter un son de manière robuste et compacte. Des méthodes, telles que la distance DTW (*dynamic time warping*), ont aussi été introduites permettant la comparaison des séquences de ces descripteurs sonores [Sakoe 1978][Bahl 1983]. L'utilisation de modèles cachés de Markov a aussi été très largement utilisée dans le domaine de la reconnaissance vocale [Rabiner 1986][Juang 1991]. Ces modèles statistiques permettent d'encoder les séquences de sons représentées sur une échelle de temps à court-terme.

Cependant, les systèmes de reconnaissance vocale souffrent encore de certaines limitations. Tout d'abord, ces systèmes utilisent généralement la structure des phrases ou des suites de mots (par ex. la probabilité de la présence d'un mot après un autre) pour améliorer la robustesse de la reconnaissance. Ils sont donc moins performants pour traiter des mots isolés. Or c'est ce type d'exemple qui nous intéresse dans cette thèse. De plus, les systèmes de reconnaissance vocale se basent également sur des corpus de mots pré-définis [Perzanowski 2001][Lauria 2002]. La reconnaissance vocale est alors limitée aux mots appartenant à ces corpus. Il nous semble néanmoins intéressant de fournir la possibilité aux utilisateurs d'apprendre des types de mots acoustiques très variés à un robot. Ces systèmes peuvent aussi être sensibles au locuteur. L'utilisateur doit alors entrainer la reconnaissance vocale avec sa propre voix avant de pouvoir l'utiliser. Nous allons donc proposer, dans la section suivante, de développer notre propre système de reconnaissance vocale afin de contourner certaines de ces limites. Plus précisément, nous allons étudier comment l'interface, si elle est bien conçue, peut permettre à l'utilisateur de participer au processus de reconnaissance, et ainsi d'en améliorer les performances.

9.1.3 Développement d'un système de perception/reconnaissance de mots acoustiques non nécessairement linguistiques

Comme expliqué ci-dessus, nous souhaitons fournir la possibilité à l'utilisateur d'entrer n'importe quel type de mots acoustiques. Il pourrait, en effet, être intéressant que l'humain puisse utiliser des mots vocaux, comme le nom de l'objet, mais aussi des noms propres, des mots inventés ou même des mots non-linguistiques tels que des onomatopées. Nous avons donc choisi de développer notre propre système de reconnaissance vocale permettant de reconnaître ces différents types de mots acoustiques isolés. En plus d'améliorer l'utilisabilité et l'expérience utilisateur, l'emploi de mots acoustiques, sans supposer l'utilisation de symboles, soulève aussi des questions de recherche intéressantes.

Enseigner conjointement des objets visuels associés à des mots acoustiques à un robot requiert, tout d'abord, d'avoir des mécanismes de perception et de regroupement des exemples d'objets visuels et des mots acoustiques. Ensuite, il faut pouvoir extraire des catégories de ces différents groupes (visuels et acoustiques) et enfin créer des associations entre les catégories visuelles et sonores. Ces différentes problématiques sont résumées par le schéma de la figure 9.1.

Dans les sections suivantes, nous allons présenter un système complet et intégré s'attaquant à l'ensemble de ces problèmes. Nous allons plus précisément étudier le rôle de l'interface dans un tel système d'apprentissage. Nous présenterons tout d'abord les algorithmes de perception et de groupement de mots acoustiques utilisés. Les mécanismes de perception et de reconnaissance visuelle utilisés dans ce chapitre sont identiques à ceux présentés en section 6.2, et ne seront donc pas détaillés à nouveau. Nous présenterons aussi le modèle proposé pour représenter les associations entre les groupes d'objets visuels et les groupes de mots acoustiques.

Nous décrirons ensuite plus particulièrement l'interface utilisateur conçue pour

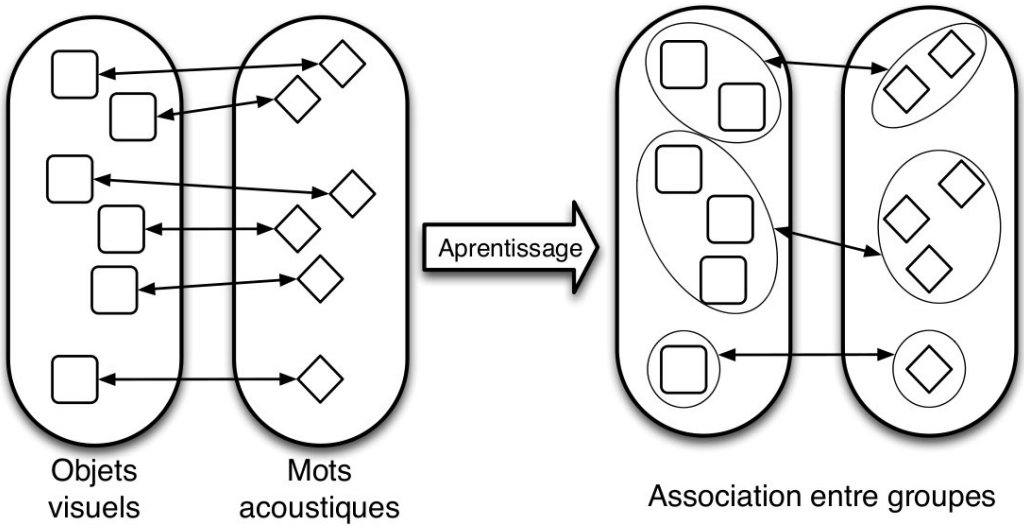


FIGURE 9.1 – Ce schéma présente le problème de l'apprentissage conjoint d'objets visuels et de mots acoustiques ainsi que leurs associations lorsqu'on ne suppose pas l'utilisation de symboles dans l'une de ces deux modalités.

entrer des mots acoustiques, mais aussi aider à la reconnaissance vocale et permettre à l'utilisateur de participer au processus de regroupement des catégories. Enfin, nous présenterons une étude de l'impact de cette interface sur les performances du système d'apprentissage, en reproduisant une tâche où un robot doit déterminer si un objet visuel correspond bien au mot acoustique prononcé par un utilisateur. Cette expérience, ainsi que l'utilisation de l'interface, ont été simulées pour des raisons pratiques. En effet, nous n'avons pas eu le temps de réaliser d'études utilisateurs de ces interactions durant cette thèse.

9.2 Description du système

Le système décrit ici se base sur les mêmes algorithmes standards de perception visuelle et d'apprentissage machine que ceux présentés dans les chapitres précédents. La contribution principale de ce chapitre repose sur la conception d'une interface dédiée, qui fournit un ensemble de mécanismes à l'utilisateur, l'aidant à résoudre les problèmes de catégorisation et de recherche soulevés par l'utilisation de mots acoustiques nouveaux. Nous décrirons, dans un premier temps, les algorithmes de perception et reconnaissance de mots acoustiques utilisés ici. Nous allons ensuite décrire comment cette interface permet aux utilisateurs de surveiller le processus d'association. Ils peuvent y participer directement, à travers des interactions transparentes et intuitives leur permettant de fusionner des groupes correspondants à une même catégorie. L'interface permet ainsi au robot de tirer profit des capacités de perception et de clusterisation du cerveau humain. L'interface et l'utilisateur font

partie intégrante du système de reconnaissance vocale présenté ci-dessous. L'impact de ces interactions de l'utilisateur sera évalué et comparé avec des algorithmes standard de clusterisation automatique.

9.2.1 Perception auditive

Les mots acoustiques fournis à notre système d'apprentissage sont représentés à l'aide du descripteur MFCCs (*Mel-frequency cepstrum coefficients*) [Zheng 2001]. Cette représentation est fréquemment utilisée dans des applications de reconnaissance vocale ou d'identification de l'orateur [Reynolds 1994][Chetouani 2009a]. Cependant, ce modèle souffre toujours de problèmes de robustesse au bruit. Nous avons choisi ce descripteur car il représente bien un standard en recherche actuelle et permet donc une comparaison facile avec d'autres travaux. Les coefficients MFCCs représentent l'énergie du spectre d'un son à court terme. Un son est décrit comme la suite temporelle de ces coefficients.

Afin de pouvoir comparer deux sons, et donc deux séquences de coefficients MFCCs, nous utilisons une mesure de déformation temporelle dynamique : DTW (*dynamic time warping*). Cette mesure de distance permet de mesurer la similarité entre deux séquences dont la vitesse peut varier. Plus précisément, cette méthode cherche la correspondance optimale entre deux suites en s'autorisant à supprimer ou insérer des éléments dans les séquences à comparer. Cette distance a été largement utilisée dans des applications de type reconnaissance vocale car elle est robuste aux variations de vitesse ainsi qu'aux informations manquantes [Sakoe 1978].

Pour comparer deux mots acoustiques, nous commençons par extraire la suite des coefficients MFCCs correspondant à chacun de ces sons. Nous utilisons ensuite la mesure DTW pour chercher la distance minimale entre les deux séquences de coefficients. Plus cette distance est faible, plus les deux mots acoustiques comparés sont proches.

Cette mesure de similarité, couplée avec l'interface décrite ci-dessous, constitue notre système de reconnaissance de mots acoustiques. Bien que très simple, il nous permet déjà d'obtenir des performances très satisfaisantes, en particulier lorsqu'il est utilisé avec un nombre de mots différents peu important (12 mots dans les expériences décrites ci-dessous).

9.2.2 Apprentissage machine

Contrairement au système d'apprentissage décrit dans la section 6.2, les représentations visuelles des objets ne sont, ici, plus étiquetées, et il n'est donc plus possible de regrouper directement les exemples se rapportant à un même objet. Nous avons donc proposé un système permettant de regrouper incrémentalement ces différents exemples.

Le système d'apprentissage proposé ici peut être vu comme une machine d'association d'exemples d'apprentissage (la figure 9.2 en est une représentation graphique). L'utilisateur fournit à ce système des couples du type $L_{ex} = (M_i, I_j)$ où

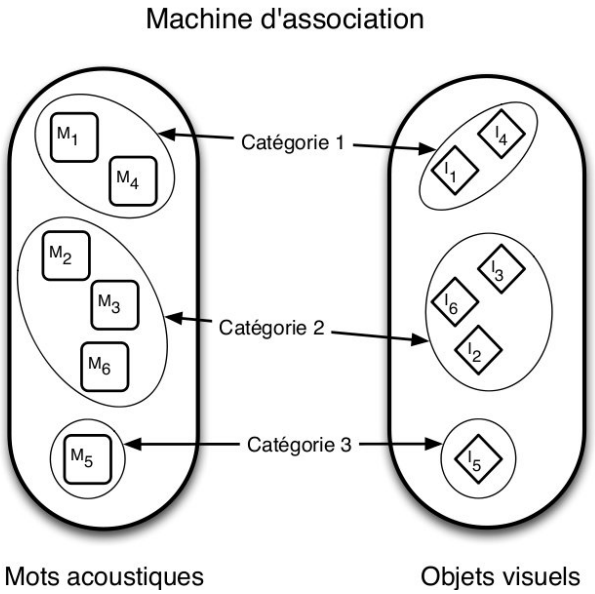


FIGURE 9.2 – Notre machine d'association permet de regrouper les différents couples (M_i, I_j) entre eux. Il est aussi possible de rechercher le groupe $C_l = \{(M_i, I_j)\}$ correspondant à un mot vocal recherché (la catégorie C_l choisie est celle contenant le mot acoustique le plus proche de celui recherché.

M_i est la représentation d'un mot acoustique, telle que décrit dans la section précédente, et I_j est la représentation d'un objet visuel, c'est-à-dire l'histogramme des caractéristiques visuelles SURF après quantification (voir section 6.2 pour plus de détails). Ces différents couples peuvent être regroupés entre eux lorsque le système « estime » qu'ils correspondent réellement à un même objet.

Afin de maximiser les performances en reconnaissance de notre système, il est important que les exemples d'apprentissage, correspondant à un même objet, soient, autant que possible, regroupés entre eux. En effet, comme nous l'avons montré dans les expériences précédentes, pour obtenir de bonnes performances, notre système de reconnaissance visuelle a besoin d'être entrainé avec plusieurs exemples d'apprentissage. Nous expliquerons dans la section suivante comment la conception de l'interface peut jouer un rôle dans ce processus, en permettant aux utilisateurs d'indiquer au système d'association lorsque deux exemples correspondent à un même objet.

Il faut aussi pouvoir demander au système d'association de rechercher le groupe d'exemples correspondant à un mot vocal. En effet, cette requête est nécessaire pour qu'un utilisateur puisse demander au robot de rechercher un objet déjà appris. Pour cela, il prononce le nom acoustique de l'objet à rechercher : M_{search}. Avec une approche naïve, le système d'association irait chercher la catégorie C_l qui contient

le mot le plus proche de celui entré par l'utilisateur :

$$\arg\min_k = DTW(M_{search}, M_k)$$

L'ensemble des images appartenant à la catégorie C_l pourrait alors être utilisé par le système de reconnaissance visuelle pour rechercher l'objet dans l'environnement immédiat du robot :

$$T = \{I_m\}, \exists L_{ex} = (M_k, I_m) \in C_l$$

Cependant, ici les mécanismes d'apprentissage et l'interface sont intégrés en un seul système, où les N résultats les plus proches sont présentés à l'utilisateur via l'interface (voir section 9.2.3.2 pour plus de détails). L'utilisateur peut alors sélectionner le(s) résultat(s) correspondant réellement à sa requête. Il participe ainsi, à travers l'interface, au processus de reconnaissance vocale, et aide le robot à bien reconnaître un mot acoustique.

9.2.3 Interface conçue pour l'aide à la reconnaissance vocale et aux regroupement d'exemples d'apprentissage

L'interface présentée ici est une extension de l'interface iPhone décrite à la section 6.3.1. Cette interface permet de déplacer le robot, d'attirer son attention vers un objet et de déclencher la capture d'un exemple d'apprentissage visuel grâce à différents gestes effectués par les utilisateurs directement sur l'écran tactile de l'appareil. En plus de ces fonctionnalités nous allons ajouter ici les possibilités suivantes :

- prononcer un mot vocal, c'est-à-dire une onde acoustique pouvant, mais pas nécessairement, correspondre à un nom/mot, qui sera associé à la représentation visuelle de l'objet
- demander au robot de rechercher un objet déjà appris en prononçant le mot vocal associé
- regrouper différents exemples d'apprentissage correspondant à un même objet visuel

Comme pour les interfaces précédentes, nous avons essayé de rendre cette interface intuitive, facile à utiliser et peu contraignante, tout en permettant aux utilisateurs de maximiser les performances du système d'apprentissage à travers des interactions peu nombreuses et aussi transparentes que possible.

Le schéma de la figure 9.3 représente la suite des interactions de l'utilisateur lui permettant de fournir un nouvel exemple d'apprentissage à un robot (objet visuel et mot acoustique) ainsi que de demander au robot de rechercher un objet déjà appris. Un exemple de regroupement d'exemples à travers l'interaction de recherche est également présenté. Chacune des interactions va être décrite plus en détails dans les sections ci-dessous.

9.2.3.1 Associer un nouveau mot vocal

Après avoir encerclé un objet sur l'écran de l'appareil afin de demander au robot d'en prendre une photo, apparait maintenant un bouton de type « pousser pour par-

FIGURE 9.3 – Pour fournir un nouvel exemple d'apprentissage (objet visuel et un mot acoustique) à un robot, l'utilisateur doit d'abord encercler l'objet visuel, puis enregistrer le mot vocal. Pour l'interaction de recherche, l'utilisateur prononce d'abord le mot associé à l'objet visuel qu'il veut que le robot cherche. Puis, le système lui présente les N résultats les plus proches (ici $N = 3$). Enfin, l'utilisateur sélectionne ceux qui correspondent réellement à sa requête, permettant ainsi d'améliorer la reconnaissance vocale, et aussi de regrouper les exemples se rapportant à un même objet.

ler ». Il permet à l'utilisateur d'enregistrer le mot acoustique qu'il veut associer avec cet objet. Cette interaction est très classique, couramment utilisée dans la vie de tous les jours, très intuitive et très facile à utiliser (une illustration est visible sur la figure 9.4). Un système plus complexe pourrait être utilisé, permettant par exemple de déclencher/arrêter automatiquement l'enregistrement lorsque l'utilisateur commence/arrête de parler. Nous avons souhaité conserver ici l'interaction la plus simple possible afin d'éviter toute source potentielle de mauvais enregistrements.

FIGURE 9.4 – Après avoir fait prendre un objet en photo à un robot afin de collecter un exemple d'apprentissage, l'utilisateur peut enregistrer un mot acoustique nouveau qui sera associé à cet objet, à l'aide d'un bouton de type « pousser-pour-parler ».

9.2.3.2 Recherche active

Les utilisateurs peuvent aussi demander au robot de rechercher un objet déjà appris simplement en pressant un bouton et en prononçant le son précédemment associé. Nous n'avons pas implémenté d'algorithme complexe de recherche d'objet, ce problème ne faisant pas partie des objectifs de cette thèse. Le robot se contentait ici de tourner sur lui même jusqu'à apercevoir l'objet.

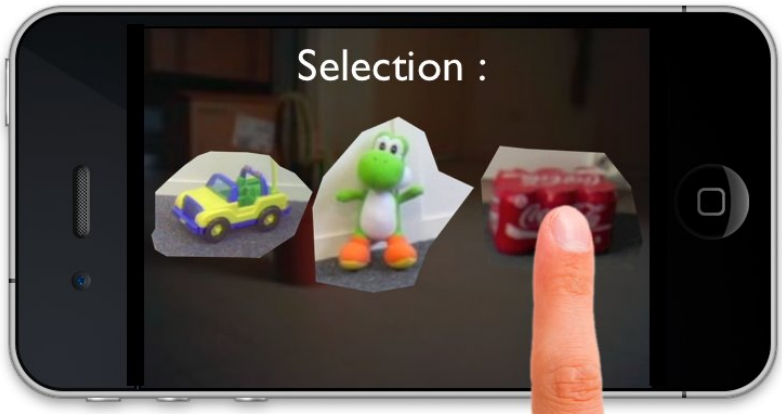


FIGURE 9.5 – Lorsque l'utilisateur prononce un mot vocal déjà appris pour demander au robot de rechercher l'objet visuel associé, un menu apparaît présentant les trois images associées aux mots les plus proches. Ce système permet d'améliorer facilement la reconnaissance vocale, mais surtout permet à l'utilisateur d'être sûr que le robot va bien aller chercher l'objet qu'il veut.

Même en utilisant des algorithmes de reconnaissance de la parole « *off-the-shelf* » très perfectionnés et/ou en utilisant un système de plus proche voisin, du type de celui décrit ci-dessus, sans le combiner à une interface bien conçue, il est impossible d'obtenir un taux de reconnaissance de 100%. Ces fausses détections sont problématiques car elles peuvent amener à des malentendus pour l'utilisateur. En effet, s'il prononce un mot pour demander au robot de chercher l'objet visuel associé, l'utilisateur ne va pas comprendre si le robot va chercher un autre objet. En effet, pour l'utilisateur ces problèmes de reconnaissance vocale sont triviaux. Il lui est donc difficilement compréhensible que des mots, qui lui semblent si différents à entendre, puissent être confondus par le robot. De plus, alors que ces mots ne sont associés qu'à des objets visuels pour le robot, ils peuvent correspondre à des concepts très différents pour l'humain, renforçant l'incompréhension de l'utilisateur. Ce phénomène risque de dégrader fortement l'expérience utilisateur.

Ce type de recherche, où l'utilisateur prononce un mot et le robot choisit directement l'objet qui lui semble correspondre le mieux peut, selon nous, être vu comme une utilisation systématique du bouton « J'ai de la chance » lors d'une recherche Google [1]. Nous avons développé ici une approche différente, où lorsque l'utilisateur

1. `http://www.google.fr/`

entre un mot vocal pour demander au robot de rechercher l'objet associé, le système lui affiche les représentations visuelles associées aux N mots les plus proches. Ces représentations visuelles sont simplement les images encerclées par l'utilisateur, comme on peut le voir sur l'exemple de la figure 9.5. Dans les expériences présentées ci-dessous, nous avons généralement présenté les trois exemples les plus proches aux utilisateurs (i.e. $N = 3$). Des expériences sur l'influence de ce paramètre sur les performances du système seront présentées dans la section 9.3.4.

L'utilisateur peut alors directement choisir, en tapant sur l'image voulue, l'objet qui correspond réellement à sa requête. En plus d'augmenter le taux de reconnaissance (il y a une très grande probabilité que le son correspondant soit parmi les N plus proches voisins), cela permet également à l'utilisateur de toujours décider quel objet le robot va aller chercher, et donc de toujours bien comprendre ce qui se passe. Si l'objet recherché n'est pas parmi les N plus proches voisins, l'utilisateur peut simplement cliquer sur le fond, c'est-à-dire n'importe où sauf sur une des images, pour annuler sa requête (des implémentations alternatives seront discutées dans la section 9.5).

9.2.3.3 Clusterisation incrémentale des exemples

En plus de permettre d'éviter les incompréhensions lors de la recherche, ce système permet également d'obtenir de manière quasi-transparente d'autres informations de l'utilisateur. En effet, si parmi les images présentées à l'utilisateur plusieurs correspondent au mot prononcé par l'utilisateur, il peut toutes les sélectionner. Ainsi, il indique au système que ces différents exemples se rapportent à un même objet, et qu'ils peuvent donc être regroupés. L'utilisateur peut ainsi participer de manière intuitive et non-contraignante à la clusterisation des exemples d'apprentissage. Un exemple de cette interaction est visible sur l'image 9.6. Cette information permet de lever ainsi l'ambigüité inhérente à la comparaison de mots vocaux.

Comme nous pouvons le voir sur le schéma de la figure 9.7, l'itération de ce processus sur plusieurs interactions va permettre au robot de se construire une clusterisation de plus en plus complète. Initialement, notre système d'apprentissage n'est constitué que d'associations entre un seul exemple d'objet visuel et de mot acoustique associé. Chaque fois que l'utilisateur clique sur deux (ou plus) images, cela permet de fusionner les deux (ou plus) groupes contenant ces images. Ainsi, l'utilisateur va pouvoir incrémentalement regrouper les exemples correspondants à un même objet et donc réduire le nombre de groupes existants.

Ce système peut également être utilisé lors de la collecte d'un nouvel exemple. En effet, après avoir fourni un nouveau couple (image / son) au système d'apprentissage, l'interface peut présenter les N représentations visuelles des mots les plus proches de celui prononcé par l'utilisateur, afin de lui demander de grouper ce nouvel exemple avec d'autres précédemment entrés. Dans les expériences décrites ci-dessous, nous avons utilisé cette possibilité supplémentaire.

Dans la prochaine section, nous allons évaluer à quel point, à travers ces interactions simples, transparentes et non-contraignantes, l'utilisateur peut grouper

FIGURE 9.6 – Lors de la recherche (ou l'ajout d'un nouvel exemple) l'utilisateur peut sélectionner plusieurs exemples - s'ils correspondent au même objet - afin d'indiquer au système d'apprentissage qu'ils peuvent être regroupés. L'utilisateur peut ainsi incrémentalement catégoriser les exemples d'apprentissage à travers une interaction intuitive et transparente.

incrémentalement l'ensemble des exemples d'apprentissage fournis. Nous étudierons aussi l'impact sur les performances de notre système de reconnaissance dans son ensemble.

9.3 Évaluation

Nous avons cherché à évaluer l'impact de la clusterisation incrémentale effectuée par l'utilisateur sur les performances de reconnaissance visuelle du système. Ici, nous n'allons ni évaluer l'utilisabilité perçue de l'interface, ni l'expérience utilisateur. Nous allons simplement simuler son utilisation, afin d'étudier les gains de performance qu'il est possible d'atteindre avec ces nouveaux éléments de l'interface dans le contexte de l'apprentissage conjoint d'objets visuels et de mots acoustiques. Il est important de noter que les résultats présentés ci-dessous ont été obtenus en modélisant les interactions de l'utilisateur et pourraient donc être légèrement différents avec de vrais utilisateurs.

9.3.1 Scénario expérimental

Comme expliqué ci-dessus, nous n'allons pas présenter d'études utilisateurs globales, comme celles décrites au chapitre 7, où des utilisateurs recrutés en dehors du laboratoire utiliseraient l'interface dans des conditions réalistes, mais simplement une simulation de l'utilisation de l'interface. Plus précisément, lors des scénarios d'interactions décrits ci-dessous, nous avons généré les informations qu'un utilisateur aurait fournies via l'interface, telles que la fusion de deux exemples, et les avons utilisé dans notre apprentissage.

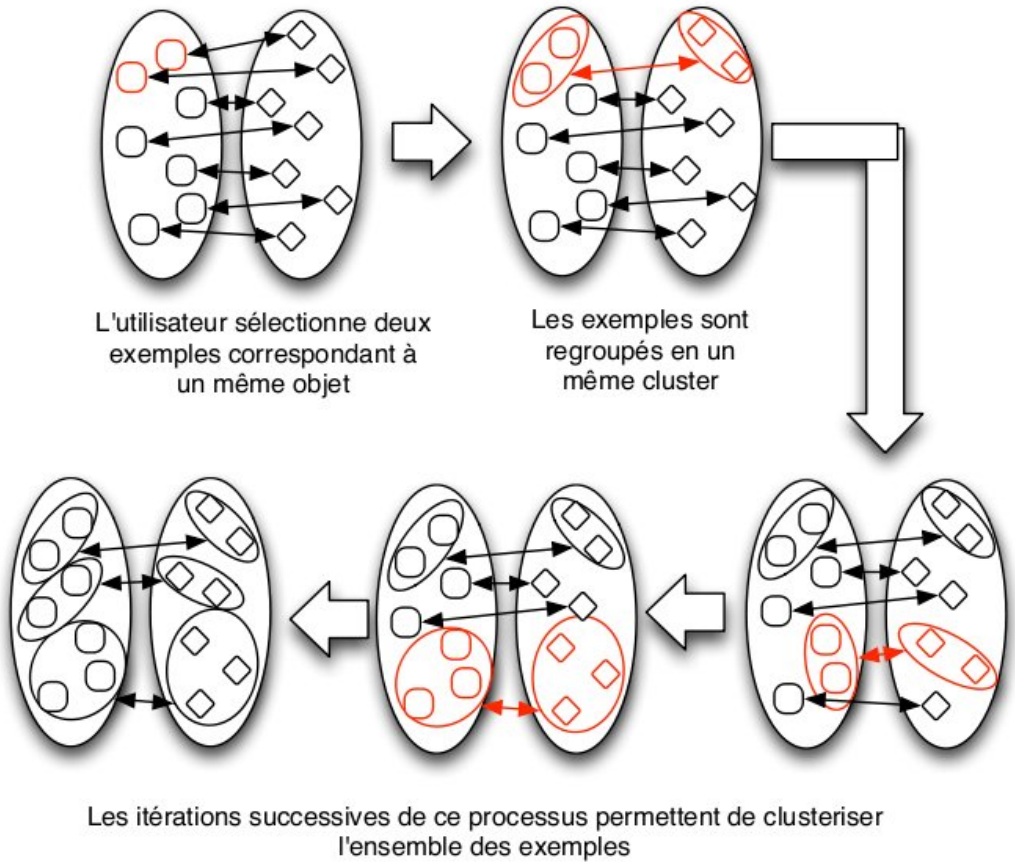


FIGURE 9.7 – Les itérations successives de l'interaction, où l'utilisateur peut sélectionner les images se rapportant à un même objet, vont permettre au robot de se construire une clusterisation de plus en plus complète.

Afin d'évaluer le rôle de l'interface dans l'apprentissage conjoint d'objets visuels et de mots acoustiques, nous avons effectué des tests sur une base de données d'exemples d'apprentissage (des détails sur sa création seront données en section 9.3.2) simulant une interaction de type élève / maître. Nous avons modélisé deux types d'interactions :

- **Interaction d'apprentissage** : Le maître fournit un nouvel exemple d'apprentissage à l'élève, c'est-à-dire qu'il montre au robot un objet visuel et y associe un mot acoustique.
- **Interaction de test** : Le maître évalue le savoir de son élève en prononçant un mot acoustique et en lui présentant un objet visuel. L'élève doit déterminer si le mot prononcé et l'objet visuel correspondent.

Les détails techniques de la réalisation effective de ces deux interactions dans nos expériences seront donnés dans la section 9.3.3.

Afin de reproduire un scénario d'interaction plausible entre un humain et un robot, nous avons choisi d'alterner ces deux interactions. Dans un premier temps, le maître favorisait les interactions d'apprentissage, les mélangeant avec quelques interactions de test. Dans un second temps, le maître se concentrait sur les interactions de test, tout en conservant quelques interactions d'apprentissage. Notre scénario exact d'interaction était le suivant :

Phase 1 : 80% d'apprentissage et 20% de test

Phase 2 : 20% d'apprentissage et 80% de test (le passage de la phase 1 à la phase 2 était effectué après 50 interactions)

Il est important de noter que, comme indiqué précédemment, lors des expériences, le déroulement de ce scénario était simulé. Par exemple, lors d'une interaction d'enseignement, un couple (représentation visuelle, mot vocal) était tiré aléatoirement parmi les exemples disponibles et fourni à notre système d'apprentissage (de même pour les interactions de tests). De plus, les utilisateurs « simulés » avaient un comportement optimal, c'est-à-dire qu'ils ne se trompaient jamais dans l'utilisation de l'interface. Les informations de clusterisation fournies étaient toujours exactes et ils fournissaient toujours le maximum d'informations possibles au système d'apprentissage (sélection de l'ensemble des images correspondant à un même objet).

9.3.2 Base de données d'exemples

Les exemples d'apprentissage (image et son) utilisées lors de nos tests proviennent des expériences utilisateurs effectuées à Cap Sciences, décrites au chapitre 7. En effet, lors de ces expériences, en plus de demander aux participants de collecter des exemples d'apprentissage d'objets visuels avec les différentes interfaces, nous leur demandions aussi d'enregistrer un mot pré-défini associé à ces objets. 12 objets différents et donc 12 catégories différentes de mots vocaux ont été utilisés pour constituer cette base de données. Afin de réduire le bruit ambiant, les utilisateurs étaient équipés d'un micro-casque.

Parmi tous les exemples d'apprentissage collectés, seuls les exemples visuels collectés par l'interface iPhone ont été conservés (ce sont les exemples donnant les

meilleurs résultats). Les exemples dits « gold » ont également été conservés comme base de tests. Les mots vocaux enregistrés par les utilisateurs ont été vérifiés et les exemples où les utilisateurs n'avaient pas enregistré le mot indiqué ont été supprimés de la base de données (e.g. « fifa » au lieu de « jeu vidéo »). La base de données d'exemples utilisée dans les expériences décrites ci-dessous est constituée de 282 exemples visuels (162 pour l'apprentissage et 120 pour les tests) et de 245 mots acoustiques. Les exemples étaient répartis de manière approximativement homogène entre les 12 catégories.

9.3.3 Protocole expérimental

Comme nous l'avons expliqué dans les sections précédentes, nous avons simulé une suite d'interactions entre un humain et un robot. Dans un premier temps, l'humain se focalisait sur les interactions d'apprentissage puis sur celles de test. Plus précisément, nous suivions le protocole expérimental suivant :

1. Un tirage aléatoire était effectué pour déterminer si nous allions procéder à une interaction d'apprentissage ou à une interaction de test (les probabilités étaient de 80–20 en faveur de l'apprentissage pour la phase 1, puis étaient inversées pour la phase 2).
2. Nous simulions l'interaction choisie :
   - Pour une interaction d'apprentissage, un exemple d'objet visuel et un mot acoustique, correspondant à une même catégorie, étaient tirés aléatoirement parmi les exemples encore disponibles de la base de données d'exemples. Ce couple était fourni à notre système d'apprentissage. Nous simulions également l'interaction de regroupement des exemples par l'utilisateur via l'interface, c'est-à-dire que nous trouvions les N exemples les plus proches et fusionnions ceux correspondant bien au mot acoustique.
   - Pour une interaction de test, un mot acoustique était tiré aléatoirement parmi les exemples disponibles de la base de données. Nous simulions l'interaction de recherche active de ce mot, ainsi que le regroupement des exemples comme pour l'interaction d'apprentissage.

     Les exemples visuels de la base de données de test étaient ensuite présentés, un par un, au système d'apprentissage. Pour chacun, le système devait indiquer si, oui ou non, le mot acoustique correspondait à l'exemple visuel présenté. Plus concrètement, nous cherchions d'une part, le groupe d'exemples correspondant au mot acoustique, d'autre part nous réalisions un vote, comme décrit en section 6.2.4, pour déterminer à quel groupe l'objet visuel présenté correspondait. Le système d'apprentissage indiquait que l'objet visuel correspondait au mot acoustique si et seulement si les deux groupes étaient les mêmes.

     Il est important de noter que ce test est plus sélectif qu'un test d'évaluation de recherche d'objet. En effet, ici nous demandions au système d'apprentissage de déterminer, si un mot acoustique correspondait à un exemple

d'objet visuel particulier, et non pas de trouver n'importe quel exemple visuel correspondant. Les résultats de ce type de tests seraient intéressants, cependant nous n'avons pas eu le temps de les réaliser.

Ces étapes étaient répétées jusqu'à ce qu'il n'y ait plus d'exemples disponibles dans la base de données. Le changement de phase intervenait après 50 interactions. Le scénario était répété 100 fois et les résultats présentés dans la section suivante représentent la moyenne des résultats obtenus.

9.3.4 Résultats

9.3.4.1 Aide à la reconnaissance vocale

Dans un premier temps, nous avons évalué l'efficacité de notre interface pour l'aide à la reconnaissance vocale. Nous considérons ici que la reconnaissance a été réussie, si parmi les K exemples présentés à l'utilisateur via l'interface, au moins un correspond réellement au mot prononcé par l'utilisateur. Nous avons comparé les résultats obtenus en fonction du nombre d'exemples présentés à l'utilisateur sur l'écran de l'appareil.

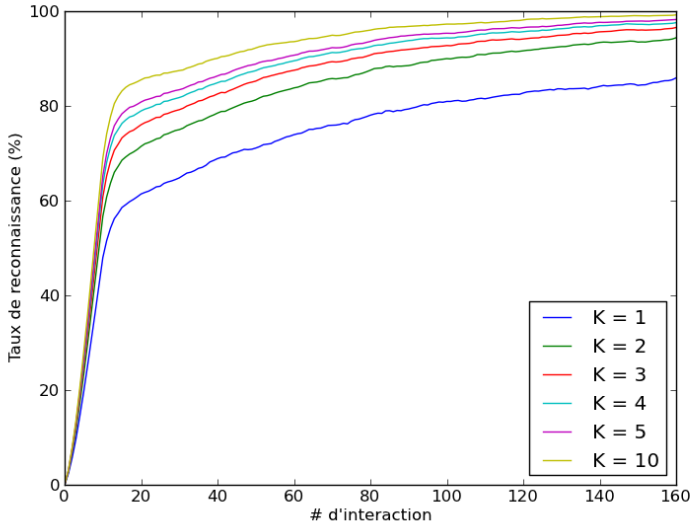


FIGURE 9.8 – L'interface de catégorisation permet, même avec un nombre d'exemples réduit présentés à l'utilisateur (e.g. $K = 3$) d'améliorer d'approximativement 20% le taux de reconnaissance vocale.

Comme nous pouvons le voir sur la figure 9.8, pour la condition $K = 1$, c'est-à-dire lorsqu'un seul exemple est présenté à l'utilisateur, et donc sans utiliser notre système de recherche active, nous obtenons environ 80% de reconnaissance. Notre base de données n'étant composée que d'un nombre limité de catégories de mots (12), ce score est déjà assez élevé. Cependant, comme expliqué plus haut, les fausses

reconnaissances peuvent mener à des incompréhensions et dégrader rapidement l'expérience utilisateur. Nous pouvons également noter qu'avec l'interface, dès la condition $K = 3$, nous obtenons quasiment 100% de reconnaissance, ce qui permet aussi de réduire le nombre d'interactions nécessaires, pour obtenir un taux de reconnaissance équivalent. Bien qu'il ne soit pas surprenant qu'en prenant en compte les K résultats les plus proches au lieu d'un seul nous améliorons les performances, il est intéressant de noter qu'en présentant seulement trois exemples à l'utilisateur, nous pouvons obtenir quasiment 100% de reconnaissance vocale.

9.3.4.2 Limitation du nombre de clusters

Nous pouvons voir, sur la figure 9.9, que l'utilisation de notre interface, et en particulier des interactions de clusterisation incrémentale des exemples, permet de limiter très significativement le nombre de clusters créés.

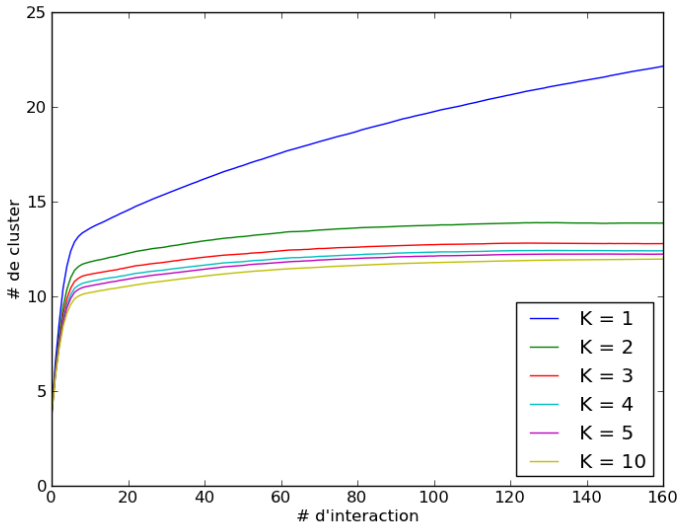


FIGURE 9.9 – À travers les interactions successives via notre interface, les utilisateurs peuvent limiter très significativement le nombre de clusters créés. Nous pouvons notamment constater que même avec $K = 3$ le nombre de clusters est quasi-constant.

Nous avons également essayé d'utiliser un algorithme de clusterisation standard (K-means [Kanungo 2002]) afin d'étudier les performances atteignables avec une clusterisation automatique des différents exemples. Nous avons utilisé une clusterisation naïve des exemples basée sur la distance DTW entre les sons et en fixant le nombre de clusters à 12 (le nombre de catégories différentes dans notre base d'exemples). Nous avantageons ici le système automatique en fixant le nombre de cluster à 12. En effet, autrement, il ne le connait pas et devrait le trouver automatiquement. Les résultats obtenus par cette méthode pourraient donc être encore

moins bons. Cette clusterisation automatique ne permet d'obtenir qu'environ 65% de bonne clusterisation, c'est-à-dire que près d'un tiers des exemples sont groupés avec des mots appartenant à une catégorie différente.

9.3.4.3 Impact sur la reconnaissance visuelle

La figure 9.10 présente le taux de succès obtenu lors de la dernière interaction de test du scénario complet. En plus des résultats obtenus en simulant l'utilisation de l'interface, nous présentons également les résultats obtenus avec différentes stratégies de clusterisation utilisées :

- **Aucune :** Les différents exemples ne sont jamais regroupés. Chacun des groupes est donc un singleton.
- **K-means :** Les exemples sont regroupés à l'aide de l'algorithme K-means, en utilisant comme mesure de similarité la distance de déformation temporelle dynamique entre les séquences de coefficients MFCCs des mots vocaux. Le nombre de clusters était fixé à 12.
- **Gold :** Les exemples sont regroupés entre eux en utilisant les étiquettes des mots acoustiques. Cette stratégie correspond donc à la clusterisation conjointe des exemples d'apprentissage idéale.

Comme nous pouvons le constater, sans *aucune* clusterisation des différents exemples d'apprentissage, le taux de succès obtenu est très faible (moins de 20%). Bien que la clusterisation automatique (via *K-means*) améliore ce résultat (environ 30% de reconnaissance), le taux de succès obtenu reste encore très loin du score obtenu avec la clusterisation *Gold*, qui permet d'obtenir aux environs de 60% de succès. Nous pouvons constater que la clusterisation incrémentale, construite par l'utilisateur via l'interface *iPhone*, permet d'obtenir des performances quasiment similaire à la condition *Gold*, et ce même en présentant seulement trois exemples aux utilisateurs.

De même que pour les résultats précédents, il est particulièrement important de noter ici que les performances absolues obtenues pourraient être améliorées en utilisant d'autres algorithmes de reconnaissance vocale et/ou visuelle, mais que l'ordre de grandeur typique du gain observé dans la littérature est plus faible que les différences observées ici. De plus, ce qui nous intéresse ici est de montrer la possibilité de faire grouper incrémentalement par l'utilisateur les différents exemples d'apprentissage de manière quasi-optimale, et ce à travers des interactions très simples, peu nombreuses et transparentes via une interface bien conçue. Il peut ainsi améliorer très significativement les performances du système de reconnaissance.

9.4 Conclusion

Dans ce chapitre, nous avons présenté un système complet permettant à un utilisateur non-expert de montrer des objets visuels à son robot, et d'associer à ces représentations visuelles des mots acoustiques eux aussi nouveaux. Par la suite, l'utilisateur pourra demander à son robot de rechercher ces objets. La principale

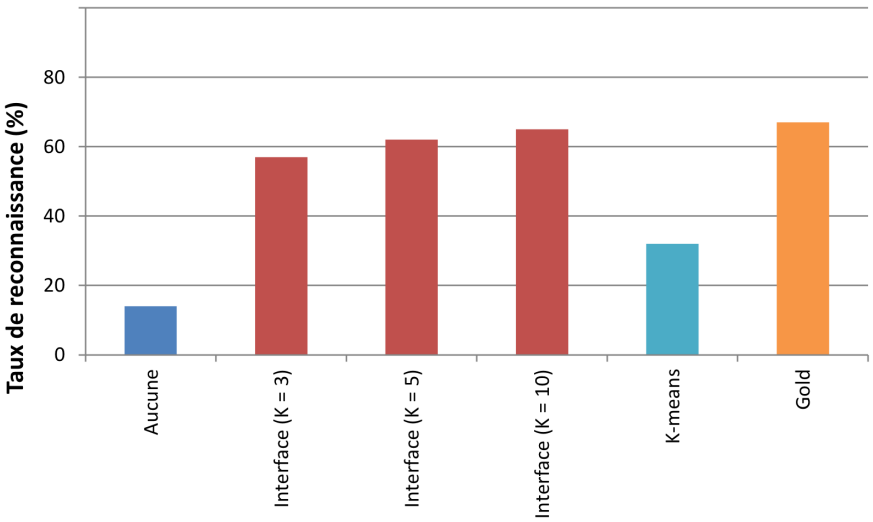


FIGURE 9.10 – La clusterisation incrémentale, construite par les utilisateurs via notre interface, permet d'obtenir, même en présentant seulement trois exemples, des performances quasiment similaires à celles obtenues avec la clusterisation optimale.

contribution de ce système réside dans la conception d'une interface dédiée permettant, d'une part, à l'utilisateur de regrouper les différents exemples d'apprentissage correspondant à un même objet à travers des interactions intuitives, transparentes et peu nombreuses, et d'autre part, de demander au robot de chercher un objet et de participer au processus de reconnaissance vocale, le rendant ainsi plus robuste.

Nous avons évalué l'impact potentiel de cette interface à travers une expérience simulant son utilisation sur une base de données d'exemples construite auparavant. Les résultats ont montré que même avec très peu d'interactions de l'utilisateur, l'interface permet de très rapidement regrouper l'ensemble des exemples correspondants à un même objet. Ce problème de groupement conjoint reste un problème difficile d'apprentissage. L'utilisateur peut ainsi, à travers l'utilisation d'une interface spécialement conçue, améliorer très significativement les performances de reconnaissance du système.

9.5 Limites et travaux futurs

Bien qu'à notre avis très prometteuse, cette étude présente encore un nombre important de limites. En particulier, il serait absolument indispensable de faire tester notre interface à des participants non-experts, dans des conditions réalistes hors-laboratoire, du type de celles décrites au chapitre 7. En effet, bien qu'il nous semble que l'interaction soit intuitive, transparente et non-contraignante, seule une étude utilisateurs avec des participants non-familiers de l'interface nous permettrait de l'affirmer. De même, il serait intéressant d'étudier l'impact du nombre d'exemples

présentés aux participants sur leur charge de travail et sur l'expérience utilisateur. Il serait, par exemple, peut-être préférable de lui présenter peu d'exemples à la fois mais de lui donner la possibilité de les faire défiler à la manière des pages de résultats d'une recherche Google.

Il serait également intéressant de tester l'impact de notre interface en utilisant d'autres algorithmes d'apprentissage multi-modaux plus perfectionnés et plus robustes. Nous avons par exemple commencé à étudier l'approche NMF (*Non-negative Matrix Factorization*). Cette technique a déjà été utilisée dans le contexte de l'acquisition du langage [Driesen 2009], notamment dans le cadre du projet européen ACORNS [2] (*Acquisition of Communication and Recognition Skills*). Les informations, fournies par l'utilisateur via l'interface, pourraient alors être encodées dans des dimensions supplémentaires. Nous n'avons cependant pas à l'heure actuelle testé cette approche.

Enfin, il serait également intéressant d'étudier la possibilité de combiner notre interface avec une approche automatique, afin de réduire le nombre d'interventions de l'utilisateur, en ne le sollicitant que lorsque le système a un doute quant à l'appartenance de deux exemples à un même objet (voir la discussion de la section 10.3 pour plus de détails).

2. `http://www.acorns-project.org/`

Chapitre 10

Discussion

Sommaire

Résumé du chapitre

Dans ce chapitre, nous allons commencer par présenter une synthèse des contributions principales de cette thèse. Nous allons ensuite discuter des limites et extensions possible de notre système pour l'enseignement d'objets visuels nouveaux associés à des mots nouveaux à un robot. Nous discuterons aussi de la transposition de notre approche, focalisée sur l'étude et le développement d'interface pour l'apprentissage social de mots et d'objets visuels nouveaux chez les robots, à d'autres domaines d'applications tels que la recherche et la préhension d'objets visuels, la télé-opération ou encore l'apprentissage social des mouvements chez les robots. Enfin, nous présenterons une ouverture possible de notre travail sur l'évaluation des interfaces à la création de jeu robotique.

10.1 Contributions principales

Dans cette thèse et plus particulièrement dans les chapitres 5 à 9, nous avons étudié le rôle de l'interface dans l'enseignement d'objets visuels nouveaux associés à des mots nouveaux à un robot.

Nous avons notamment montré comment des interfaces basées sur des objets médiateurs peuvent aider un utilisateur non-expert, si elles sont bien conçues, à attirer l'attention d'un robot, à désigner des objets visuels particuliers et, également à faciliter les situations d'attention partagée avec un robot. Ces interfaces permettent aux utilisateurs de faire collecter à un robot des exemples d'apprentissage de très bonne qualité, c'est-à-dire d'une qualité comparable aux exemples collectés par des utilisateurs experts dans des conditions idéales. Nous avons aussi montré que la collecte d'exemples de bonne qualité permettait d'améliorer de manière très significative les performances générales de notre système de reconnaissance, d'un ordre de magnitude important en comparaison des améliorations obtenues lors du changement d'algorithmes de perception et/ou de méthodes d'inférences statistiques [Mikolajczyk 2003][Bay 2008][Rouanet 2009c][Rouanet 2010a][Rouanet 2010b].
Nous avons aussi proposé une approche de collecte semi-automatique d'exemples d'apprentissage permettant de réduire le nombre d'interventions de l'utilisateur.

Nous avons également étudié comment l'interface pouvait permettre d'améliorer les performances du système de reconnaissance vocale. Plus précisément, lors de la recherche d'un objet visuel, une liste de résultats les plus proches était présentée à l'utilisateur. En plus, d'améliorer les performances du système de reconnaissance vocale, cette interaction permet surtout de rendre le processus de recherche plus transparent et permet donc une meilleure compréhension mutuelle.

Enfin, nous avons montré qu'une interface spécialement conçue pouvait permettre à des utilisateurs de catégoriser incrémentalement, efficacement et conjointement les différents exemples d'apprentissage audio-visuels fournis au système. Nous avons plus particulièrement souligné, via une simulation de l'utilisation de cette interface, qu'à travers des interactions simples et peu nombreuses, l'utilisateur pouvait regrouper de manière quasi-optimale l'ensemble des exemples d'apprentissage présentés à un robot lors d'un scénario d'interaction plausible. Ainsi, nous pouvons permettre l'apprentissage conjoint des catégories visuelles et auditives et leur association aussi bien que lorsqu'on suppose l'utilisation de symboles dans une des modalités.

Nous nous sommes aussi intéressés à l'utilisabilité perçue de nos interfaces par des utilisateurs non-experts ainsi qu'à l'expérience utilisateur lors de l'interaction. Nous avons ainsi pu évalué que nos interfaces basées sur des objets médiateurs étaient jugées intuitives, faciles à utiliser mais aussi non-contraignantes voir plaisantes. En particulier, elles ont été jugées plus intuitives qu'une interface basée sur des gestes des mains et des bras de l'utilisateur, et donc a priori plus naturelle. Cependant, à cause des différences d'appareil sensori-moteurs entre l'humain et son robot, les interactions directes souffrent d'un manque de robustesse, ce qui entraine une interaction en milieu ouvert non-satisfaisante pour les utilisateurs. L'évaluation de l'interface gestuelle, où la reconnaissance de gestes était réalisée à travers un protocole de type magicien d'Oz, nous a permis de souligner que même avec des algorithmes de reconnaissance et d'interprétation aussi performants qu'un humain, ce type d'interaction souffre toujours de problèmes d'utilisabilité.

D'autre part, nous avons montré, particulièrement dans le chapitre 7, que la

conception d'études utilisateurs sous forme de jeu robotique pouvait être une solution à l'évaluation d'interface, lors d'une interaction humain-robot avec des utilisateurs non-experts. Plus précisément, nous avons présenté les spécificités de notre protocole expérimental (tel que le scénario, l'interface de jeu ou les tutoriels) intégrées sous forme de jeu, qui permettaient de justifier une tâche abstraite pour les participants, ainsi que de les maintenir motivés et concentrés durant toute l'expérience [Rouanet 2010a].

Après avoir présenté les principales contributions de ce travail, nous allons maintenant discuter de ses limites et des extensions possibles de notre étude du rôle de l'interface dans l'enseignement d'objets visuels nouveaux associés à des mots nouveaux à un robot. Nous allons notamment présenter un scénario possible de développement d'un système complet et intégré, ainsi que son évaluation. Nous présenterons aussi des extensions possibles, telles que l'utilisation d'algorithmes plus perfectionnés de catégorisation multi-modaux (son et image), automatique ou semi-automatique, ou encore la transposition de notre démarche à d'autres contextes applicatifs tels que l'apprentissage social de mouvements à un robot.

10.2 Développement d'une solution complète et intégrée

10.2.1 Limites des simulations de l'interface pour la reconnaissance vocale et la clusterisation d'exemples d'apprentissage

Dans les chapitres précédents, nous avons présenté un système mêlant perception visuelle, apprentissage machine et surtout différentes interfaces basées sur des objets médiateurs. L'interface iPhone, celle permettant aux utilisateurs de collecter les meilleurs exemples, jugée la plus satisfaisante par les utilisateurs, mais aussi celle offrant le plus de possibilités d'interaction grâce à son écran tactile, a été étendue à l'utilisation de mots vocaux. Bien que des prototypes de cette interface aient été présentés, cette interface n'a pas à l'heure actuelle été implémentée entièrement ni intégrée avec le reste du système (perception/reconnaissance vocale et visuelle). Les résultats présentés au chapitre 9 ont été obtenus en simulant l'utilisation de cette interface, mais pas par son utilisation réelle par des utilisateurs. Ce procédé nous a permis d'explorer le potentiel de clusterisation de cette interface et en particulier nous avons pu montrer que lors d'un scénario d'interaction entre un humain et son robot, l'interface présentée pouvait permettre à des utilisateurs de regrouper incrémentalement l'ensemble des exemples d'apprentissage de manière quasi-optimale. Bien que seule une utilisation réelle de cette interface permettrait de valider son efficacité, cette simulation nous a permis de présenter une preuve de concept et nous a aussi permis de comparer facilement et rapidement différents paramètres de l'interface. Nous avons notamment pu évaluer l'impact du nombre d'exemples présentés à l'utilisateur sur les performances du système de reconnaissance vocale et sur la catégorisation.

Bien qu'apportant de précieuses indications quant aux possibilités de notre interface pour l'aide à la reconnaissance vocale et à la catégorisation, ces simulations

ne permettent ni de l'évaluer de manière complète, ni d'évaluer la manière dont les utilisateurs se serviront réellement de l'interface. Il serait nécessaire d'évaluer si l'interface est effectivement utilisable par des humains non-experts, et en particulier :

- s'ils parviennent à utiliser l'interface,
- s'ils font des erreurs lors de son utilisation (regrouper des exemples ne correspondant pas au même objet) et
- s'ils fournissent le maximum d'informations de clusterisation (l'utilisateur pourrait, par exemple, ne sélectionner qu'un sous ensemble des exemples à grouper s'il trouve l'interaction trop contraignante).

Il faudrait également évaluer l'utilisabilité perçue par les humains non-experts de notre interface lors de ces interactions supplémentaires (reconnaissance vocale et clusterisation). Ces critères sont en effet primordiaux ici. Les performances obtenues, spécialement pour la clusterisation, dépendront directement de l'implication et de la motivation des utilisateurs à sélectionner l'ensemble des exemples se rapportant à un même objet qui lui sont présentés. Il faudrait plus particulièrement étudier les points suivants :

- la compréhension de l'interface (aide à la reconnaissance vocale et à la clusterisation)
- la facilité d'utilisation (aide à la reconnaissance vocale et à la clusterisation)
- la charge cognitive de travail (suivant le nombre d'exemples présentés)
- la pénibilité de l'interaction pour la clusterisation des exemples (à court mais aussi plus long terme)
- la motivation des utilisateurs : il serait intéressant d'étudier si les participants sont enclins à fournir naturellement ces informations ou s'il serait indispensable de justifier ces interactions, par exemple en leur expliquant leur utilité.

10.2.2 Difficultés de l'évaluation de l'utilisabilité réelle et perçue

L'évaluation valide de ces différents critères ne pourrait se faire qu'à travers une étude utilisateur à grande échelle du type de celle présentée au chapitre 7. La conception d'un jeu robotique nous avait permis de justifier la collecte d'exemple d'objets visuels nouveaux ainsi que de mots vocaux nouveaux. Il semble clair que la justification de l'interaction lors de la reconnaissance vocale et de la clusterisation passe nécessairement par la recherche effective des objets visuels par le robot (ou plus simplement en pointant ces objets). Cependant, cette possibilité n'est pas actuellement fonctionnelle dans notre système. De plus, pour pouvoir reconnaître et suivre un objet visuel, il serait nécessaire d'avoir plusieurs exemples d'apprentissage par objets visuels, ce qui nécessiterait des expériences plus longues. Ce problème pourrait être contourné en utilisant les exemples collectés par les autres participants, ou en simulant la recherche grâce à un protocole de type magicien d'Oz.

En plus, des difficultés mentionnées ci-dessus, ce type d'étude présente aussi un ensemble de limites propres à notre contexte d'utilisation : la robotique personnelle et sociale. Nous avons déjà présenté certaines de ces limites dans les chapitres 4 et 7. Nous allons maintenant en discuter plus en détails.

Pour pouvoir réaliser une étude utilisateur valide d'une interface, il faut pouvoir apporter une réponse aux questions suivantes [Nielsen 1994] :

- quel utilisateur ?
- pour quel usage ?
- dans quel contexte ?

Or, comme nous l'avons expliqué au chapitre 4, l'enseignement d'objets visuels nouveaux associés à des mots nouveaux à un robot personnel n'est pas à l'heure actuelle une tâche existante hors des laboratoires de recherche. Cette activité deviendra probablement cruciale dans les années à venir (voir chapitre 1 pour plus de détails), mais il n'existe pas encore d'utilisateurs ni de contexte d'utilisation réaliste. Nous ne pouvons donc pas réellement garantir la validité écologique de nos études (faire correspondre les conditions d'évaluation aux conditions d'utilisations réelles) mais nous devons essayer d'imaginer, le plus objectivement possible, ce à quoi pourrait ressembler l'utilisation future de nos interfaces.

Tout d'abord, en recrutant des participants dans un lieu grand public comme le musée Cap Sciences, nous avons pu avoir des utilisateurs présentant des profils très variés et donc couvrir un vaste panel de types d'utilisateurs possibles. Ensuite, nous avons essayé de recréer un scénario d'interaction plausible, basé sur la conception d'un jeu robotique dont l'objectif était de justifier l'usage de nos interfaces. Cependant, la conception de cette étude présentait un certains nombre de limites. Tout d'abord, nous ne cherchions à justifier l'utilisation de nos interfaces que pour une interaction très limitée dans le temps. Réaliser des expériences utilisateur de plus de 30 minutes soulève en effet de très nombreux problèmes. D'autre part, dans l'étude présentée au chapitre 7, les participants n'étaient pas nécessairement au courant de l'utilité potentielle de notre système pour le développement de la robotique personnelle. Pour eux, nos interfaces étaient conçues pour le jeu et pas nécessairement pour l'enseignement conjoint de mots nouveaux et d'objets visuels. Ce problème introduit donc un biais clair de notre évaluation mais qui ne semble pas possible, ou tout du moins très difficile, à éviter actuellement.

Il est toutefois intéressant de noter que, même lorsque que la robotique personnelle deviendra quotidienne, l'enseignement de mots nouveaux à un robot pourrait se faire à travers des interactions sous forme de jeu, telles que celles proposées lors de nos expériences. La collecte des exemples d'apprentissage pourrait toujours être considéré par les utilisateurs comme un jeu n'ayant pas nécessairement d'autres objectifs que l'amusement. Les parents jouant avec leur nourrisson et lui montrant des objets visuels présents autour d'eux, ne pensent probablement pas seulement à permettre à leur bébé de se développer mais surtout à interagir et jouer avec lui. Notre contexte de jeu robotique pourrait donc au final, s'avérer un contexte d'utilisation plausible. Les jeux robotiques n'ont jusqu'à présent fait l'objet que de très peu d'études et il serait sans doute intéressant de les étudier plus en détails (voir section 10.6).

Cependant, une différence majeure subsiste. Nous n'avons évalué nos interfaces que sur des périodes d'utilisation très courtes (moins d'une heure dans tous les cas). Il est pourtant envisageable que les utilisateurs interagissent avec leur robot plusieurs

fois par jour pendant des années. Il serait donc très important de pouvoir réaliser des études utilisateurs à plus long terme. Si notre jeu a été jugé amusant pendant une demi-heure, il est probable que les utilisateurs se lasseraient très rapidement au delà de cette durée. Cette limite de l'évaluation de notre système est partagée par la plupart des travaux réalisés dans le domaine de l'interaction homme-robot, mais avec l'arrivée de la robotique personnelle il devient de plus en plus nécessaire de lever cette barrière.

10.2.3 Extension à la recherche d'objets visuels

Les interfaces présentées dans cette thèse permettent à un utilisateur d'enseigner des objets visuels nouveaux associés à des mots nouveaux à un robot, d'aider à la reconnaissance vocale et de regrouper les différents exemples d'apprentissage. Cependant, dans les expériences décrites précédemment la reconnaissance de ces objets visuels était effectuée hors ligne, c'est-à-dire après les expériences et en utilisant des bases de données de tests pré-enregistrées. Ce protocole de test nous a permis d'évaluer et de comparer les exemples d'apprentissage visuels collectés avec les différentes interfaces dans des conditions similaires. De plus, réaliser des tests hors-ligne permet d'effectuer de nombreux tests et donc de tester facilement une grande variété de paramètres. Cette méthode d'évaluation présentait cependant un ensemble de limites :

- La tâche de reconnaissance utilisée consistait à identifier parmi un ensemble d'images celles qui correspondaient à un objet visuel donné, c'est-à-dire simplement une tâche de classification. De plus, notre système ne s'autorisait pas à ne reconnaître aucun des objets dans une image. Il faudrait donc introduire une mesure de confiance de classification. Nous ne cherchions pas non plus à localiser l'objet visuel dans l'image.
- Nous n'avons pas non plus examiné le coût en calcul de notre système de reconnaissance visuelle et donc à son utilisation possible en temps réel. Il existe cependant de nombreuses implémentations des techniques de sacs de mots visuels temps réel (e.g. pour la navigation [Filliat 2008][Angeli 2008]).

Il serait donc intéressant de coupler notre système de classification visuelle à des algorithmes de recherche/localisation des objets. Bien que la mise en place de ces algorithmes dépasse le cadre de cette thèse, nous discuterons dans la section 10.5.1 du rôle qu'une interface du type de celles présentées ici pourrait jouer dans la détection et la recherche d'objets visuels.

10.3 Une extension possible : la catégorisation semi-automatique

Dans le chapitre 9, nous avons montré comment l'interface pouvait permettre aux utilisateurs de regrouper incrémentalement l'ensemble des exemples de manière quasi-optimale. Nous avons aussi montré qu'une méthode automatique de catégori-

sation non-supervisée simple (telle que l'utilisation de l'algorithme K-Means sur les sons en utilisant la distance DTW) ne permettait pas d'obtenir des résultats comparables. Notre démarche visait à montrer qu'une interface, bien conçue, permettait à l'utilisateur de résoudre facilement ce problème de clusterisation et de catégorisation, qui reste complexe à résoudre avec les méthodes de catégorisation automatique standard.

Il serait cependant intéressant d'étudier les différentes techniques de catégorisations automatiques multi-modales (nous n'avons pour l'instant utilisé que les informations auditives pour la catégorisation automatique) afin de voir les performances qu'elles permettent d'atteindre. Parmi les différentes techniques possibles, l'approche NMF (*non negative matrix factorization*) a déjà été utilisée pour l'association automatique d'images et de sons, notamment dans le cadre du projet européen ACORNS [Driesen 2009].

Les expériences présentées dans le chapitre 9 opposent notre technique « manuelle », où l'utilisateur regroupe lui-même les exemples, à une approche automatique, où l'utilisateur n'intervient pas du tout. Il semble pourtant intéressant de combiner ces deux approches. En effet, elles ont chacune des avantages et inconvénients complémentaires :

- L'approche basée sur l'interface permet de lever toutes les ambiguïtés inhérentes à la reconnaissance visuelle et vocale. Par contre, même si nous avons essayé de proposer une interaction aussi simple et transparente que possible, l'utilisation pourrait devenir contraignante pour les humains lors des interactions à long terme, plus particulièrement si elle est systématique.
- Les approches automatiques, quant à elles, peuvent permettre de réduire la charge de travail demandée à l'utilisateur mais elles peuvent amener à des erreurs de catégorisation pouvant dégrader les performances du système de reconnaissance vocale et/ou visuelle.

La combinaison de ces deux approches pourraient donc potentiellement limiter les interactions utilisateurs aux cas les plus ambigus. Une telle approche nécessiterait cependant la mise en place d'une mesure de confiance robuste permettant de déterminer quand regrouper automatiquement plusieurs exemples et quand demander de l'aide à l'utilisateur. Il serait intéressant d'étudier la possibilité de trouver un compromis entre la réduction des interventions de l'utilisateur et le maintien des performances de reconnaissance.

10.4 Limites technologiques : utilisation du Nao

Lors de nos tests en laboratoires et surtout lors de nos expériences utilisateurs à Cap Sciences nous n'avons utilisé que le robot Nao (le robot Aibo fut utilisé mais seulement lors des expériences sur la navigation seule). Le choix de ce robot correspond à notre vision de la robotique personnelle du futur et donc du domaine d'applications dans lequel nous nous situons. En effet, comme expliqué dans la section 4.2.2, ce robot présente un ensemble de caractéristiques importantes pour l'utilisa-

tion grand public. Tout d'abord, son aspect de jouet et sa forme humanoïde facilitent les interactions sociales avec les humains non-experts. Son utilisation ne nécessite pas l'utilisation de périphériques externes tels qu'une caméra fixée au plafond ou des détecteurs de présence. Ce robot pourrait donc facilement être mis en place dans un environnement du quotidien, sans contrainte particulière. Enfin, son coût, bien qu'à l'heure actuelle encore important, peut permettre un accès par le grand public. Ce coût limité entraine naturellement l'utilisation de capteurs et d'effecteurs limités. Ce robot n'est, par exemple, pas équipé d'une caméra 3D omnidirectionnelle ni de télémètre laser rendant plus difficile le partage d'attention entre l'humain et le robot. Sa petite taille ($\sim 60cm$) ainsi que sa caméra avec un angle d'ouverture limité rendent aussi sa perception de l'environnement très différente de celle d'un humain.

Les spécificités de Nao et ses limites technologiques ont nécessairement eu un impact sur nos interfaces. Utiliser un robot avec des capacités différentes nous aurait permis d'explorer d'autres possibilités. Par exemple, l'utilisation d'un robot équipé d'un micro-projecteur ouvre des perspectives intéressantes. Ishii et al. ont, par exemple, utilisé un projecteur permettant d'afficher directement dans le monde réel des informations à l'utilisateur telles que le tracé d'un pointeur laser ou encore des zones sélectionnées [Ishii 2009]. Un projecteur monté sur le robot pourrait nous permettre de présenter à l'utilisateur le champ de vision du robot sous forme de halo lumineux et aurait pu rendre le pointage avec l'interface basée sur le pointeur laser plus facilement compréhensible par les utilisateurs, et donc potentiellement plus performante.

Il est donc important de se demander si nos interfaces pourront facilement être transposés à d'autres robots ayant des capacités différentes. Nous pensons que, indépendamment des contraintes imposées par le robot, l'étude des questions d'attention partagée par exemple, centrale dans notre démarche, bien qu'influencée par l'appareil sensori-moteur du robot utilisé, resteront cruciales, même avec un robot doté de capacités visuelles plus performantes. L'utilisation d'une caméra omnidirectionnelle permettrait sûrement de faciliter l'attraction de l'attention d'un robot. De même, le pointage serait facilité si le robot avait une caméra 3D permettant d'extraire de manière relativement précise la position pointée par un geste de l'humain. Cependant, en plus d'améliorer les capacités du robot, il faudrait aussi permettre à l'humain de comprendre ce que le robot perçoit. Par exemple, si un robot n'a pas une apparence anthropomorphique, et s'il n'est pas doté de capteurs semblables à ceux d'un humain, savoir précisément ce qu'il regarde restera difficile pour un humain non-expert. Nous pensons donc, que tant que les robots personnels ne seront pas dotés de capacités sociales aussi sophistiquées que à celles d'un humain, l'interface jouera donc toujours un rôle crucial dans l'apprentissage social. De plus, les problèmes tels que la segmentation ou la catégorisation d'exemples d'apprentissage contournés par nos interfaces sont indépendants du type de robot utilisé.

D'autres problématiques, telles que la navigation, sont très influencées par le choix du robot. Il serait donc très intéressant de tester nos interfaces (ou de nouvelles) avec d'autres types de robot. L'utilisation d'un robot à roues aurait probablement modifié complètement la perception de l'interface lors de la navigation par

exemple.

10.5 Transposition de notre approche à d'autres domaines d'applications

Dans cette thèse, nous avons présenté une démarche générale visant à étudier le rôle de l'interface et des facteurs humains dans l'apprentissage social chez les robots et notamment pour faciliter les situations d'attention partagée. Nous nous sommes, en particulier, attachés à permettre un apprentissage en milieu non-contraint à des utilisateurs non-experts. Nous avons en particulier présenté un cycle de développement visant à garantir l'utilisabilité réelle et perçue d'une interface. Nous avons aussi montré comment les jeux robotiques pouvaient être un moyen efficace de concevoir des expériences utilisateurs en justifiant une tâche parfois abstraite pour les participants. Ils restaient ainsi impliqués, concernés et motivés jusqu'à la fin de l'expérience.

Cette approche générale a été appliquée à un problème d'apprentissage social particulier : l'enseignement d'objets visuels nouveaux associés à des mots nouveaux à un robot personnel. Nous avons choisi de nous attaquer à ce problème car il représente un levier très important pour le développement de la robotique personnelle et le guidage d'un robot dans sa découverte de son environnement immédiat. De plus, les problématiques d'attention jouent un rôle crucial dans la résolution de ce problème. Il était donc particulièrement intéressant d'étudier l'interface et les facteurs humains dans ce contexte. Nous avons ainsi pu montrer qu'une interface bien conçue et adaptée à cette tâche permettait d'améliorer de manière très sensible la qualité des exemples d'apprentissage et donc les performances du système de reconnaissance vocale et visuelle dans son ensemble.

Bien qu'appliquée à un domaine spécifique, notre démarche pourrait être transposée à d'autres domaines de l'interaction humain-robot et à l'apprentissage social chez les robots, où l'interface et les facteurs humains jouent aussi un rôle fondamental et ont donc également un impact potentiellement important. L'identification des mécanismes utilisés par l'humain pour faciliter l'apprentissage, leur transposition fonctionnelle, mais aussi le cycle de développement centré utilisateur visant à réellement prendre en compte les besoins de l'utilisateur, ainsi que la conception d'études utilisateurs valides, sont autant de clés qui pourraient permettre d'obtenir le même type de résultats que ceux présentés dans cette thèse dans d'autres domaines. Nous allons dans les sections suivantes présenter des pistes possibles de transposition de notre démarche à d'autres domaines.

10.5.1 Montrer des objets visuels à un robot

L'enseignement d'objets visuels nouveaux associés à des mots nouveaux à un robot nous a notamment permis d'étudier le rôle de l'interface et des facteurs humains lorsqu'on souhaite permettre à un humain de montrer des objets à un robot. En effet, attirer l'attention d'un robot vers un objet particulier est la première étape

nécessaire à son association avec un mot nouveau. Montrer des objets visuels à un robot est un pré-requis à de nombreux autres domaines applicatifs : e.g. désigner un objet visuel à un robot afin qu'il l'attrape. Choi et al. ont par exemple présenté un système basé sur un pointeur laser ou un écran tactile pour désigner des objets visuels à un bras robotisé monté sur une chaise roulante afin qu'il les attrapent [Choi 2008]. De même, Tsui et al. ont proposé différentes interfaces basées sur un écran tactile pour permettre à un utilisateur de désigner des objets à un bras robotique fixé à une chaise roulante [Tsui 2008]. Dans ces travaux, les auteurs supposent que le robot puisse attraper les objets simplement à partir d'un emplacement 3D donné par l'utilisateur et supposent donc que le robot puisse séparer automatiquement l'objet visuel du reste de la scène. Or, comme discuté dans la section 6.5.2, la segmentation automatique nécessite une connaissance a priori des objets. Nos interfaces, et particulièrement l'interface iPhone où l'utilisateur encercle l'objet visuel à montrer, permettent d'obtenir directement une segmentation faite par l'utilisateur. Plus généralement, contourner ce problème, en utilisant l'encerclement d'un objet visuel par les utilisateurs, pourrait s'appliquer à d'autre applications, telles que la robotique de service. L'utilisateur pourrait, à travers ce geste, désigner une zone à nettoyer ou désigner un objet visuel à surveiller ou à éviter.

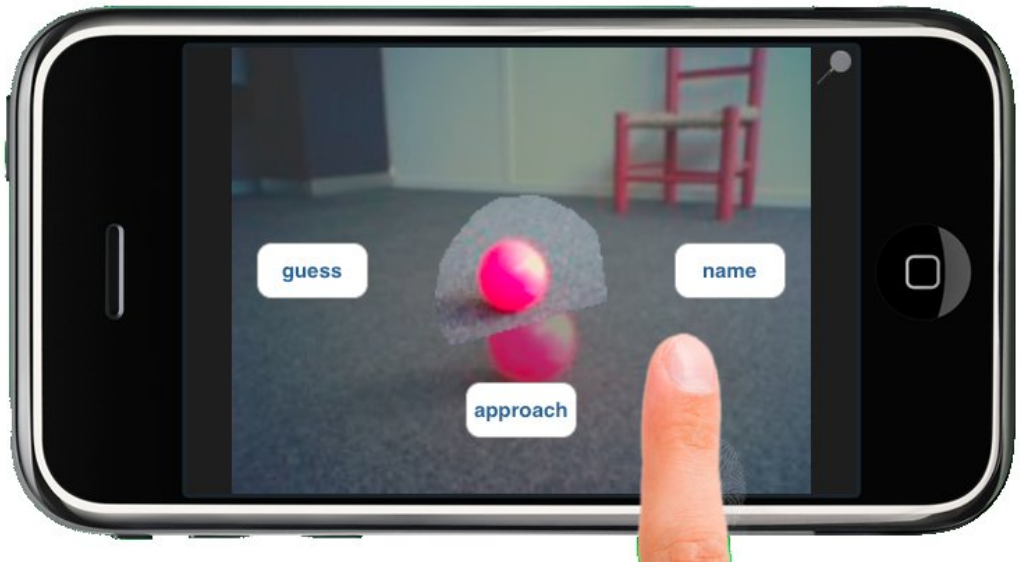


FIGURE 10.1 – Cette figure présente un prototype d'interface où un menu contextuel apparait après que l'utilisateur ait entouré un objet visuel afin de le montrer à un robot. Ce menu présente différentes possibilités d'interaction avec cet objet visuel. Pour chacune d'elle l'interface pourrait permettre à l'utilisateur de superviser son exécution mais aussi d'intervenir dans le processus afin d'aider le robot.

En plus d'aider à la segmentation et de faciliter les situations d'attention partagée entre un humain et un robot sur un objet visuel, une interface bien conçue pourrait aussi permettre à l'utilisateur de superviser les différentes actions liées à un

objet, comme l'attraper ou le suivre. Cela permettrait, à la fois que l'utilisateur ait une meilleure compréhension du comportement du robot (à travers différents retours sur l'écran d'un smartphone par exemple) mais aussi lui permettre d'intervenir dans ce processus et ainsi faciliter son exécution. Nous pourrions, par exemple, imaginer présenter un ensemble de retours visuels à l'utilisateur lui décrivant la manière dont le robot va attraper l'objet. En effet, la préhension reste un problème difficile sans connaissance a priori de la forme d'un objet [Bicchi 2002][Saxena 2008]. L'utilisateur pourrait alors corriger, si besoin, la manière dont le robot prévoit de se saisir de l'objet, en précisant par exemple, où le robot devrait placer ses doigts sur l'objet. Une telle interface pourrait permettre à un utilisateur d'enseigner à un robot comment se saisir d'objets visuels nouveaux ayant des formes a priori inconnues. Une interface bien conçue pourrait aussi aider l'utilisateur à comprendre la détection/recherche d'un objet visuel par un robot. Comme présenté dans le chapitre ASMAT, l'humain peut suivre le processus de recherche d'un objet visuel via l'interface iPhone, à travers des retours visuels tels que le dessin de la zone correspondant à l'objet visuel détecté (voir la figure 8.3).

La figure 10.1 présente un prototype d'interface où un menu contextuel apparait lorsque l'utilisateur encercle un objet visuel. Il peut alors sélectionner une des options possibles lui permettant de demander au robot d'attraper l'objet, de le surveiller ou encore de le nommer. L'utilisateur pourrait aussi apprendre à son robot que certaines de ces actions ne peuvent pas s'appliquer à certains objets visuels (e.g. attraper un objet fixe).

10.5.2 Télé-opération et télé-présence

Dans cette thèse, nous nous sommes concentrés sur des interactions humain-robot en co-location, c'est-à-dire où l'utilisateur est présent physiquement à côté de son robot. Ce type d'interaction correspond en effet à notre contexte d'apprentissage social. Notre démarche et plus précisément l'interface basée sur un iPhone, peut cependant être transposée à une utilisation en télé-opération, où l'utilisateur interagit à distance avec le robot.

Les problématiques liées à l'interface ont été particulièrement prises en considération dans le domaine de la télé-opération d'un robot. En effet, lors d'une interaction en télé-opération, l'utilisateur ne peut connaître la situation du robot qu'à travers une interface et non par vérification directe. Il est donc nécessaire de proposer à l'utilisateur d'autres moyens de connaître la position du robot, son état fonctionnel, ce qu'il est en train de faire, etc. Cette question peut être particulièrement critique, par exemple dans le cadre du sauvetage-déblaiement (*USAR : Urban Search And Rescue*) [Yanco 2004a].

La capacité de savoir ce que le robot perçoit ainsi que la possibilité de le diriger efficacement sont des caractéristiques cruciales en télé-opération. L'interface iPhone pourrait donc être utilisée dans ce cadre là. Notre démarche ne visait cependant pas à proposer des interfaces maximisant l'efficacité de la navigation mais plutôt à la rendre intuitive et ludique. Plutôt que d'utiliser cette interface dans le cadre

d'applications où la robustesse est critique, nous pourrions imaginer la transposer à des applications de télé-opération où l'interaction sociale joue aussi un rôle important. Par exemple, la télé-présence robotique, un sous domaine de la télé-opération, cherche à permettre à un utilisateur d'avoir l'impression d'être présent sur un site distant à travers un robot (e.g. ANYBOTS [1], JAZZ [2], VGo [3]). En plus des aspects de localisation et de perception de la situation, il est important dans ce cas de transmettre également des informations relatives à l'aspect social de l'interaction telles que des indicateurs de l'état émotionnel ou encore la capacité de pouvoir suivre le regard d'une personne. L'interface iPhone ainsi que l'étude des questions liées à la robotique sociale pourraient donc s'appliquer ici.

10.5.3 Apprentissage social de mouvements et d'actions

Comme nous l'avons expliqué tout au long de cette thèse, nous avons étudié le rôle de l'interface dans l'apprentissage social chez les robots. Nous avons choisi d'appliquer cette étude à un exemple particulier : l'apprentissage social du langage. Certaines des problématiques rencontrées lors du travail de cette thèse, telles que l'importance de faciliter le partage d'attention entre un humain et un robot, ou encore les problèmes de différences d'appareil sensori-moteur, ne sont pas spécifiques à l'apprentissage du langage chez les robots. Nous pouvons aussi imaginer que les solutions proposées, comme l'interface iPhone par exemple, pourraient être transposées plus ou moins directement à d'autres types d'apprentissage social, et en particulier à l'apprentissage social de mouvements et d'actions à un robot.

L'apprentissage de mouvements et d'actions est également un des défis majeurs de la robotique personnelle et, en particulier, pour l'adaptation d'un robot à son environnement. Une des réponses les plus couramment apportées à ce problème est l'apprentissage par démonstration [Billard 2008]. L'imitation en robotique a principalement traitée les questions d'encodage, de reproduction et de généralisation des démonstrations. Pourtant, comme l'a souligné Calinon, le démonstrateur joue aussi un rôle majeur dans l'apprentissage par imitation [Calinon 2007b]. Plus généralement, l'interface et les facteurs humains ont une influence majeure dans l'apprentissage social de mouvements et d'actions.

Une interface bien conçue pourrait, par exemple, aider l'humain à indiquer à un robot, quelle(s) composante(s) d'une démonstration sont importantes. De la même manière que l'interface basée sur un iPhone, décrite dans cette thèse, permet à un humain et un robot de focaliser leur attention sur un même objet, nous pourrions imaginer une interface, où un retour vidéo de la démonstration, telle que perçue par le robot, serait présentée à l'utilisateur. Cela permettrait tout d'abord d'aider l'humain à mieux comprendre ce que le robot a perçu de la démonstration et d'adapter les prochaines démonstrations en conséquence. D'autre part, l'utilisateur pourrait, par exemple à l'aide d'un geste d'encerclement, sélectionner les éléments clés du

1. https://www.anybots.com/
2. http://www.gostai.com/connect/
3. http://www.vgocom.com/

mouvement. Par exemple, pour apprendre à un robot à se saisir d'une bouteille, l'humain pourrait entourer la bouteille et l'extrémité du bras sur le retour vidéo, indiquant ainsi au robot que ce sont là les éléments pertinents de la démonstration.

L'interface pourrait aussi jouer un rôle important dans l'enseignement de mouvements via l'apprentissage par renforcement. Dans ce contexte, le robot effectue, par lui-même, des mouvements. L'humain ne fait que guider le robot à travers des retours sur les mouvements/actions effectués. Ces retours sont généralement du type « bien » ou « pas bien ». L'interface pourrait permettre d'enrichir ces retours. Une question centrale de cette problématique est, par exemple, de savoir comment le robot peut identifier ce sur quoi s'appliquent les retours de l'utilisateur : au mouvement entier ? à une sous-partie ? si oui, à laquelle ? Là encore, une interface présentant un schéma des actions effectuées par le robot, permettrait à l'utilisateur de sélectionner les composantes du mouvement sur lesquelles ses commentaires s'appliquent. L'interface pourrait aussi être utilisée par l'humain pour contraindre un ensemble d'articulations du robot lors de l'exploration. De la même manière, qu'un professeur de tennis indique à ses élèves qu'ils doivent bloquer leur poignet, l'humain pourrait via l'interface, indiquer à un robot apprenant à frapper des coups-droits, qu'il doit maintenir son poignet fixe et dans une position donnée.

10.6 Le jeu robotique

Au chapitre 7, nous avons décrit un jeu robotique, conçu pour rendre une étude utilisateurs plus intéressante pour les participants. Ce jeu représente, selon nous, l'un des premiers jeux robotiques développés. En effet, à notre connaissance, très peu de jeu mettant en scène des interactions avec un robot ont été proposés. Parmi les exemples existants, nous pouvons citer le drone Parrot [4]. Il est possible de contrôler ce drone à l'aide d'un iPhone. Des jeux en réalité augmenté ont été spécialement conçus permettant à l'utilisateur de piloter le drone et de combattre des ennemis virtuels. Bien qu'intéressant, il ne s'agit pas ici d'un exemple de robotique sociale et les interactions que l'utilisateur peut avoir avec son drone sont très limitées. Un autre exemple est le robot Keepon [5]. Ce robot très simple peut danser en rythme sur une musique. Les utilisateurs peuvent aussi interagir physiquement avec lui, afin de lui indiquer un rythme à suivre [Kozima 2009]. Le robot Nao a aussi été doté de la capacité de jouer au Puissance 4 [6]. Xin et Sharlin ont conçu un jeu de plateau en réalité augmenté où des robots, représentants les pions, se déplaçaient sur un damier [Xin 2007]. Ce jeu a été conçu comme un moyen d'évaluation de l'interaction humain-robot.

Dans cette thèse, nous avons montré comment le jeu robotique pouvait permettre de concevoir une expérience robotique, hors du laboratoire, qui maintienne les utilisateurs impliqués et motivés. Nous avons aussi utilisé les mécanismes clas-

4. http://ardrone.parrot.com/parrot-ar-drone/fr/
5. http://beatbots.net/
6. http://www.generationrobots.com/site/program-nao-robot/

siques des jeux vidéos, tels que les tutoriels ou les vidéos de présentation, afin de créer un protocole expérimentale qui soit à la fois simple, efficace et reproductible [Rouanet 2010a]. Bien que non centrale dans cette thèse, nous avons commencé à explorer la conception de jeu robotique, dont le seul but est le jeu lui-même. Dans les questionnaires, décrits en section 7.3.3, certaines questions portaient sur le jeu robotique. Nous avons ainsi pu constater que les participants ont trouvé notre jeu distrayant, et ils ont indiqué souhaiter participer à d'autres jeux de ce type dans le futur. Il est aussi intéressant de remarquer que les participants ont très rapidement adhéré au concept de jeu robotique, même s'il était inconnu pour la plupart d'entre eux. Les personnes ayant participé à notre jeu devaient en effet comprendre l'objectif du jeu, apprendre à y jouer et enfin y participer dans un temps très limité (moins de 30 minutes en moyenne). Le jeu robotique semble donc recevoir un accueil très positif de la part du public. Cependant, nous avons aussi pu constater que notre jeu était probablement trop complexe. Les utilisateurs devaient surveiller, le robot, leur interface ainsi que l'interface du jeu. De plus, comme nous l'avons indiqué ci-dessus, l'expérience était très courte. Bien que notre jeu était distrayant pendant 30 minutes, il est probable qu'il devienne très rapidement lassant au delà de cette durée. En particulier, notre jeu était entièrement pré-codé et les répliques du robot était pré-écrites. Il n'avait donc aucune rejouabilité, c'est-à-dire que les joueurs n'éprouvaient aucun intérêt à recommencer le jeu une fois terminé.

Il serait très intéressant d'explorer les critères à prendre en compte lors de la conception de jeux robotiques. Avec le développement de la robotique personnelle, ce type de jeux pourrait devenir de plus en plus populaire et occuper une place importante dans nos sociétés. Comme souligné par Brooks et al., après les jeux vidéo, les jeux en réalité augmenté et les jeux avec retours haptiques, les jeux robotiques pourraient être la prochaine étape, offrant une variété encore plus grande de possibilités d'interactions à l'utilisateur [Brooks 2004]. Nous allons proposer ici une liste de critères/défis à prendre en compte, selon nous, lors de la conception de jeux robotiques. Il est important de noter que cette liste n'est que le résultat de nos observations informelles et d'une étude très exploratoire. Elle n'est donc constituée que d'hypothèses et n'est pas exhaustive. Il serait indispensable de mener d'autres expériences pour pouvoir étudier ces questions plus en détails.

1. **Robustesse et intuitivité de l'interaction :** Il semble crucial de permettre aux utilisateurs d'interagir avec un robot de manière intuitive, transparente et robuste. En effet, dans ce contexte de jeu, les utilisateurs souhaiteront sûrement pouvoir interagir directement avec un robot, sans avoir besoin de s'entrainer auparavant. De plus, les utilisateurs devraient être capable d'identifier facilement les capacités du robot, afin d'éviter les malentendus et/ou les déceptions lors de l'interaction, qui pourraient dégrader l'expérience utilisateur très sensiblement.
2. **Interaction physique sûre :** Il est probable que la plupart des jeux robotiques se dérouleront en co-location. Les contacts entre l'humain et le robot seront inévitables. Les utilisateurs devront pouvoir interagir physiquement avec

le robot (e.g. le toucher, le pousser) sans être blessés. Le robot devra aussi pouvoir se cogner ou tomber sans être détruit ou devoir être réparé.

3. **Interactions sur le long-terme :** Un autre défi majeur est de permettre à un humain et à un robot d'interagir sur des périodes longues et répétées. Or, très peu d'expériences en HRI ont étudié la question de l'interaction avec un robot sur de longues périodes. Il serait sans doute indispensable que le robot ait différents degrés d'autonomie, c'est-à-dire qu'il puisse jouer avec l'humain lorsque celui-ci est disponible, mais également qu'il puisse jouer « dans son coin » lorsque l'utilisateur est occupé, ou n'a pas envie de jouer. Le robot devrait aussi être capable d'adapter son comportement à la situation, à l'utilisateur et à leur humeur. Pour éviter que les humains se lassent rapidement, le robot devrait aussi être doté de mécanismes d'apprentissage, afin qu'il acquiert des capacités nouvelles.

Malgré de très nombreux défis à relever, la perspective de développer des jeux robotiques nous semblent passionnante, aussi bien du point de vue des défis scientifiques et technologiques que du point de vue sociétal. Le développement de la RoboCup en représente une illustration parfaite [7].

7. `http://www.robocup.org/`

Conclusion

Dans cette thèse, nous avons étudié le rôle de l'interface dans l'apprentissage social en robotique. Plus précisément, nous avons examiné comment une interface bien conçue peut aider les utilisateurs non-experts à guider l'apprentissage social d'un robot, notamment en facilitant les situations d'attention partagée. Cette étude a été réalisée dans le cadre de l'enseignement conjoint de mots et d'objets visuels nouveaux à un robot, qui peut jouer un rôle majeur dans le développement de la robotique personnelle.

Nous avons commencé par montrer comment des interfaces, basées sur des objets médiateurs, peuvent aider un utilisateur non-expert à attirer l'attention d'un robot, à désigner des objets visuels particuliers, et également, à faciliter les situations d'attention partagée. L'interface, si elle est bien conçue, peut aussi pousser les utilisateurs à fournir des exemples d'apprentissage de bonne qualité, qui amélioreront les performances générales du système de reconnaissance visuelle. Nous avons aussi proposé un système de collecte semi-automatique d'exemples d'apprentissage permettant de réduire le nombre d'interventions de l'utilisateur.

Nous avons ensuite étudié comment l'interface pouvait permettre d'améliorer les performances du système de reconnaissance vocale, en proposant à l'utilisateur une liste de résultats les plus proches lors de la recherche d'un objet visuel. Cette interaction permet également de rendre la recherche plus transparente et donc plus compréhensible pour l'utilisateur.

Enfin, nous avons montré comment, à travers des interactions transparentes et peu nombreuses, les utilisateurs peuvent regrouper incrémentalement les différents exemples d'apprentissage audio-visuels et ainsi fournir au robot une clusterisation de plus en plus complète. Une simulation de l'utilisation de cette interface a montré qu'une telle approche permet d'obtenir un apprentissage presque aussi performant que lorsque l'on suppose l'utilisation de symboles dans une des modalités.

Nous avons également évalué l'utilisabilité perçue de nos interfaces par des humains non-experts, ainsi que l'expérience utilisateur lors de l'interaction. Nos interfaces, basées sur des objets médiateurs, ont été jugées intuitives, faciles à utiliser et plaisantes. Une interface gestuelle, utilisant un système de reconnaissance et d'interprétation de gestes aussi performant qu'un humain, a été jugée moins intuitive et moins utilisable par les participants.

Nous avons montré que la conception d'études utilisateurs sous forme de jeu robotique permet, en plus de définir un protocole d'expérimentation strict et reproductible, de maintenir les participants motivés et concernés tout au long des expériences d'interaction humain-robot.

Nous avons finalement discuté des extensions possibles au travail de cette thèse, telles que l'étude d'une solution de clusterisation semi-automatique, l'utilisation d'autres robots, ou enfin la transposition de notre approche à d'autres domaines d'applications.

Bibliographie

[Adams 1994] R. Adams et L. Bischof. *Seeded Region Growing.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 16, no. 6, pages 641–647, 1994. 84

[Adams 2000] Bryan Adams, Cynthia Breazeal, Rodney A. Brooks et Brian Scassellati. *Humanoid Robots : A New Kind of Tool.* IEEE Intelligent Systems, vol. 15, pages 25–31, July 2000. 27

[Aimetti 2009] Guillaume Aimetti. *Modelling Early Language Acquisition Skills : Towards a General Statistical Learning Mechanism.* Computational Linguistics, no. April, pages 1–9, 2009. 30

[Alami 2006] R. Alami, A. Albu-Schaeffer, A. Bicchi, R. Bischoff, R. Chatila, A. De Luca, A. De Santis, G. Giralt, J. Guiochet, G. Hirzinger, F. Ingrand, V. Lippiello, R. Mattone, D. Powell, S. Sen, B. Siciliano, G. Tonietti et L. Villani. *Safe and Dependable Physical Human-Robot Interaction in Anthropic Domains : State of the Art and Challenges.* In A. Bicchi et A. De Luca, editeurs, Procceedings IROS Workshop on pHRI - Physical Human-Robot Interaction in Anthropic Domains, Beijing, China, Octobre 2006. 3

[Angeli 2008] A. Angeli, D. Filliat, S. Doncieux et J.-A. Meyer. *Real-Time Visual Loop-Closure Detection.* In Proceedings of the International Conference on Robotics and Automation (ICRA), 2008. 67, 138

[Argall 2009] Brenna D. Argall, Sonia Chernova, Manuela Veloso et Brett Browning. *A survey of robot learning from demonstration.* Robot. Auton. Syst., vol. 57, pages 469–483, May 2009. 6, 16

[Armon–Jones 1986] C. Armon–Jones. *The Social Functions of Emotions.* In R. harre, editeur, The social construction of emotions, pages 57–82. Blackwell, Oxford, 1986. 19

[Arsenio 2003] Artur Arsenio, Paul Fitzpatrick, Charles C. Kemp et Giorgio Metta. *The Whole World in Your Hand : Active and Interactive Segmentation*, 2003. 84

[Atienza 2003] R Atienza et A Zelinsky. *Intuitive human-robot interaction through active 3d gaze tracking.* In Proceedings of the International Symposium of Robotics Research, pages 172–181. Springer, 2003. 21, 33

[Bab-Hadiashar 2006] A. Bab-Hadiashar et N. Gheissari. *Range image segmentation using surface selection criterion.* Image Processing, IEEE Transactions on, vol. 15, no. 7, pages 2006–2018, July 2006. 85

[Bahl 1983] L.R. Bahl, F. Jelinek et R.L. Mercer. *A maximum likelihood approach to continuous speech recognition.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, no. 2, pages 179–190, 1983. 115

[Bartneck 2009] Christoph Bartneck, Takayuki Kanda, Omar Mubin et Abdullah Al Mahmud. *Does the Design of a Robot Influence Its Animacy and Perceived*

Intelligence ? International Journal of Social Robotics, vol. 1, no. 2, pages 195–204, Avril 2009. 5

[Bay 2008] Herbert Bay, Andreas Ess, Tinne Tuytelaars et Luc Van Gool. *Speeded-Up Robust Features (SURF).* Comput. Vis. Image Underst., vol. 110, no. 3, pages 346–359, 2008. 37, 68, 134

[Bicchi 2002] A. Bicchi et V. Kumar. *Robotic grasping and contact : a review.* In Robotics and Automation, 2000. Proceedings. ICRA '00. IEEE International Conference on, volume 1, pages 348–353, Août 2002. 4, 143

[Billard 1997] A. Billard et K. Dautenhahn. *Grounding communication in situated, social robots.* In Proceedings Towards Intelligent Mobile Robots Conference, Report No. UMCS-97-9-1, Department of Computer Science, Manchester University. Citeseer, 1997. 31

[Billard 2008] A. Billard, S. Calinon, R. Dillmann et S. Schaal. *Survey : Robot Programming by Demonstration.* Handbook of Robotics, . chapter 59, 2008, 2008. 6, 7, 16, 144

[Bogdan 2011] Cristian Bogdan, Dominik Ertl, Helge Hüttenrauch, Michael Göller, Anders Green, Kerstin Severinson-Eklundh, Jürgen Falb et Hermann Kaindl. New frontiers in human–robot interaction, pages 185–210. John Benjamins Publishing Company, 2011. 19

[Bouchard 2005] Guillaume Bouchard. *Hierarchical part-based visual object categorization.* In Proc. CVPR, pages 710–715, 2005. 37, 38

[Breazeal 2002a] Cynthia Breazeal. Designing sociable robots. Bradford book - MIT Press, Cambridge, MA, 2002. 3, 4, 7, 10, 16, 17, 18, 19, 27, 28, 33

[Breazeal 2002b] Cynthia Breazeal et Lijin Aryananda. *Recognition of Affective Communicative Intent in Robot-Directed Speech.* Auton. Robots, vol. 12, pages 83–104, January 2002. 32

[Breazeal 2003] Cynthia Breazeal. *Emotion and sociable humanoid robots.* Int. J. Hum.-Comput. Stud., vol. 59, pages 119–155, July 2003. 19

[Breazeal 2005] Cynthia Breazeal, Daphna Buchsbaum, Jesse Gray, David Gatenby et Bruce Blumberg. *Learning From and About Others : Towards Using Imitation to Bootstrap the Social Understanding of Others by Robots.* Artif. Life, vol. 11, pages 31–62, January 2005. 6, 7

[Brooks 1999] Rodney A. Brooks, Cynthia Breazeal, Matthew Marjanovic, Brian Scassellati et Matthew M. Williamson. *The Cog project : Building a humanoid robot.* In Lecture Notes in Computer Science, pages 52–87. Springer-Verlag, 1999. 27

[Brooks 2004] Andrew G. Brooks, Jesse Gray, Guy Hoffman, Andrea Lockerd, Hans Lee et Cynthia Breazeal. *Robot's play : interactive games with sociable machines.* Comput. Entertain., vol. 2, no. 3, pages 10–10, 2004. 146

[Bumby 1999] K. E. Bumby et K. Dautenhahn. *Investigating children's attitudes towards robots : A case study.* In K. Cox, B. Gorayska et J. Marsh, editeurs,

Proceedings of the Third Cognitive Technology Conference, CT'99, pages 391–410. M.I.N.D. Lab, Michigan State University, East Lansing, MI, 1999. 18

[Cakmak 2009] M. Cakmak, N. DePalma, A.L. Thomaz et R. Arriaga. *Effects of social exploration mechanisms on robot learning.* In Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on, pages 128 –134, 27 2009-oct. 2 2009. 8, 16

[Calinon 2003] S. Calinon. Pda interface for humanoid robots using speech and vision processing. 2003. 24

[Calinon 2007a] S. Calinon, F. Guenter et A. Billard. *On Learning, Representing, and Generalizing a Task in a Humanoid Robot.* Systems, Man, and Cybernetics, Part B : Cybernetics, IEEE Transactions on, vol. 37, no. 2, pages 286 –298, april 2007. 7, 16

[Calinon 2007b] Sylvain Calinon et Aude G. Billard. *What is the Teacher's Role in Robot Programming by Demonstration ? Toward Benchmarks for Improved Learning.* Interaction Studies. Special Issue on Psychological Benchmarks in Human-Robot Interaction, vol. 8, no. 3, 2007. 8, 16, 144

[Canamero 2001] L Canamero et J Fredslund. *I Show You How I Like You-Can You Read it in My Face ?* IEEE Transactions on Systems Man and Cybernetics Part A, vol. 31, no. 5, pages 454–459, 2001. 19

[Cangelosi 2006] A. Cangelosi et T. Riga. *An embodied model for sensorimotor grounding and grounding transfer : experiments with epigenetic robots.* Cognitive science, vol. 30, no. 4, pages 673–689, 2006. 31

[Cangelosi 2010] Angelo Cangelosi, Giorgio Metta, Gerhard Sagerer, Stefano Nolfi, Chrystopher Nehaniv, Kerstin Fischer, Jun Tani, Tony Belpaeme, Giulio Sandini, Francesco Nori, Luciano Fadiga, Britta Wrede, Katharina Rohlfing, Elio Tuci, Kerstin Dautenhahn, Joe Saunders et Arne Zeschel. *Integration of Action and Language Knowledge : A Roadmap for Developmental Robotics.* IEEE Transactions on Autonomous Mental Development, vol. 2, no. 3, pages 167–195, Septembre 2010. 4, 31

[Chetouani 2009a] M. Chetouani, M. Faundez-Zanuy, B. Gas et J.L. Zarader. *Investigation on LP-Residual Representations For Speaker Identification.* Pattern Recognition, vol. 42, no. 3, pages 487–494, 2009. 3, 118

[Chetouani 2009b] M. Chetouani, A. Mahdhaoui et F. Ringeval. *Time-scale feature extractions for emotional speech characterization.* Cognitive Computation, vol. 1, no. 2, pages 194–201, June 2009. 3

[Choi 2008] Young Sang Choi, Cressel D. Anderson, Jonathan D. Glass et Charles C. Kemp. *Laser pointers and a touch screen : intuitive interfaces for autonomous mobile manipulation for the motor impaired.* In Assets '08 : Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility, pages 225–232, New York, NY, USA, 2008. ACM. 24, 142

[Correa 2009] Mauricio Correa, Javier Ruiz-Del-Solar et Fernando Bernuy. Face recognition for human-robot interaction applications : A comparative study, pages 473–484. Springer-Verlag, Berlin, Heidelberg, 2009. 3

[Cowie 2001] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz et J. G. Taylor. *Emotion recognition in human-computer interaction.* Signal Processing Magazine, IEEE, vol. 18, no. 1, pages 32–80, Janvier 2001. 3

[Csurka 2004] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski et Cedric Bray. *Visual categorization with bags of keypoints.* In Workshop on Statistical Learning in Computer Vision, ECCV, pages 1–22, 2004. 37, 67

[Dahl 2011] TorbjÃ¸rn S Dahl, Erick A R Swere et Andrew Palmer. New frontiers in human–robot interaction, pages 281–303. John Benjamins Publishing Company, 2011. 27

[Dalgalarrondo 2004] Dalgalarrondo, Dufourd et Filliat. *Controlling the autonomy of a reconnaissance robot.* In SPIE Defense & Security 2004 Symposium. Unmanned Ground Vehicle Technology VI Conference, 2004. 23

[Dautenhahn 1999] Kerstin Dautenhahn et Aude Billard. *Bringing up robots or the psychology of socially intelligent robots : from theory to implementation.* In Proceedings of the third annual conference on Autonomous Agents, AGENTS '99, pages 366–367, New York, NY, USA, 1999. ACM. 15

[Dautenhahn 2002a] K. Dautenhahn et A. Billard. *Games children with autism can play with robota, a humanoid robotics doll.* In Proceedings of the 1st Cambridge Workshop on Universal Access and Assistive Technology, numéro 1, 2002. 5, 17, 27

[Dautenhahn 2002b] Kerstin Dautenhahn, Bernard Ogden et Tom Quick. *From embodied to socially embedded agents - Implications for interaction-aware robots.* Cognitive Systems Research, pages 397–428, 2002. 10, 18

[Dominey 2004a] Peter Ford Dominey et Jean-David Boucher. *Developmental Stages of Perception and Language Acquisition in a Perceptually Grounded Robot*, 2004. 31, 32

[Dominey 2004b] P.F. Dominey, J.-D. Boucher et T. Inui. *Building an adaptive spoken language interface for perceptually grounded human-robot interaction.* 4th IEEE/RAS International Conference on Humanoid Robots, 2004., pages 168–183, 2004. 31

[Dominey 2007] P. F. Dominey, A. Mallet et E. Yoshida. *Progress in Programming the HRP-2 Humanoid Using Spoken Language.* In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2007. 21, 32, 33

[Dominey 2009] PF Dominey et Anthony Mallet. *Real-time spoken-language programming for cooperative interaction with a humanoid apprentice.* International Journal of Humanoid, vol. 6, no. 2, pages 147–171, 2009. 21, 32, 33

[Dowson 2005] N. D. H. Dowson et R. Bowden. *Simultaneous Modeling and Tracking (SMAT) of Feature Sets.* In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02, CVPR '05, pages 99–105, Washington, DC, USA, 2005. IEEE Computer Society. 108

[Driesen 2009] Joris Driesen, Louis ten Bosch et Hugo Van hamme. *Adaptive non-negative matrix factorization in a computational model of language acquisition.* In INTERSPEECH, pages 1731–1734. ISCA, 2009. 132, 139

[Durrant-Whyte 2006] Hugh Durrant-Whyte et Tim Bailey. *Simultaneous Localisation and Mapping (SLAM) : Part I The Essential Algorithms.* IEEE Robotics & Automation Magazine, vol. 13, no. 2, pages 99–110, Juin 2006. 3

[Edin 2008] B. B. Edin, L. Ascari, L. Beccai, S. Roccella, J-J J. Cabibihan et M. C. Carrozza. *Bio-inspired sensorization of a biomechatronic robot hand for the grasp-and-lift task.* Brain research bulletin, vol. 75, no. 6, pages 785–795, Avril 2008. 4

[Eklundh 2003] Kerstin Severinson Eklundh, Anders Green et Helge HÃ$\frac{1}{4}$ttenrauch. *Social and collaborative aspects of interaction with a service robot.* In Special Issue on Socially Interactive Robots, Robotics and Autonomous Systems 42 (34, 2003. 27

[Everingham 2010] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn et A. Zisserman. *The Pascal Visual Object Classes (VOC) Challenge.* International Journal of Computer Vision, vol. 88, no. 2, pages 303–338, Juin 2010. 38

[Evrard 2009] P. Evrard, E. Gribovskaya, S. Calinon, A. Billard et A. Kheddar. *Teaching physical collaborative tasks : Object-lifting case study with a humanoid.* In Proc. IEEE-RAS Intl Conf. on Humanoid Robots (Humanoids), pages 399–404, December 2009. 7

[Filliat 2003] D. Filliat et J.-A. Meyer. *Map-based navigation in mobile robots - I. A review of localisation strategies.* Journal of Cognitive Systems Research, vol. 4, no. 4, pages 243–282, 2003. 3

[Filliat 2007] D. Filliat. *A visual bag of words method for interactive qualitative localization and mapping.* In Proceedings of the International Conference on Robotics and Automation (ICRA), 2007. 37, 67

[Filliat 2008] D. Filliat. *Interactive learning of visual topological navigation.* In Proceedings of the 2008 IEEE International Conference on Intelligent Robots and Systems (IROS 2008), 2008. 67, 138

[Fong 2001] Terrence Fong, Nathalie Cabrol, Charles Thorpe et Charles Baur. *A Personal User Interface for Collaborative Human-Robot Exploration.* In 6th International Symposium on Artificial Intelligence, Robotics, and Automation in Space (iSAIRAS), Montreal, Canada, June 2001. 22, 23

[Fong 2002] T. Fong, I. Nourbakhsh et K. Dautenhahn. *A Survey of Socially Interactive Robots : Concepts, Design, and Applications.* 2002. 10, 16, 17

[Fong 2003a] Terrence W. Fong, Illah Nourbakhsh et Kerstin Dautenhahn. *A survey of socially interactive robots.* Robotics and Autonomous Systems, 2003. 2, 28

[Fong 2003b] Terrence W Fong, Chuck Thorpe et Betty Glass. *PdaDriver : A Handheld System for Remote Driving.* In IEEE International Conference on Advanced Robotics 2003. IEEE, July 2003. 23

[Gams 2008] Andrej Gams et Pierre-André Mudry. *Gaming controllers for research robots : controlling a humanoid robot using a WIIMOTE.* In Proc. of the 17th Int. Electrotechnical and Computer Science Conference (ERK08), pages 191–194, 2008. 25

[Gates 2007] B. Gates. *A Robot in Every Home.* Scientific American, January 2007. http ://www.sciam.com/article.cfm ?id=a-robot-in-every-home. 2

[Goetz 2002] Jennifer Goetz et Sara Kiesler. *Cooperation with a robotic assistant.* In CHI '02 extended abstracts on Human factors in computing systems, CHI EA '02, pages 578–579, New York, NY, USA, 2002. ACM. 19

[Goodrich 2007] Michael A. Goodrich et Alan C. Schultz. *Human-robot interaction : a survey.* Found. Trends Hum.-Comput. Interact., vol. 1, pages 203–275, January 2007. 4, 10, 20

[Green 2004] A. Green, H. Huttenrauch et K.S. Eklundh. *Applying the Wizard-of-Oz framework to cooperative service discovery and configuration.* In Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on, pages 575–580. IEEE, 2004. 79

[Guo 2008a] Cheng Guo et Ehud Sharlin. *Exploring the use of tangible user interfaces for human-robot interaction : a comparative study.* In CHI '08 : Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, pages 121–130, New York, NY, USA, 2008. ACM. 23, 25

[Guo 2008b] Cheng Guo et Ehud Sharlin. *Utilizing Physical Objects and Metaphors for Human Robot Interaction.* University of Calgary, Computer Science, Science, 2008. 25

[Haasch 2004] A. Haasch, S. Hohenner, S. Huwel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G. Fink, J. Fritsch, B. Wrede et G. Sagerer. *BIRON – The Bielefeld Robot Companion.* In Proc. Int. Workshop on Advances in Service Robotics Stuttgart Germany 2004 pp. 27–32., 2004. 4, 21, 25, 33

[Hachet 2008] Martin Hachet, Fabrice Dècle, Sebastian Knödel et Pascal Guitton. *Navidget for Easy 3D Camera Positioning from 2D Inputs.* In Proceedings of IEEE 3DUI - Symposium on 3D User Interfaces, 2008. best paper award. 74

[Haddadin 2010] S. Haddadin, A. Albu-Schaffer, O. Eiberger et G. Hirzinger. *New insights concerning intrinsic joint elasticity for safety.* In Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on, pages 2181 –2187, oct. 2010. 5, 27

[Hafner 2004] Verena V. Hafner et Frederic Kaplan. *Learning to Interpret Pointing Gestures : Experiments with four-legged autonomous Robots.* In in Proceedings of the KI2004 Workshop on Neurobotics, pages 225–234. Springer, 2004. 20, 33

[Hafner 2011] Verena Hafner, Manja Lohse, Joachim Meyer, Yukie Nagai et Britta Wrede. *The role of expectations in intuitive human-robot interaction.* In Proceedings of the 6th international conference on Human-robot interaction, HRI '11, pages 7–8, New York, NY, USA, 2011. ACM. 17

[Heerink 2006] Marcel Heerink, Ben Krose, Vanessa Evers et Bob Wielinga. *The Influence of a Robot's Social Abilities on Acceptance by Elderly Users.* In proceedings of RO-MAN, Hertfordshire, 2006. 5, 27

[Hermansky 1992] H. Hermansky, N. Morgan, A. Bayya et P. Kohn. *RASTA-PLP speech analysis technique.* In Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on, volume 1, pages 121–124. IEEE, 1992. 115

[Hinds 2004] Pamela J. Hinds, Teresa L. Roberts et Hank Jones. *Whose job is it anyway ? a study of human-robot interaction in a collaborative task.* Hum.-Comput. Interact., vol. 19, no. 1, pages 151–181, Juin 2004. 10

[Huttenrauch 2001] Helge Huttenrauch et Mikael Norman. *PocketCERO – mobile interfaces for service robots.* In In Proceedings of the Mobile HCI, International Workshop on Human Computer Interaction with Mobile Devices, 2001. 24

[Huttenrauch 2002] H Huttenrauch et K S Eklundh. *Fetch-and-carry with CERO : observations from a long-term user study with a service robot.* Proceedings 11th IEEE International Workshop on Robot and Human Interactive Communication, pages 158–163, 2002. 26

[Ishiguro 2001] Hiroshi Ishiguro, Tetsuo Ono, Michita Imai, Takeshi Maeda, Takayuki Kanda et Ryohei Nakatsu. *Robovie : an interactive humanoid robot.* Industrial Robot : An International Journal, pages 498–504, 2001. 27

[Ishii 2009] Kentaro Ishii, Shengdong Zhao, Masahiko Inami, Takeo Igarashi et Michita Imai. *Designing Laser Gesture Interface for Robot Control.* In Proceedings of the 12th IFIP Conference on Human-Computer Interaction, INTERACT2009, pages 479–492, 2009. 23, 24, 77, 140

[Iwahashi 2003] Naoto Iwahashi. *Language acquisition through a human-robot interface by combining speech, visual, and behavioral information.* Inf. Sci. Inf. Comput. Sci., vol. 156, pages 109–121, November 2003. 30, 31

[Iwahashi 2007] N. Iwahashi. Human-robot interaction, chapitre Robots That Learn Language : A Developmental Approach to Situated Human-Robot Conversations, pages 995–118. I-Tech Education and Publishing, 2007. 4, 31

[Johnson 2005] Mark Johnson. Developmental cognitive neuroscience. Blackwell publishing, 2nd édition, 2005. 8

[Juang 1991] B.H. Juang et L.R. Rabiner. *Hidden Markov models for speech recognition*. Technometrics, pages 251–272, 1991. 115

[Jurafsky 2000] D. Jurafsky, J.H. Martin, A. Kehler, K. Vander Linden et N. Ward. Speech and language processing : An introduction to natural language processing, computational linguistics, and speech recognition, volume 163. MIT Press, 2000. 115

[Kahn 2010] Peter H. Kahn Jr., Brian T. Gill, Aimee L. Reichert, Takayuki Kanda, Hiroshi Ishiguro et Jolina H. Ruckert. *Validating interaction patterns in HRI*. In Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction, HRI '10, pages 183–184, New York, NY, USA, 2010. ACM. 26

[Kanda 2009] Takayuki Kanda, Masahiro Shiomi, Zenta Miyashita, Hiroshi Ishiguro et Norihiro Hagita. *An affective guide robot in a shopping mall*. In Proceedings of the 4th ACM/IEEE international conference on Human robot interaction, HRI '09, pages 173–180, New York, NY, USA, 2009. ACM. 26

[Kanungo 2002] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman et Angela Y. Wu. *An Efficient k-Means Clustering Algorithm : Analysis and Implementation*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, pages 881–892, July 2002. 129

[Kaplan 2001] Frederic Kaplan, Pierre yves Oudeyer, Eniko Kubinyi et Adam Miklosi. *Taming robots with clicker training : A solution for teaching complex behaviors*. In in : Proceedings of the European Workshop on Learning Robots. Springer, 2001. 7

[Kaplan 2004] Frederic Kaplan et Verena Hafner. *The Challenges of Joint Attention*, 2004. 9, 32, 35

[Kaplan 2005] Frédéric Kaplan. Les machines apprivoisées comprendre les robots de loisir. vuibert, 2005. 34

[Kaplan 2008] Frederic Kaplan, Pierre-Yves Oudeyer et Benjamin Bergen. *Computational models in the debate over language learnability*. Infant and Child Development / formerly Early Development and Parenting, page n/a, 2008. 30

[Kaymaz 2003] Hande Kaymaz, Keskinpala Julie, A. Adams et Kazuhiko Kawamura. *PDA-Based Human-Robotic Interface*. In Proceedings of the IEEE International Conference on Systems, Man & Cybernetics : The Hague, Netherlands, 10-13 October 2004, 2003. 23

[Kazuhiro 2009] Nakadai Kazuhiro, Nakajima Hirofumi, Yuji Hasegawa et Hiroshi Tsujino. *Sound source separation of moving speakers for robot audition*. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '09, pages 3685–3688, Washington, DC, USA, 2009. IEEE Computer Society. 3

[Kemp 2008] Charles C. Kemp, Cressel D. Anderson, Hai Nguyen, Alexander J. Trevor et Zhe Xu. *A point-and-click interface for the real world : laser designation of objects for mobile manipulation.* In HRI '08 : Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction, pages 241–248, New York, NY, USA, 2008. ACM. 24, 77

[Khan 1998] Z. Khan. *Attitudes towards intelligent service robots.* NADA KTH, Stockholm, 1998. 17

[Kobayashi 1994] H. Kobayashi, F. Hara et A. Tange. *A basic study on dynamic control of facial expressions for Face Robot.* In Robot and Human Communication, 1994. RO-MAN '94 Nagoya, Proceedings., 3rd IEEE International Workshop on, pages 168 –173, jul 1994. 19

[Kozima 2009] H. Kozima, M.P. Michalowski et C. Nakagawa. *Keepon : A Playful Robot for Research, Therapy, and Entertainment.* International Journal of Social Robotics, vol. 1, no. 1, pages 3–18, 2009. 145

[Kupferberg 2011] Aleksandra Kupferberg, Markus Huber et Stefan Glasauer. New frontiers in human–robot interaction, pages 165–183. John Benjamins Publishing Company, 2011. 27

[Lauria 2002] S Lauria, G Bugmann, T Kyriacou et E Klein. *Mobile robot programming using natural language.* Robotics and Autonomous Systems, vol. 38, no. 3-4, pages 171–181, Mars 2002. 7, 21, 32, 116

[Lazebnik 2005] Svetlana Lazebnik, Cordelia Schmid et Jean Ponce. *A sparse texture representation using local affine regions.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, pages 1265–1278, 2005. 37

[Lazebnik 2009] Svetlana Lazebnik, Cordelia Schmid et Jean Ponce. *Spatial pyramid matching.* In Aleš Leonardis, Bernt Schiele, Sven J. Dickinson et Michael J. Tarr, editeurs, Object Categorization : Computer and Human Vision Perspectives, pages 401–415. Cambridge University Press, Novembre 2009. 37

[Li 2010] Dingjun Li, P. Rau et Ye Li. *A Cross-cultural Study : Effect of Robot Appearance and Task.* International Journal of Social Robotics, vol. 2, pages 175–186, 2010. 10.1007/s12369-010-0056-9. 5

[Lockerd 2004] A. Lockerd et C. Breazeal. *Tutelage and socially guided robot learning.* In Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on, volume 4, pages 3475 – 3480 vol.4, sept.-2 oct. 2004. 6, 7

[Lömker 2002] Frank Lömker et Gerhard Sagerer. *A Multimodal System for Object Learning.* In Proceedings of the 24th DAGM Symposium on Pattern Recognition, pages 490–497, London, UK, 2002. Springer-Verlag. 35

[Lowe 1999] David G. Lowe. *Object Recognition from Local Scale-Invariant Features.* In Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2, ICCV '99, pages 1150–, Washington, DC, USA, 1999. IEEE Computer Society. 3

[Lowe 2004] David G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints.* Int. Journal of Computer Vision, vol. 60, no. 2, pages 91–110, 2004. 37, 68

[Lungarella 2003] M Lungarella, G. Metta, R Pfeifer et G Sandini. *Developmental Robotics : A Survey.* Connection Science, vol. 15, no. 4, pages 151–190, 2003. 8

[Mataric 2001] Maja J Mataric. *Learning in behavior-based multi-robot systems : Policies, models, and other agents.* Cognitive Systems Research, vol. 2, pages 81–93, 2001. 16

[Matsumoto 2001] Y. Matsumoto, T. Ino et T. Ogasawara. *Development of Intelligent Wheelchair System with Face and Gaze Based Interface.* In Proc. of 10th IEEE Int. Workshop on Robot and Human Communication (ROMAN 2001), pages 262–267, 2001. 21

[Maulsby 1993] D. Maulsby, S. Greenberg et R. Mander. *Prototyping an intelligent agent through Wizard of Oz.* In Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems, pages 277–284. ACM, 1993. 79

[Mcguire 2002] P. Mcguire, J. Fritsch, J. J. Steil, F. Rothling, G. A. Fink, S. Wachsmuth, G. Sagerer et H. Ritter. *Multi-modal human-machine communication for instructing robot grasping tasks.* In In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS, pages 1082–1088, 2002. 21, 32

[Meerbeek 2009] B. Meerbeek, Martin Saerbeck et Christoph Bartneck. *Iterative design process for robots with personality.* In Kerstin Dautenhahn, editeur, AISB2009 Symposium on New Frontiers in Human-Robot Interaction, pages 94–101. SSAISB : The Society for the Study of Artificial Intelligence and the Simulation of Behaviour, 2009. 19

[Meger 2008] David Meger, Per E. Forssén, Kevin Lai, Scott Helmer, Sancho McCann, Tristram Southey, Matthew Baumann, James J. Little et David G. Lowe. *Curious George : An attentive semantic robot.* Robot. Auton. Syst., vol. 56, no. 6, pages 503–511, Juin 2008. 3

[Michael 2005] David R. Michael et Sandra L. Chen. Serious games : Games that educate, train, and inform. Muska & Lipman/Premier-Trade, 2005. 89

[Mikolajczyk 2003] Krystian Mikolajczyk et Cordelia Schmid. *A performance evaluation of local descriptors.* In International Conference on Computer Vision & Pattern Recognition, volume 2, pages 257–263, 2003. 37, 68, 134

[Miller 2001] P.H. Miller. Theories of developmental psychology. New York : Worth, 4th édition, 2001. 8

[Miyashita 2005] Takahiro Miyashita, Taichi Tajika, Hiroshi Ishiguro, Kiyoshi Kogure et Norihiro Hagita. *Haptic Communication Between Humans and Robots.* In ISRR'05, pages 525–536, 2005. 3

[Morimoto 2005] Carlos H. Morimoto et Marcio R. M. Mimica. *Eye gaze tracking techniques for interactive applications.* Comput. Vis. Image Underst., vol. 98, pages 4–24, April 2005. 21

[Nagai 2003] Yukie Nagai, Koh Hosoda Y, Akio Morita et Minoru Asada Y. *A constructive model for the development of joint attention.* Connection Science, vol. 15, pages 211–229, 2003. 9

[Nehaniv 2004] Chrystopher L. Nehaniv et Kerstin Dautenhahn, editeurs. Imitation and social learning in robots, humans, and animals : behavioural, social and communicative dimensions. Cambridge University Press, 2004. 6

[Newman 2000] Rhys Newman, Yoshio Matsumoto, Sebastien Rougeaux et Alexander Zelinsky. *Real-Time Stereo Tracking for Head Pose and Gaze Estimation.* In Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000, FG '00, pages 122–, Washington, DC, USA, 2000. IEEE Computer Society. 21

[Nguyen 2011] Sao Mai Nguyen, Adrien Baranes et Pierre-Yves Oudeyer. *Bootstrapping Intrinsically Motivated Learning with Human Demonstrations.* In proceedings of the IEEE International Conference on Development and Learning, page Nguyen, Frankfurt, Allemagne, 2011. ERC Grant EXPLORERS 240007. 7, 16

[Nickel 2004] Kai Nickel et Rainer Stiefelhagen. *Real-time Recognition of 3D-Pointing Gestures for Human-Machine-Interaction.* In International Workshop on Human-Computer Interaction HCI 2004, May 2004, Prague (in conjunction with ECCV 2004), 2004. 3, 20

[Niehaus 2012] L. Niehaus et S.E. Levinson. *Online Learning of Word and Action Lexicons for Grounding Language : Experiments with the iCub Humanoid Robot.* In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2012. 30

[Nielsen 1994] Jakob Nielsen. *Usability inspection methods.* In Conference companion on Human factors in computing systems, CHI '94, pages 413–414, New York, NY, USA, 1994. ACM. 48, 137

[Nomura 2006] Tatsuya Nomura, Takayuki Kanda et Tomohiro Suzuki. *Experimental investigation into influence of negative attitudes toward robots on human-robot interaction.* AI Soc., vol. 20, pages 138–150, February 2006. 18

[Nourbakhsh 1999] Illah R. Nourbakhsh, Alvaro Soto, Judith Bobenage, Sebastien Grange, Roland Meyer et Ron Lutz. *An effective mobile robot educator with a full-time job.* Artif. Intell., vol. 114, pages 95–124, October 1999. 19

[Oudeyer 2003] Pierre-Yves Oudeyer. *The production and recognition of emotions in speech : features and algorithms.* International Journal of Human-Computer Studies, vol. 59, pages 157–183, 2003. 32

[Oudeyer 2006] Pierre-Yves Oudeyer et Frédéric Kaplan. *Discovering communication.* Connection Science, vol. 18, no. 2, pages 189–206, Juin 2006. 31, 34

[Oudeyer 2007] Pierre-Yves Oudeyer, Frederic Kaplan et Verena Hafner. *Intrinsic motivation systems for autonomous mental development.* IEEE Transactions on Evolutionary Computation, vol. 11, pages 265–286, 2007. 6

[Park 2008] Alex S Park et James R Glass. *Unsupervised Pattern Discovery in Speech.* Language, vol. 16, no. 1, pages 186–197, 2008. 30

[Perzanowski 2001] Dennis Perzanowski, Alan C. Schultz, William Adams, Elaine Marsh et Magda Bugajska. *Building a Multimodal Human-Robot Interface.* IEEE Intelligent Systems, vol. 16, no. 1, pages 16–21, 2001. 21, 25, 32, 116

[Piaget 1945] J. Piaget. Play, Dreams and Imitation in Childhood. Norton, New York, 1945. 6

[Pontil 1998] Massimiliano Pontil et Alessandro Verri. *Support Vector Machines for 3D Object Recognition.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, pages 637–646, June 1998. 37

[Pylyshyn 2001] Z.W. Pylyshyn. *Visual indexes, preconceptual objects, and situated vision.* Cognition, vol. 80, no. 1-2, pages 127–158, 2001. 85

[Quine 1960] Willard Van Orman Quine. Word and Object, volume 22. MIT Press, 1960. 31, 36

[Rabiner 1986] L. Rabiner et B. Juang. *An introduction to hidden Markov models.* ASSP Magazine, IEEE, vol. 3, no. 1, pages 4–16, 1986. 115

[Rahimi 2009] Ali Rahimi, Joshua R. Smith, David I. Ferguson et Siddhartha S. Srinivasa. *Personal Robots : A Personal Computer Industry Perspective*, 2009. 2

[Raibert 2008] Marc Raibert. *BigDog, the Rough-Terrain Quadruped Robot.* In Myung J. Chung, editeur, Proceedings of the 17th IFAC World Congress, volume 17, 2008. 3

[Revel 2004] A Revel et J Nadel. Imitation and social learning in robots, humans and animals : Behavioural, social and communicative dimensions, chapitre How to build an imitator ? Cambridge University Press, 2004. 7

[Reynolds 1994] D.A. Reynolds. *Experimental evaluation of features for robust speaker identification.* Speech and Audio Processing, IEEE Transactions on, vol. 2, no. 4, pages 639–643, 1994. 118

[Robbel 2007] Phillip Robbel. *Exploiting object dynamics for recognition and control.* PhD thesis, Massachusetts Institute of Technology. Dept. of Architecture. Program in Media Arts and Sciences., 2007. 107

[Rouanet 2009a] Pierre Rouanet, Jérome Béchu et Pierre-Yves Oudeyer. *A comparison of three interfaces using handheld devices to intuitively drive and show objects to a social robot : the impact of underlying metaphors.* IEEE International Symposium on Robots and Human Interactive Communications RO-MAN, 2009. 52

[Rouanet 2009b] Pierre Rouanet et Pierre-Yves Oudeyer. *Exploring the use of a handheld device in language teaching human-robot interaction.* In Proceedings of the AISB 2009 Workshop : New Frontiers in Human-Robot Interaction, 2009. 65, 114, 115

[Rouanet 2009c] Pierre Rouanet, Pierre-Yves Oudeyer et David Filliat. *An integrated system for teaching new visually grounded words to a robot for non-expert users using a mobile device.* In Proceedings of the IEEE-RAS Humanoids 2009 Conference, 2009. 65, 108, 114, 134

[Rouanet 2010a] Pierre Rouanet, Pierre-Yves Oudeyer et David Filliat. *A study of three interfaces allowing non-expert users to teach new visual objects to a robot and their impact on learning efficiency.* In Proceedings of the ACM/IEEE Human-Robot Interaction HRI 2010 Conference, 2010. 65, 115, 134, 135, 146

[Rouanet 2010b] Pierre Rouanet, Pierre-Yves Oudeyer et David Filliat. *Using mediator objects to easily and robustly teach visual objects to a robot.* In ACM SIGGRAPH 2010 Posters. ACM, 2010. 65, 134

[Rouanet 2011] Pierre Rouanet, Fabien Danieau et Pierre-Yves Oudeyer. *A Robotic Game to Evaluate Interfaces used to Show and Teach Visual Objects to a Robot in Real World Condition.* In proceedings of the ACM/IEEE Human-Robot Interaction HRI 2011 conference, 2011. 89

[Roy 1999] Deb Kumar Roy. *Learning words from sights and sounds : a computational model.* PhD thesis, Architecture and Planning, 1999. 30

[Roy 2000] Deb Roy et Alex Pentland. *Learning Words from Sights and Sounds : A Computational Model.* Cognitive Science, vol. 26, pages 113–146, 2000. 31

[Roy 2003] Deb Roy. *Grounded spoken language acquisition : experiments in word learning.* IEEE Transactions on Multimedia, vol. 5, no. 2, pages 197–209, 2003. 21, 32, 33

[Rumbaugh 1996] Sue S. Rumbaugh et Roger Lewin. Kanzi : The ape at the brink of the human mind. Wiley, September 1996. 12, 35

[Russell 1997] J. A. Russell. *Reading emotions from and into faces : Resurrecting a dimensional-contextual perspective.* In J. A. Russell et J. M. Fernández-Dols, editeurs, The Psychology of Facial Expression, pages 295–320. Cambridge : Cambridge University Press, 1997. 19

[Sakamoto 2009] Daisuke Sakamoto, Koichiro Honda, Masahiko Inami et Takeo Igarashi. *Sketch and run : a stroke-based interface for home robots.* In CHI '09 : Proceedings of the 27th international conference on Human factors in computing systems, pages 197–200, New York, NY, USA, 2009. ACM. 24

[Sakoe 1978] Hiroaki Sakoe. *Dynamic programming algorithm optimization for spoken word recognition.* IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, pages 43–49, 1978. 115, 118

[Saxena 2008] Ashutosh Saxena, Justin Driemeyer et Andrew Y. Ng. *Robotic Grasping of Novel Objects using Vision.* Int. J. Rob. Res., vol. 27, pages 157–173, February 2008. 143

[Scassellati 1996] Brian Scassellati. *Mechanisms of Shared Attention for a Humanoid Robot.* In Embodied Cognition and Action : Papers from the 1996 AAAI Fall Symposium, 1996. 4, 9, 21, 32, 33

[Scassellati 2001] Brian Michael Scassellati. *Foundations for a theory of mind for a humanoid robot.* PhD thesis, 2001. AAI0803464. 4, 10, 17, 18, 27

[Scheeff 2000] M. Scheeff, J. Pinto, K. Rahardja, S. Snibbe et R. Tow. *Experiences with Sparky : A social robot.* In Proceedings of the Workshop on Interactive Robot Entertainment, 2000. 18

[Schiele 1996] Bernt Schiele et James L Crowley. *Object recognition using multidimensional receptive field histograms.* Energy, vol. 1064, no. section 6, pages 610–619, 1996. 37

[Schmalstieg 1999] Dieter Schmalstieg, L. Miguel Encarnação et Zsolt Szalavári. *Using transparent props for interaction with the virtual table.* In I3D '99 : Proceedings of the 1999 symposium on Interactive 3D graphics, pages 147–153, New York, NY, USA, 1999. ACM. 74

[Schmuedderich 2010] J. Schmuedderich, N. Einecke, S. Hasler, A. Gepperth, B. Bolder, R. Kastner, M. Franzius, S. Rebhan, B. Dittes, H. Wersing, J. Eggert, J. Fritsch et C. Goerick. *System approach for multi-purpose representations of traffic scene elements.* pages 1677–1684, 09 2010. 37

[Schulte 1999] Jamieson Schulte, Chuck Rosenberg et Sebastian Thrun. *Spontaneous Short-term Interaction with Mobile Robots in Public Places.* 1999. 19

[Sivic 2003] J. Sivic et A. Zisserman. *Video Google : A Text Retrieval Approach to Object Matching in Videos.* In IEEE International Conference on Computer Vision (ICCV), 2003. 37, 67, 71

[Skubic 2002] Marjorie Skubic, Sam Blisard, Andy Carle et Pascal Matsakis. *Hand-Drawn Maps for Robot Navigation.* In AAAI Spring Symposium, Sketch Understanding Session, March, 2002., 2002. 23

[Sofge 2004] Donald Sofge, J. Gregory Trafton, Nicholas Cassimatis, Dennis Perzanowski, Magdalena Bugajska, William Adams et Alan Schultz. *Human-Robot Collaboration and Cognition with an Autonomous Mobile Robot.* In Proceedings of the 8th Conference on Intelligent Autonomous Systems (IAS-8, pages 80–87. IOS Press, 2004. 10

[Staudte 2009] Maria Staudte et Matthew W. Crocker. *Visual attention in spoken human-robot interaction.* In Proceedings of the 4th ACM/IEEE international conference on Human robot interaction, HRI '09, pages 77–84, New York, NY, USA, 2009. ACM. 21, 33

[Steels 2000] Luc Steels et Frederic Kaplan. *AIBO's first words : The social learning of language and meaning.* Evolution of Communication, vol. 4, no. 1, pages 3–32, 2000. 4, 31, 32

[Steels 2002] L. Steels et F. Kaplan. *Bootstrapping grounded word semantics.* In Ted Briscoe, editeur, Linguistic Evolution through Language Acquisition :

Formal and Computational Models, chapitre 3. Cambridge University Press, 2002. 30, 31

[Steinfeld 2006] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz et Michael Goodrich. *Common metrics for human-robot interaction.* In Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction, HRI '06, pages 33–40, New York, NY, USA, 2006. ACM. 11, 26

[Sutton 1998] Richard S. Sutton et Andrew G. Barto. Reinforcement Learning : An Introduction (Adaptive Computation and Machine Learning). The MIT Press, Mars 1998. 6

[Tapus 2007] Adriana Tapus, Maja Mataric et Brian Scassellati. *Socially assistive robotics [Grand Challenges of Robotics].* IEEE Robotics Automation Magazine, vol. 14, no. 1, pages 35–42, 2007. 5, 17, 26

[ten Bosch 2008] Louis ten Bosch, Hugo Van hamme et Lou Boves. *Unsupervised detection of words-questioning the relevance of segmentation.* In ISCA ITRW, Speech Analysis and Processing for Knowledge Discovery, pages 4–7. Citeseer, 2008. 30

[Thomaz 2007] A.L. Thomaz et C. Breazeal. *Robot learning via socially guided exploration.* In Development and Learning, 2007. ICDL 2007. IEEE 6th International Conference on, pages 82 –87, july 2007. 6, 7, 16

[Thomaz 2008] Andrea L. Thomaz et Cynthia Breazeal. *Teachable robots : Understanding human teaching behavior to build more effective robot learners.* Artificial Intelligence Journal, vol. 172, pages 716–737, 2008. 6, 7, 8, 16

[Tomasello 1995] Michael Tomasello. Joint attention as social cognition, volume 16, pages 103–130. Erlbaum, 1995. 8, 11, 12, 20, 32

[Tomasello 1999] Michael Tomasello. The cultural origins of human cognition. Harvard University Press, 1999. 11, 30

[Tomasello 2004] Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne et Henrike Moll. *Understanding and sharing intentions : The origins of cultural cognition.* Behavioral and Brain Sciences, vol. In Press, 2004. 8, 12

[Tsui 2008] Katherine Tsui, Holly Yanco, David Kontak et Linda Beliveau. *Development and evaluation of a flexible interface for a wheelchair mounted robotic arm.* In HRI '08 : Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction, pages 105–112, New York, NY, USA, 2008. ACM. 24, 142

[Vogt 1998] Paul Vogt. *The evolution of a lexicon and meaning in robotic agents through self-organization.* In Proceedings of the NetherlandsBelgium Conference on Artificial Intelligence Amsterdam CWI Amsterdam. Citeseer, 1998. 31

[Wada 2005] K. Wada, T. Shibata, T. Saito, K. Sakamoto et K. Tanie. *Psychological and Social Effects of One Year Robot Assisted Activity on Elderly People at a*

Health Service Facility for the Aged. Proceedings of the 2005 IEEE International Conference on Robotics and Automation, no. April, pages 2785–2790, 2005. 27, 28

[Walters 2005] M. L. Walters, S. N. Woods, K. L. Koay et K. Dautenhahn. *Practical and Methodological Challenges in Designing and Conducting Human-Robot Interaction Studies.* In Proceedings of the AISB'05 Symposium on Robot Companions Hard Problems and Open Challenges in Human-Robot Interaction, pages 110–119. University of Hertfordshire, University of Hertfordshire, April 2005. 4, 11, 26, 48, 79

[Walters 2007] Michael L. Walters, Kerstin Dautenhahn, Sarah N. Woods et Kheng Lee Koay. *Robot etiquette : Results from user studies involving a fetch and carry task.* In in ACM/IEEE International Conference on Human-Robot Interaction, pages 317–324, 2007. 19

[Walters 2008] Michael L. Walters, Dag S. Syrdal, Kerstin Dautenhahn, René Te Boekhorst et Kheng Lee Koay. *Avoiding the uncanny valley : robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion.* Auton. Robots, vol. 24, pages 159–178, February 2008. 5

[Wang 2005] Junqiu Wang, R. Cipolla et Hongbin Zha. *Vision-based Global Localization Using a Visual Vocabulary.* In Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA), 2005. 67

[Weiss 2009] Astrid Weiss, Regina Bernhaupt, Michael Lankes et Manfred Tscheligi. *The usus evaluation framework for human-robot interaction.* In AISB2009 : Proceedings of the Symposium on New Frontiers in Human-Robot Interaction, 2009. 4, 11, 26, 49

[Weng 2001] J Weng, J McClelland, A Pentland, O Sporns, I Stockman, M Sur et Esther Thelen. *Autonomous mental development by robots and animals.* Science, vol. 291, pages 599–600, 2001. 8

[Wersing 2006] Heiko Wersing, Stephan Kirstein, Michael Götting, Holger Brandl, Mark Dunn, Inna Mikhailova, Christian Goerick, Jochen J. Steil, Helge Ritter et Edgar Körner. *A Biologically Motivated System for Unconstrained Online Learning of Visual Objects.* In ICANN (2), pages 508–517, 2006. 35

[Woods 2005] Sarah Woods, Kerstin Dautenhahn, Christina Kaouri et René te Boekhorst Kheng Lee Koay. *Is this robot like me ? Links between human and robot personality traits.* In 5th IEEE-RAS International Conference on Humanoid Robots, pages 375–380, 2005. 18

[Wu 1999] Ying Wu et Thomas S. Huang. *Vision-Based Gesture Recognition : A Review.* In Proceedings of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction, GW '99, pages 103–115, London, UK, 1999. Springer-Verlag. 20

[Xin 2007] Min Xin et Ehud Sharlin. *Playing Games with Robots - A Method for Evaluating Human-Robot Interaction.* Education, 2007. 145

[Yanco 1993] H. Yanco et L.A. Stein. *An adaptive communication protocol for cooperating mobile robots.* From animals to animats, vol. 2, pages 478–485, 1993. 31

[Yanco 2004a] Holly A. Yanco. *where Am I ? Acquiring Situation Awareness Using a Remote Robot Platform.* In In IEEE Conference on Systems, Man and Cybernetics, pages 2835–2840, 2004. 143

[Yanco 2004b] Holly A. Yanco, Jill L. Drury et Jean Scholtz. *Beyond usability evaluation : analysis of human-robot interaction at a major robotics competition.* Hum.-Comput. Interact., vol. 19, no. 1, pages 117–149, 2004. 48

[Yu 2004a] Chen Yu et Dana H. Ballard. *A Multimodal Learning Interface for Grounding Spoken Language in Sensory Perceptions.* ACM Transactions On Applied Perception, vol. 1, pages 57–80, 2004. 30, 31

[Yu 2004b] Chen Yu et Dana H. Ballard. *On the integration of grounding language and learning objects.* In AAAI'04 : Proceedings of the 19th national conference on Artifical intelligence, pages 488–493. AAAI Press / The MIT Press, 2004. 31

[Zheng 2001] Fang Zheng, Guoliang Zhang et Zhanjiang Song. *Comparison of different implementations of MFCC.* Journal of Computer Science and Technology, vol. 16, pages 582–589, 2001. 10.1007/BF02943243. 115, 118

Appendices

Annexe A

Liens vidéos

Présentation de notre démarche générale

http://pier.rouanet.free.fr/videos/interfaces/general.m4v

Présentation des problématiques liées aux interactions directes et au waving

http://pier.rouanet.free.fr/videos/interfaces/gestures.mov
http://pier.rouanet.free.fr/videos/interfaces/waving.mov

Présentation de nos différentes interfaces basées sur des objets médiateurs

http://pier.rouanet.free.fr/videos/interfaces/mediateurs.mov

Présentation de notre installation expérimentale dans le musée Cap Sciences

http://pier.rouanet.free.fr/videos/interfaces/capsciences.mov

Annexe B

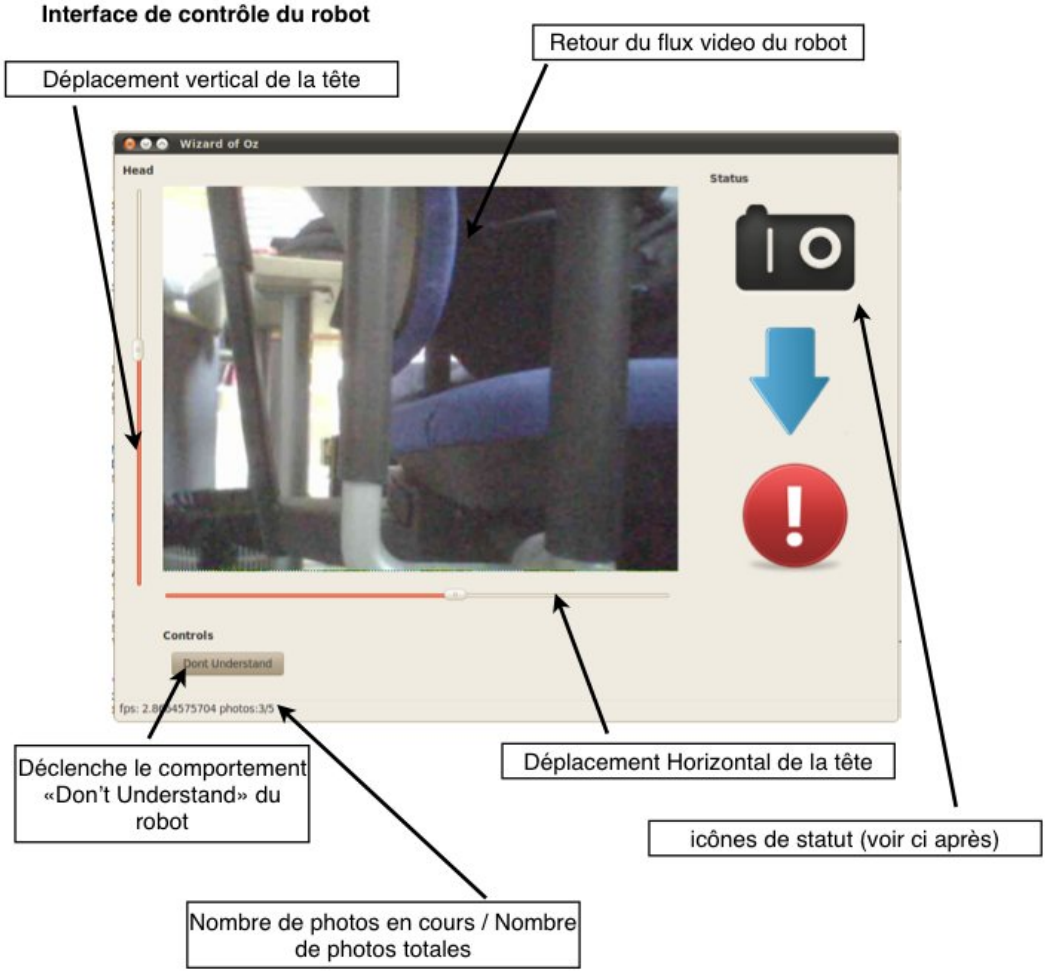
Instructions fournies aux magiciens pour l'interprétation de l'interface gestuelle

Protocole du Magicien d'Oz

Objectif: Interpréter des gestes uniquement via la camera du robot et le piloter en fonction.

Description de l'expérience: un jeu robotique est proposé à l'utilisateur. Il doit guider un robot pour lui montrer certains objets. La première partie du jeu est une phase de tutoriel où l'utilisateur se familiarise avec le robot. Il va apprendre que le robot peut se déplacer et regarder autour de lui. Pour finir il mettra le robot en face de lui pour qu'il voie son visage, puis il lui caressera la tête pour prendre la première photo.
Ensuite l'utilisateur devra montrer 4 objets au robot. Pour chacun il devra faire en sorte que le robot voie l'objet, puis il devra lui caresser le tête pour qu'il le prenne en photo.
Le tutoriel dure environ 5 min. Le reste du jeu dure environ 10 minutes.

Interface de contrôle du robot



Déplacement du robot

le déplacement du robot se fait via les touches directionnelles du clavier:
- haut = avancer
- bas = reculer
- gauche = rotation sur la gauche
- droite = rotation sur la droite

le mouvement se fait tant que la touche est maintenant enfoncée.

Icônes de statut

Le jeu vient de commencer. Le robot est allongé sur le sol, l'utilisateur doit lui caresser le tête.
Consigne : attendre que l'icône disparaisse.

Le robot est en train de prendre une photo.
Consigne : attendre que l'icône disparaisse.

Le robot effectue le mouvement «don't understand».
Consigne: attendre que l'icône disparaisse.

Une touche directionnelle du clavier est pressée, le robot se déplace.

Le robot est allongé ou ses pieds ne touchent pas le sol.
Consigne: attendre que l'icône disparaisse.

Annexe C

Formulaire de consentement

Formulaire de Consentement libre et de Renonciation au droit à l'image

Etude sur les jeux robotiques

Equipe FLOWERS
INRIA Bordeaux Sud-Ouest, Bat. A29
351 Cours de la Libération, 33405 Talence, Cedex, France

Je certifie avoir donné mon accord pour participer à une étude sur les jeux robotiques. J'accepte volontairement de participer à cette étude et je comprends que ma participation n'est pas obligatoire et que je peux stopper ma participation à tout moment sans avoir à me justifier ni encourir aucune responsabilité. Mon consentement ne décharge pas les organisateurs de la recherche de leurs responsabilités et je conserve tous mes droits garantis par la loi.

Au cours de cette expérience, j'accepte que soient recueillies des données chronométriques et un enregistrement audio de ma voix. Je comprends que les informations recueillies sont strictement confidentielles et à usage exclusif des investigateurs concernés. J'accepte que les données enregistrées à l'occasion de cette étude puissent être conservées dans une base de données et faire l'objet d'un traitement informatisé non nominatif par l'INRIA.

J'accorde également à l'INRIA la permission irrévocable de publier toutes les photographies ou vidéos que vous avez prises de moi. Ces images peuvent être exploitées dans le cadre d'articles scientifiques ou de publications, d'espaces Internet servant de présentation ou de promotion des activités de l'INRIA. Je m'engage à ne pas tenir responsable l'INRIA, l'expérimentateur cité ci-dessous ainsi que ses représentants et toute personne agissant avec sa permission en ce qui relève de la possibilité d'un changement de cadrage, de couleur et de densité qui pourrait survenir lors de la reproduction.

Etabli en double exemplaire, le ..... à .....

Nom du volontaire (et de son tuteur légal le cas échéant): .....

Signature du volontaire (ou de son tuteur légal le cas échéant) précédée de la mention « lu et approuvé » :

Nom de l'expérimentateur : .....

Signature de l'expérimentateur:

Annexe D

Questionnaires

Pre Questionnaire

Critère		Question
Infos	Sexe	Masculin / Féminin
	Age	Nombre
Background	Ordinateur	A quelle fréquence utilisez-vous un ordinateur?
	Robotique	Estes-vous familier avec la robotique ?
	Jeu Video	A quelle fréquence jouez-vous aux jeux vidéos?
	Wiimote	Avez-vous déjà utilisé une wiimote?
	Iphone / Tactile	Avez-vous déjà utilisé un appareil à interface tactile (type iphone) ?
Utilisabilité	Efficience	Je pense qu'interagir avec un robot est compliqué *
Experience Utilisateur	Emotion	Je suis enthousiaste à l'idée d'interagir avec un robot
		Je me sens à l'aise à côté du robot
	Interaction Sociale	Je n'envisage pas les robots comme des créatures sociales *
		Je pense que le robot peut être un partenaire de jeu

Post Questionnaire

Critère		Question
Utilisabilité	Efficience	Demande au robot de se déplacer était facile
		Réussir le jeu était facile
		J'ai facilement interagi avec le robot
		Prendre une photo avec le robot était simple
		Je me suis senti(e) confiant lorsque j'interagissais avec le robot
	Erreur	Le robot répondait correctement aux directives

Critère		Question
	Apprentissage	Les comportements du robot étaient peu compréhensibles *
		Comprendre comment interagir avec le robot m'a demandée beaucoup d'efforts *
		Les consignes données par le robot lors du jeu m'ont permis de bien prendre en main l'interface
Expérience Utilisateur	Emotion	Je pense que j'aimerais rejouer avec le robot
		J'ai aimé interagir avec le robot
		Je n'ai pas aimé les réactions du robot *
	Perception de l'animation	J'ai perçu le robot comme "vivant"
		Je pense que le robot n'a pas compris ce que je lui demandais *
	Interaction Sociale	J'ai l'impression d'avoir fait équipe avec le robot pour réussir le jeu
Jeu	Challenge	Le jeu m'a intéressé jusqu'à la fin
	Fun	Le jeu était amusant
	Réactivité	Le robot mettait beaucoup de temps à réagir *
	Immersion	Je n'ai pas vu le temps passer durant le jeu
		Les réactions du robot n'apportaient rien au jeu *
		Le scénario était intéressant
		La phase d'entrainement m'a permis d'être impliqué(e) dès le début du jeu
		Le jeu aurait été aussi intéressant sans le robot