Perception multimodale et interaction sociable

Dominique Vaufreydaz

Communauté UNIVERSITÉ Grenoble Alpes

HABILITATION À DIRIGER DES RECHERCHES

Spécialité : **Informatique et Mathématiques**

Présentée par : **Dominique Vaufreydaz**

Préparée au sein du **Laboratoire d'Informatique de Grenoble**
de **l'Inria**
et de **MSTII**

PERCEPTION MULTIMODALE ET INTERACTION SOCIABLE

Habilitation soutenue publiquement le 24 juillet 2018, devant le jury composé de :

Charpillet François

Directeur de Recherche, Inria (Nancy, France), Rapporteur

Morency Louis-Philippe

Professeur Associé, Carnegie Mellon University, (Pittsburgh, États-Unis), Rapporteur

Vinciarelli Allessandro

Professeur à l'Université de Glasgow (Écosse, Royaume-Uni), Rapporteur

Pelachaud Catherine

Directrice de Recherche, ISIR (Paris, France), Examinatrice

Chateau Thierry

Professeur à l'Université Blaise Pascal (Clermont-Ferrand, France), Examinateur

Crowley James

Professeur à Grenoble INP (France), Examinateur

Remerciements

Pour cette habilitation à diriger les recherches, je tenais tout d'abord à remercier François Charpillet, Louis-Philippe Morency et Allessandro Vinciarelli pour avoir accepté d'en être rapporteurs. Je remercie également Catherine Pelachaud et Thierry Chateau pour avoir accepté d'être examinateurs. Je remercie Laurence Boissieux, Éric Castelli et Philippe Dessus pour leur relecture attentive de ce manuscrit.

Je remercie tous mes collègues (vacataires, moniteurs/DCE, ATER, enseignants-chercheurs, chercheurs, ingénieurs, administratifs, …) grenoblois, français ou étrangers, mais aussi les étudiants et doctorants avec lesquels j'ai échangé de manière si constructive au cours de ces années. Vous avez su faire de nos rencontres professionnelles, de nos collaborations, et parfois de nos amitiés, un environnement de travail où ont pu s'épanouir mes recherches. Pour être certain de n'oublier personne, je ne mentionnerai aucun nom, la liste étant bien trop longue et ma mémoire des noms trop capricieuse. Je tiens simplement à ce qu'ils sachent que, sans eux, rien de ce qui est présenté dans cette habilitation n'aurait été possible.

Pour la période couverte par ce manuscrit, je remercie spécialement James L. Crowley pour son accueil au sein de l'équipe PRIMA en 2002, pour l'environnement toujours propice qu'il a su mettre en place autour de mes recherches, et pour m'avoir fait confiance au cours de ces années.

Pour terminer, je remercie Sonia, Célia et Esteban, ma famille, pour leur soutien inconditionnel même si mon poste de Maître de conférences me vole parfois trop de temps.

Dédicace

Cette habilitation à diriger les recherches est dédiée à Sonia, Célia et Esteban.

Résumé

L'une des tâches les plus complexes pour laquelle les ordinateurs ont été programmés concerne le mimétisme des capacités de perception et d'interaction des humains en utilisant tout d'abord des informations monomodales (acoustiques, visuelles, tactiles, de proprioception, …) puis multimodales en combinant plusieurs modalités. À partir de ces capacités de perception, les systèmes interactifs, c'est-à-dire les systèmes interagissant avec des humains, peuvent être sensibles à l'environnement qui les entoure, aux utilisateurs présents, à la situation courante… Cela leur permet de percevoir, comprendre et prédire pour agir en conséquence, voire d'agir d'une manière sociable pour être un partenaire des humains à part entière.

La perception multimodale par ordinateur et les interactions sociables sont les problématiques de fond de mes travaux depuis mon recrutement en tant que Maître de conférences en 2005, le traitement du signal (« *signal processing* ») et l'apprentissage automatique (« *machine learning* ») en étant les fondements. Ce manuscrit présente mes travaux sur la perception multimodale et les interactions sociables dans plusieurs contextes en les regroupant autour de mes thématiques de recherche principales.

Ce manuscrit aborde tout d'abord la perception multimodale ubiquitaire au sein d'espaces perceptifs multimodaux tels les salles de réunions augmentées, les appartements équipés pour le maintien de personnes âgées/fragiles à domicile ou des espaces à plus grande échelle comme les bâtiments d'un campus universitaire. Faisant suite aux progrès en robotique, cette perception s'est naturellement déplacée des environnements perceptifs vers les robots mobiles, permettant des interactions sociables entre les humains et des robots compagnons (*Human Robot Interaction - HRI*) mais aussi avec des robots particuliers que sont les véhicules autonomes. Les travaux de recherche concernant la perception des humains et de leurs affects sont ensuite présentés *via* mes recherches sur la perception en champ proche (< 1 m) et sur la détection des humains et de leurs comportements autour de nos systèmes interactifs, base nécessaire à leur fonctionnement. Nos travaux préliminaires sur la détection de personnes en utilisant de l'apprentissage profond (« *Deep Learning* ») sont décrits. Ce manuscrit se clôt en présentant les directions et les perspectives de mon projet de recherche intitulé « Perception multimodale et interaction sociable ».

Mots clés : traitement du signal, vision par ordinateur, perception multimodale, apprentissage machine, interaction sociable.

Table des matières

Chapitre I
Introduction

I.A Contexte

En introduction de mon manuscrit de doctorat [1], avec le préambule ci-dessous, je mettais en perspective la vision de la science-fiction et l'état de l'art en reconnaissance de la parole à l'époque :

- bonsoir Dave.

- ça va HAL ?

- tout est en ordre de marche, et vous ça va ?

- pas trop mal.

- je vois que vous avez travaillé...

- quelques croquis...

- je peux les voir ?

- bien sûr.

- c'est très bien rendu Dave... je crois que vous avez fait beaucoup de progrès... un peu plus près je vous prie.

- bien sûr [...]

Cet extrait de dialogue est tiré de « *2001, l'odyssée de l'espace* », l'adaptation cinématographique du premier tome de la célèbre tétralogie de *Arthur C. Clarke* réalisée par *Stanley Kubrick*. Voilà l'avenir tel que l'auteur du livre le dépeignait dans les années soixante. Un ordinateur capable de comprendre le langage oral, d'interagir directement avec les humains en utilisant la parole comme modalité d'entrée et de sortie. Cela allait même bien plus loin puisque celui-ci était finalement animé de capacités supérieures qui lui permettaient d'apprendre la lecture labiale afin d'espionner ses passagers et d'assouvir ses pulsions paranoïaques. C'est ainsi qu'était *HAL*, un ordinateur personnifié comme une tierce personne dans un futur qui est aujourd'hui notre présent. *Clarke* n'était bien entendu pas seul à avoir cette vision de l'avenir. Nombre d'auteurs de science-fiction avaient prédit que l'ordinateur prendrait une place prépondérante dans la société et que son intégration, dans la vie quotidienne, serait complète.

Qu'en est-il 15 ans plus tard ? *Google* vient de présenter *Google Duplex*, une intelligence artificielle de reconnaissance, d'analyse linguistique et de synthèse de la parole que ses performances, notamment en synthèse, rendent difficile à distinguer d'un humain. Ce système est un candidat à la réussite du test de *Turing*[1]. Ce test, élaboré par *Alan Turing* en 1950 [2], vise à mesurer l'intelligence d'une machine en lui proposant d'imiter un humain lors d'un dialogue avec un juge humain. Si la machine ne peut être distinguée d'un être humain dans ce jeu d'imitation, elle est considérée comme intelligente. *Google Duplex* est-elle intelligente ? *Turing* cite dans son article [2] le Professeur *G. Jefferson* qui exprimait en 1949 :

> *« Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain-that is, not only write it but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants. »*

Cette citation est un condensé de ce qui fait l'intelligence humaine : être conscient de soi, apprendre et inventer dans de nombreux domaines, tout ceci teinté d'émotions ressenties et/ou exprimées. Les systèmes informatiques les plus performants de nos jours ne sont donc pas intelligents en ce sens que l'on nomme l'intelligence artificielle forte[2]. Ils appartiennent à ce que l'on qualifie d'intelligence artificielle faible : une simulation mathématique de l'intelligence (humaine).

Depuis son émergence, l'informatique a été utilisée dans ce but. Non pas pour supplanter les hommes mais pour permettre aux machines de les remplacer dans certaines tâches laborieuses. Au cours du temps, ce mimétisme des machines s'est accentué et s'est dirigé vers le mimétisme du vivant avec notamment de nombreux travaux en perception par ordinateur, l'une de nos préoccupations dans ce manuscrit.

Au cours de la dernière décennie, la recherche scientifique est en train de vivre ce que d'aucuns qualifient de révolution. L'apprentissage profond (« *Deep Learning* ») [3], un retour au premier plan des réseaux de neurones, a permis de nombreux progrès en perception, notamment en vision par ordinateur et en reconnaissance de parole. *Google Duplex* en est une illustration. Dans la vie courante, de nombreux algorithmes emploient des réseaux de neurones profonds sans que les utilisateurs n'en aient conscience. Les plus utilisés étant certainement les détecteurs/identificateurs de visages [4] sur les réseaux sociaux ou la reconnaissance de la parole des téléphones portables ou des assistants vocaux connectés [5].

---

[1] Voir http://www.turingarchive.org/browse.php/B/9 (dernière visite 04/2018).

[2] Voir https://fr.wikipedia.org/wiki/Intelligence_artificielle pour plus d'informations sur les différences entre IA forte et IA faible (*strong AI/weak AI* en anglais). Dernière visite 04/2018.

Cette révolution, ainsi que les progrès antérieurs depuis le début des années 2000, ont été soutenus par des avancées technologiques fortes. L'augmentation de la puissance de calcul des ordinateurs, leur miniaturisation, l'amélioration des réseaux informatiques et l'arrivée de capteurs performants à coût raisonnable (webcam, capteurs de profondeur, etc.) ont permis ces avancées. Les progrès en robotique, et pour ceux qui nous concernent directement les progrès de la robotique mobile [6], ont apporté de nouveaux cadres applicatifs mais également de nouveaux challenges en perception, ces capteurs devenant mobiles.

Dans ces thématiques, mes recherches adressent les systèmes interactifs, c'est-à-dire des systèmes interagissant avec des humains. Pour répondre aux attentes des utilisateurs, ces systèmes se doivent d'être sensibles à l'environnement qui les entoure, aux utilisateurs présents, à la situation courante. Ils doivent *percevoir*, *comprendre* et *prédire* pour agir en conséquence [7]. La perception multimodale par ordinateur est la problématique de fond des travaux présentés dans ce manuscrit, le traitement du signal (« *signal processing* ») et l'apprentissage automatique (« *machine learning* ») en étant les fondements. Les signaux acoustiques, visuels, proxémiques ou tactiles servent à extraire des informations de plus haut niveau comme le comportement, l'état émotionnel et le but des humains partenaires de l'interaction. Ces outils permettent à nos systèmes de « comprendre » et de prédire les intentions des humains lors d'interactions explicites ou implicites.

Dans ce manuscrit, de façon non chronologique, je présente mes travaux sur la perception multimodale et les systèmes interactifs dans plusieurs contextes suivant l'évolution thématique de mes recherches (fig. 1 page 27). Mes travaux de doctorat traitaient de la reconnaissance de la parole. J'ai poursuivi en 2002 en abordant la perception acoustique en environnement perceptif, puis à partir de 2008, la perception multimodale avec une application au maintien de personnes âgées à domicile et à la localisation de personnes en utilisant des technologies sans fil. En 2012, mes travaux se sont orientés vers l'interaction avec des robots compagnons et plus récemment avec des véhicules autonomes. En parallèle, j'ai contribué aux aspects intergiciels de la perception distribuée dans les environnements perceptifs (2004-2012).

Ce manuscrit doit être lu en gardant à l'esprit l'état-de-l'art à l'époque de ces recherches, certains travaux pourraient bénéficier bien entendu des avancées récentes que nous avons évoquées précédemment. Ce manuscrit est organisé comme suit :

- la fin de ce premier chapitre présente le déroulement de ma carrière puis liste les collaborations qui ont jalonné mes recherches. La figure 1 (page 27) résume ces informations en relation avec l'évolution thématique de mes recherches.
- le chapitre II illustre mes travaux sur les espaces perceptifs multimodaux, tels les salles de réunions augmentées avec des caméras et des microphones ambiants, dans la section « Salle augmentée et perceptive ». Nous avons abordé en toile de fond les aspects intergiciels de la perception distribuée (section « Perception distribuée dans un environnement perceptif »). Nous verrons comment, à partir de ces salles augmentées, nous avons

transposé nos travaux au maintien des personnes âgées à domicile, cette problématique étant un enjeu sociétal majeur lié au vieillissement de la population dans les pays développés (section « Maintien de personnes âgées à domicile »). Enfin, pour conclure nos travaux sur les espaces perceptifs, nous présentons notre contribution à la localisation en intérieur en utilisant des smartphones (section « Localisation en intérieur avec des technologies sans fil »).

- le troisième chapitre dépeint mes travaux sur les interactions homme-robot (*Human Robot Interaction - HRI*). Faisant suite aux progrès en robotique, la perception s'est déplacée de l'environnement au robot mobile. Ce chapitre décrit nos travaux sur l'interaction sociable avec des robots, dont des robots particuliers que sont les véhicules autonomes. Dans la section « Robots compagnons d'assistance à domicile », nous présentons nos recherches sur la détection d'engagement envers un robot compagnon et sur le retour émotionnel que celui-ci peut fournir dans le cadre d'une interaction sociable. Nous présentons également nos travaux sur la modélisation et la prédiction du comportement des piétons des véhicules autonomes (section « Interaction avec des véhicules autonomes »).
- le chapitre IV aborde des thématiques autour de la perception de l'humain et de ses affects. La section « Perception des humains en champ proche » décrit les travaux que nous menons sur la capture de sessions narratives avec des enfants et l'analyse des comportements de joueurs d'échecs lors de la résolution de problèmes. Ce chapitre présente ensuite nos contributions pour la détection de personne. Tout d'abord, nous présentons MobileRGBD, notre corpus enregistré avec un robot compagnon pour évaluer la performance ces systèmes de détection. Le chapitre se clôt avec nos travaux préliminaires sur la détection de personne en utilisant de l'apprentissage profond (section « Détection de personnes »).
- le dernier chapitre présente les directions et les perspectives de mon projet de recherche intitulé « Perception multimodale et interaction sociable ».
- Après les annexes, une sélection d'articles portant sur mes principaux travaux complète ce manuscrit.

I.B Déroulement de carrière

La chronologie ci-dessous présente les grandes lignes de mon parcours. Des informations complémentaires à propos de ma carrière se trouvent en annexe.

2002 Thèse soutenue en janvier 2002 (encadrée par *Jean Caelen*, directeur de recherche CNRS) au laboratoire CLIPS sur la modélisation statistique du langage pour la reconnaissance de la parole en utilisant les documents d'Internet [1].

2002-2005 Post-doctorat sur la perception acoustique dans les environnements perceptifs multimodaux dans l'équipe Prima du laboratoire GRAVIR/Inria.

2005- Maître de Conférences en Informatique (CNU 27ème section) avec une affectation à la Faculté d'Économie de Grenoble / Chercheur dans l'équipe Prima.

Responsable des cours typés informatiques de L1 et L3 présentiels et en enseignement à distance (~750h eq TD, ~800 étudiants, recrutement et encadrement de 8 enseignants temporaires chaque année).

2012-2014 Accueil en délégation à l'Inria. Recherches sur la perception embarquée sur des robots compagnons et sur l'interaction sociable.

depuis 2014 Coresponsable de la spécialité *Graphic Vision Robotic* du Master 2 en anglais MOSIG[3] de l'U*niversité Grenoble Alpes / Grenoble INP*.

depuis 2016 Responsable de tous les cours typés informatiques de la Faculté d'Économie de Grenoble pour les campus de Grenoble et Valence (~1250 h eq TD, ~1350 étudiants, gestion d'une équipe de 2 enseignants statutaires et 10/12 enseignants temporaires chaque année).

Chercheur dans l'équipe *Pervasive Interaction* (équipe faisant suite à l'équipe *Prima*).

I.C Collaborations

Les travaux présentés dans ce manuscrit ne sont bien entendu pas de mon seul fait. Au cours de ces années en tant que post-doctorant puis Maître de Conférences, j'ai rencontré, encadré et collaboré avec de nombreuses personnes en France et à l'étranger. Cette section donne une vue globale de mes principales collaborations. Je profite de l'occasion qui m'est donnée pour remercier toutes les personnes (étudiants, enseignants, ingénieurs, doctorants, chercheurs, administratifs et autres) avec qui j'ai eu l'opportunité de travailler, de collaborer, d'apprendre

[3] *Master Of Science in Informatics at Grenoble*, voir http://mosig.imag.fr/ (dernière visite 04/2018).

et de m'enrichir. Cette section détaille les travaux de doctorat que j'ai co-encadrés, les sujets que j'ai traités avec des étudiants de Master et les principaux projets auxquels j'ai participé.

C.1) **Encadrements doctoraux**

a) « Description Sémantique de Services et d'Usines à Services pour l'Intelligence Ambiante » : (09/2006-09/2009)

Le sujet de la thèse de *Rémi Emonet*, co-encadrée avec *James Crowley* (HDR), se situe dans le cadre de l'intelligence ambiante. Les avancées dans ce domaine sont conditionnées par l'intégration de modules développés par des équipes de recherche de plusieurs disciplines. La thèse s'intéresse à la problématique de l'intégration dynamique de logiciels et de dispositifs développés indépendamment les uns des autres. La méthode proposée se base en partie sur le nouveau concept d'usines à services, usines capables d'instancier des services à la demande. Cette proposition est soutenue par un langage de description de services, un compilateur et un environnement d'exécution. (voir section « Perception distribuée dans un environnement perceptif »). Publications liées : [8, 9, 10, 11, 12, 13, 14].

b) « Localisation d'un utilisateur dans un espace perceptif à grande échelle par analyse multimodale hétérogène » (09/2008-abandon en 2011)

Cette thèse s'est déroulée avec le support financier de *Grenoble INP* (Projet *PersPos*). J'étais co-encadrant avec *Éric Castelli* (HDR, UMI *MICA* à Hanoï). La proposition de cette thèse est d'élargir la perception des utilisateurs à un environnement perceptif plus grand qu'une pièce : bâtiment ou campus intelligent. Cette thèse s'est déroulée à mi-temps entre l'équipe *Prima* à Grenoble et l'UMI *MICA*. Pour des raisons personnelles, la doctorante n'a pas pu conduire cette thèse à son terme.

c) « Smartphone-based indoor positioning using Wifi, inertial sensors and Bluetooth » : (01/2014-12/2017)

Cette thèse en co-tutelle était supervisée par *Éric Castelli* (HDR), *Trung-Kien Dao* à l'UMI *MICA* à Hanoï et moi-même à l'*Inria*. Le sujet de la thèse de *Viet Cuong Ta* fait écho à la thèse abandonnée précédemment. Le but est la localisation des utilisateurs portant un téléphone portable à l'échelle d'un groupe d'immeubles. Dans cette configuration, nous avons fait le choix de limiter les capteurs utilisés aux technologies communes des smartphones ne nécessitant pas d'ajout dans l'infrastructure des bâtiments : les accéléromètres, les gyroscopes, les magnétomètres, le Wifi et le Bluetooth. Nos résultats montrent qu'il est possible d'obtenir une précision de localisation dans un contexte multi-utilisateurs d'environ 3 mètres. (voir section « Localisation en intérieur avec des technologies sans-fil »). Publications liées : [15, 16, 17, 18].

d) « Vers des robots animés - Outils et méthodes pour l'intégration d'artistes animateurs dans la conception de robots expressifs » (10/2013-...)

La thèse d'*Étienne Balit* propose d'inclure des artistes animateurs dans la programmation de mouvements des robots sociaux. Je co-encadre cette thèse avec *Patrick Reignier* (HDR). *Étienne* a mis au point des méthodes facilitant l'animation des robots sociaux soit en utilisant des outils tirés de l'animation, soit en utilisant la manipulation tangible du robot. Cette approche permet notamment de facilement programmer un robot humanoïde pour qu'il joue de la musique sur un xylophone en lui ayant appris à frapper une seule note avec différents styles. Ses méthodes de déformations de trajectoires en préservant le style de l'animation sont très intéressantes et applicables à de nombreux cas. Publications liées : [19, 20].

e) « Human Vehicle Interaction » : (03/2016-...)

Nous co-supervisons la thèse de *Pavan Vasishta* avec *Anne Spalanzani* (HDR, équipe Chroma, *Inria*). Cette thèse du projet ANR Valet porte sur la prise en compte des piétons dans leur interaction avec les véhicules autonomes dans les centres urbains, c'est-à-dire à faible vitesse. Cette thèse se concentre sur la prédiction des comportements des piétons autour du véhicule pour améliorer la sécurité des usagers des centres-villes. Cela permet aux algorithmes de contrôle de la voiture de prendre en compte les piétons dans leurs décisions (trajectoire, vitesse, etc.). Nos modèles sont basés sur le principe sociologique de *Natural Vision* [21] avec des champs de potentiels dynamiques pour modéliser l'environnement, et des approches bayésiennes pour la prédiction. (voir section « Interaction avec des véhicules autonomes »). Publications liées : [22, 23].

f) « Chess Expertise from Eye Gaze and Emotion » : (09/2016-...)

Dans le cadre du projet *CEEGE* (*Chess Expertise from Eye Gaze and Emotion*), nous nous intéressons avec *Thomas Guntz* et *James Crowley* (HDR) à la perception de joueurs lors de parties d'échecs. Le but est de déterminer l'expertise des joueurs et, nous l'espérons à terme, leur état mental (se savent-ils en difficulté ? quelle pièce vont-ils jouer ? …). Dans ce cadre, l'originalité de notre approche multimodale vient de l'intégration, en plus du suivi du corps et des mouvements du joueur, de la perception des émotions faciales et du suivi du regard. (voir section « Perception de l'expertise de joueurs d'échecs »). Publications liées : [24, 25].

C.2) **Encadrement d'étudiants de master**

a) Perception multimodale et détection d'engagement (2011-2012)

Les premiers travaux réalisés avec *Wafa Benkaouar-Johal* étaient intitulés « *Autocalibration of Tracking Towers Equipped with an Omnidirectional camera and a set of Microphones* » [26]. Son travail a consisté en la définition, la mise au point et l'évaluation de la calibration intrinsèque d'une tour de perception multimodale avec ses microphones et sa

caméra panoramique (voir la section « Tour multimodale de suivi de personnes »). Nous avons poursuivi sur la détection d'engagement avec un robot compagnon. Les résultats ont montré qu'il était possible de détecter la volonté d'engagement en utilisant les capteurs présents sur le robot. Nous avons également validé pour un robot compagnon des travaux de Schegloff [27] en psychologie qui indiquaient que la rotation des épaules était le premier indicateur de l'intention d'interagir. (voir section « Détection d'intention d'interagir avec des robots »). Publications liées : [28, 29].

b) Capture de sessions narratives (2015)

Maxime Portaz a travaillé sur la mise au point du dispositif expérimental et des algorithmes sous-jacents pour la construction d'une représentation numérique de séances narratives d'enfants. Les enfants et leurs figurines augmentées sont suivis en utilisant des centrales inertielles (*Inertial Motion Unit* - IMU) et des capteurs RGBD. Ces résultats ont servi de base à l'équipe *IMAGINE* de l'*Inria* pour pouvoir faire une reconstruction 3D de la scène jouée avec une projection du jeu d'acteur des enfants (voir section « Perception de sessions narratives »). Publications liées : [30, 31, 32].

c) Proxémique pour les véhicules autonomes (2017)

Nous avons co-encadré avec *Anne Spalanzani*, *Eléonore Ferrier*, une étudiante en Sciences Cognitives. Nous avons questionné la notion de proxémique pour les voitures autonomes dans les centres-villes en nous intéressant à l'espace personnel que projette le passager d'un tel véhicule. Nous avons réalisé des expérimentations à l'aide d'un véritable véhicule autonome, d'une caméra 360° et d'un casque de réalité virtuelle. Nos résultats montrent que le passager étend bien son espace personnel à la voiture, mais que, contrairement à notre intuition, cette extension est symétrique et non corrélée à la vitesse du véhicule. Publication liée : [33].

d) Détection de personne tombée à terre (2016-2018)

Pour ce projet personnel non financé, j'ai eu plusieurs étudiants de Master 2 au cours des deux dernières années. *Alia Hadjar* (2016) a travaillé sur la détection de personnes tombées à terre en utilisant les données d'une caméra de profondeur et des filtres particulaires. Les résultats n'ont pas été à la hauteur de nos espérances. Nous nous sommes alors intéressés à l'apprentissage profond avec un premier stage (*Daulet Bari*, 2017) avec qui nous avons montré la difficulté de généraliser les résultats en détection de personnes sur les personnes tombées à terre. En 2018, nous continuons à avancer sur le sujet avec un autre étudiant, *Ghadeer Mohamad* (voir section « Perception des humains à distance »).

e) Chess Expertise tough Gaze Emotion and Eye (2018)

Dans le cadre du projet CEEGE, nous avons questionné avec Justin Le Louedec la possibilité de prédire l'attention visuelle des joueurs d'échecs avec des techniques de *Deep Learning*, l'utilisation de réseaux neuronaux convolutifs étant un moyen populaire de prédire l'attention

visuelle, surtout sur des photographies. Celui que nous proposons capture les caractéristiques hiérarchiques et spatiales de l'échiquier pour créer des cartes de saillance, c'est-à-dire la probabilité de fixation de chaque pixel par le joueur, représentant son attention visuelle. À l'aide d'une architecture d'autoencodeur spécifique, nous créons des caractéristiques multi-échelles décodant de multiples degrés de relation entre les pièces. En travaillant sur les données d'eye-tracking recueillies dans le projet *CEEGE* et des parties d'échec en ligne, nous avons conduit des expérimentations mettant en évidence des résultats intéressants. Notre architecture de réseau génère des cartes de saillance significatives sur des configurations d'échec non-connues avec de bons scores sur les métriques standards.

C.3) **Principaux projets**

Cette section détaille les principaux projets qui ont servi de cadre à mes recherches. Cette liste de projets n'est pas exhaustive et ne présente que ceux en lien avec les thématiques abordées dans ce manuscrit. Vous pouvez voir leur chronologie détaillée sur la figure 1 (page 27). Les responsabilités administratives sur ces projets sont détaillées en annexe.

a) IST FAME (Octobre 2001-Janvier 2005)

Description : FAME signifie *Facilitating Agent for Multicultural Exchange*. L'objectif du projet est l'analyse automatique et l'interrogation multilingue de documents audios et vidéos enregistrés automatiquement lors de présentations scientifiques. L'interface d'interrogation est une table augmentée placée au centre d'un espace perceptif couplée avec des microphones ambiants.

Dans ce projet, nous avons mis au point le Cameraman Automatique [34]. Celui-ci est un système perceptif qui analyse la situation courante et réalise en temps réel un film à partir des caméras et microphones présents dans une salle de conférence. Il fait actuellement l'objet d'un transfert industriel.

Partenaires : Universität Karlsruhe (Allemagne), Laboratoire GRAVIR (Grenoble, France), Laboratoire CLIPS (Grenoble, France), Istituto Trentino di Cultura (Trento, Italy), Universitat Politècnica de Catalunya Centre TALP (Barcelone, Espagne), Sony International (Allemagne), Applied Technologies on Language and Speech S. L (Allemagne).

b) IST CHIL (Janvier 2004 - Janvier 2007)

Description : *CHIL* : ***C****omputer in the* ***H****uman* ***I****nteraction* ***L****oop*. L'objectif de *CHIL* est d'inclure l'ordinateur dans l'interaction homme-homme de manière ubiquitaire. Le cadre applicatif concerne les séminaires et les réunions dans des salles augmentées avec des caméras et des microphones. Le système informatique agit en fonction de sa perception des besoins des utilisateurs. Il propose des services appropriés tout en étant le moins intrusif possible.

Partenaires : Fraunhofer Institut für Informations und Datenverabeitung (Allemagne), Universität Karlsruhe - Interactive Systems Laboratories (Allemagne), Daimler Chrysler AG,

(Allemagne), ELDA (France), IBM (République Tchèque), Research and Education Society in Information Systems (Grèce), Insitut National Polytechnique de Grenoble (France), Insituto Trentino di Cultura (Italie), Kungl Tekniska Högskolan (Suède), Centre National de la Recherche Scientifique (France), Technische Universiteit Eindhoven (Pays-bas), Universitat Politecnica de Catalunya (Espagne), Stanford University (USA), Carnegie Mellon University, (USA).

c) ANR CASPER (2006 –2010)

Description : *Communication, Activity Analysis and Ambient Assistance for Senior PERsons*. L'objectif recherché est d'offrir des services d'assistance contextuelle adaptative et de rupture d'isolement pour les personnes âgées et les personnes souffrant de déficits cognitifs légers. L'assistance visée par le projet se fait par un réseau d'aidants distants, qui peuvent être aussi bien des bénévoles que des professionnels.

Encadrement : 1 ingénieur.

Partenaires : *Prima* (Grenoble, France), France Telecom R&D (Grenoble, France) et H2AD (Saint-Étienne, France).

d) PersPos (2008-2011)

Description : *PersPos* (*Person Positioning*) propose le développement de larges « espaces perceptifs dynamiques » (à l'échelle d'un bâtiment, d'un campus ou d'une zone industrielle) distribuant des services adaptés aux utilisateurs, en temps réel et dépendants de leur localisation.

Encadrement : 1 doctorante.

Partenaires : *Prima* (Grenoble, France), Centre de recherche international MICA *(Institut Polytechnique de Hanoi*, CNRS, Grenoble INP).

e) AEN PAL (2009-2013)

Description : Ce projet national interne de l'*Inria* propose le développement de technologies et de services pour améliorer l'autonomie et la qualité de vie des personnes âgées et fragiles à domicile. Le but principal de ce projet est de faire collaborer les équipes concernées autour d'une infrastructure dédiée aux expérimentations collaboratives.

Encadrement : 1 ingénieur.

Partenaires : 12 équipes projets *Inria* : *Coprin*, *Demar*, *Lagadic-Sophia, Pulsar*, *Reves* (Sophia-Antipolis, France), *Flowers*, *Phoenix* (Bordeaux, France), *Maia*, *Trio* (Nancy, France), *E-Motion*, *Prima* (Grenoble, France), *Lagadic* (Rennes, France).

f) FUI PRAMAD (2011-2014)

Description : PRAMAD (Plateforme Robotique d'Assistance et de Maintien À Domicile) est un projet de recherche à visée industrielle. Les buts sont très proches de ceux du projet CASPER : le maintien de personnes âgée/fragiles à domicile en sécurité. Pour ce faire, nous avons développé des fonctions de suivi d'activités, des jeux sérieux via des interactions sociables et du maintien du lien social via des applications de communications embarquées sur le robot *Kompaï* de notre partenaire *Robosoft*.

Encadrement : 1 stage de Master, 4 ingénieurs.

Partenaires : *Orange Labs* (Grenoble, France), *Covéa Tech* (Paris, France), *Wizarbox* (Paris, France), *Robosoft* (robot), *Inria Rhône Alpes* (Grenoble, France), *ISIR* (Paris, France) et l'Hôpital *Broca* (Paris, France).

g) Figurines (2014-2015)

Description : Le nom complet du projet est « *Figurines expressives interactives pour le design narratif* ». L'objectif est de permettre la création de contenu narratif en utilisant des objets réels augmentés par un ou plusieurs conteurs. Les éléments capturés dans le monde réel (mouvement des figurines et informations à propos des narrateurs) permettent de rejouer le contenu narratif en virtuel. Les applications envisagées sont l'étude du procédé narratif chez les enfants et la mise en place d'un outil d'aide à la mise en scène pour le théâtre ou le cinéma.

Encadrement : 1 stage de Master.

Partenaires : Équipes-projets *Prima* et *IMAGINE* de l'*Inria Rhône-Alpes* (Grenoble, France).

h) ANR Valet (2016-2018)

Description : Le projet ANR VALET propose une approche novatrice pour redistribuer des véhicules en autopartage en utilisant des flottes de véhicules guidés par des conducteurs professionnels. Une fois arrivés sur leurs aires de stationnement, les véhicules sont autonomes et se garent seuls. Dans ce projet, nous nous intéressons à la prédiction du comportement des piétons lors des déplacements urbains de la flotte et lors des phases de stationnement sur les parkings.

Encadrement : 1 doctorant.

Partenaires : l'institut *Ircyyn* de l'*École Centrale de Nantes* (France), l'entreprise *AKKA* (Paris, France) et 3 équipes projets *Inria* participent : *Chroma*, *Pervasive Interaction* (Grenoble, France) et *RITS* (Paris, France).

i) ANR CEEGE (2016-2019)

Description : *CEEGE* (*Chess Expertise from Eye Gaze and Emotion*). Nous nous intéressons à l'attention visuelle, aux paramètres physiologiques, émotionnels et comportementaux de joueurs résolvant des problèmes d'échecs. L'objectif est de déterminer s'il est possible d'évaluer automatiquement l'expertise des joueurs et de prédire le déroulement de parties d'échecs à partir de nos observations multimodales.

Encadrement : 1 doctorant.

Partenaires : *Pervasive Interaction* (Grenoble, France) et *CITEC* (Bielefeld, Allemagne).

j) ANR HIANIC (2018-2021)

Description : *HIANIC* (*Human Inspired Autonomous Navigation In Crowds*) a pour but l'étude des interactions entre les véhicules autonomes et les autres usagers de la route dans des espaces partagés en centre-ville, à faible vitesse. Ce projet propose de s'intéresser au comportement des piétons et à son influence sur les déplacements des véhicules, et vice-versa. Pour améliorer l'interaction entre ces différents usagers de la route, de nouveaux dispositifs de communication véhicules-piétons seront étudiés.

Encadrement : 1 post-doctorant (à venir).

Partenaires : équipes *ARMEN* et *PACCE* du laboratoire *LS2N* (Nantes, France), équipe *RITS* de l'*Inria* (Paris, France), équipe *MAGMA* du laboratoire LIG (Grenoble, France), équipe *Chroma* et *Pervasive Interaction* de l'*Inria* (Grenoble, France).

C.4) **Résumé en chiffres**

Au cours de ces années, j'ai co-encadré 6 thèses et j'ai encadré 27 étudiants dont 6 étudiants en Master 2 recherche, 15 étudiants de Master 2 professionnel/d'École d'ingénieur ou de Master 1, 5 étudiants à niveau Bac+2.

J'ai participé à 16 projets[4] de recherche (6 projets internationaux, 6 projet nationaux dont 2 à visée industrielle, 3 projets régionaux et 3 projets de transfert technologique). Pour 7 de ces projets, j'ai eu des responsabilités scientifiques et/ou budgétaires. Ces projets m'ont permis de collaborer avec plus d'une trentaine de partenaires nationaux et internationaux.

[4] Les chiffres présentés ici prennent en compte les projets non listés dans ce manuscrit.

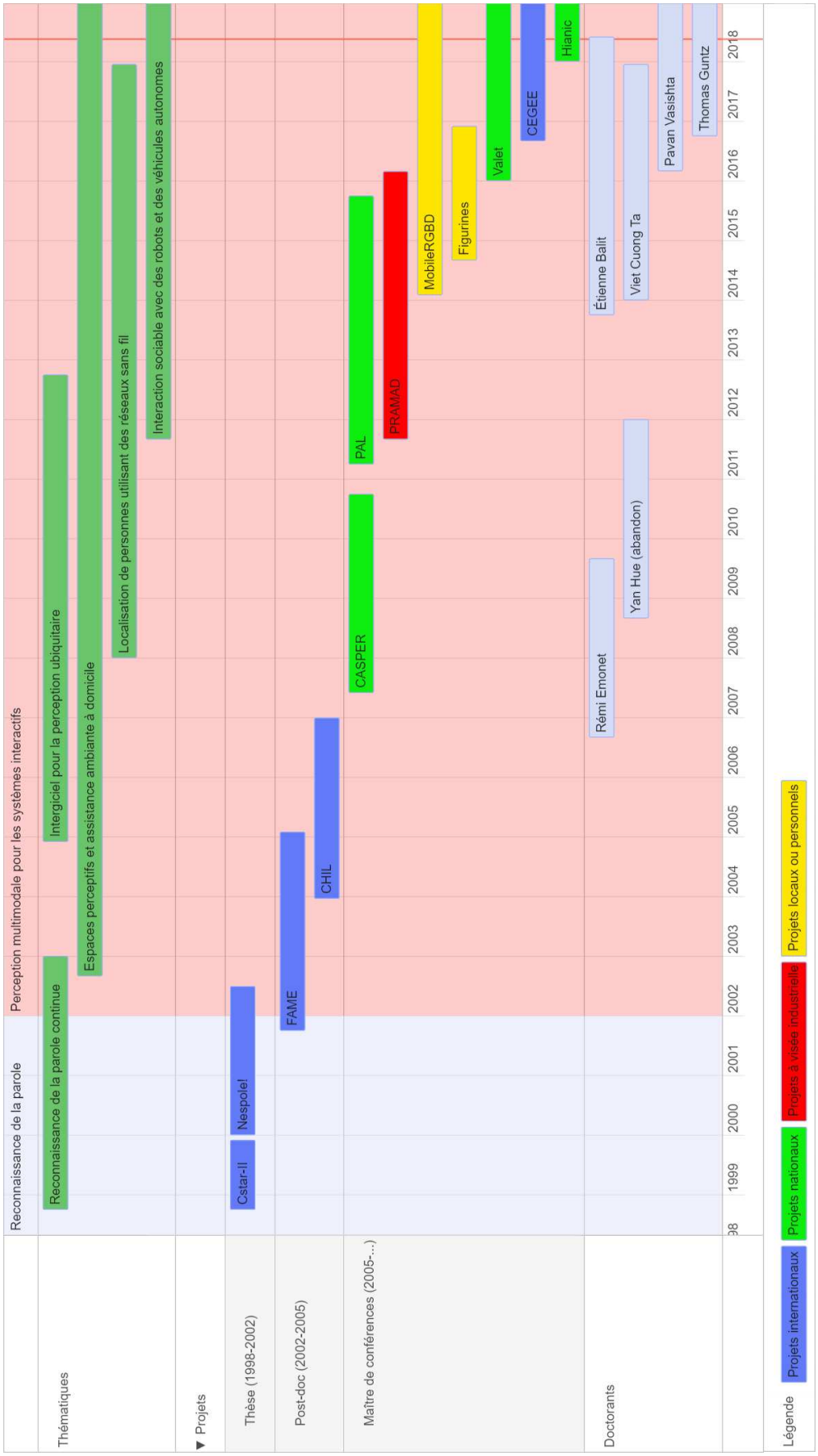


Figure 1 : Évolution thématique de mes recherches, principaux projets, encadrement de doctorants.

Chapitre II
Espaces perceptifs multimodaux

II.A Introduction

Au début des années 2000, les développements technologiques des réseaux informatiques (performance et développement des réseaux sans fil) et des capteurs telles les caméras performantes à bas coût ont fait émerger de nouveaux concepts de lieux où la technologie doit fournir des services sur mesure aux utilisateurs : les espaces perceptifs multimodaux. Les progrès simultanés du traitement du langage, de la vision par ordinateur ainsi que de la conception d'interfaces homme-machine rendaient possibles de nouveaux outils de communication homme-homme ou homme-homme médiatisés. L'intégration de la parole, de la vision, d'un système de dialogue offre la possibilité d'une nouvelle classe d'outils pour faciliter la communication entre les personnes. Dans la littérature, ces environnements perceptifs multimodaux sont souvent qualifiés d'« intelligents » (« *smart* ») avec plusieurs types de déclinaison : bureau [35], salle de réunion [36], salle de cours [37]. Le principe de fonctionnement de ces applications découle de l'informatique ubiquitaire [38] : le système informatique est distribué sur des dispositifs de plus en plus petits et se fond le plus possible dans l'environnement pour disparaître.

La littérature évoque également la notion d'« *intelligence ambiante* » (« *ambient intelligence* » ou « *AmI* »). Cette notion est au confluent des espaces perceptifs multimodaux et de l'intelligence artificielle [39]. L'intelligence ambiante, dans le cadre qui nous intéresse, se résume généralement à des applications sensibles au contexte, ce contexte étant principalement la localisation des personnes et leurs activités. L'idée fondamentale est de percevoir ce contexte pour anticiper les besoins des utilisateurs et déclencher une action [40] pour satisfaire ces besoins. D'autres actions peuvent être plus en lien avec le fonctionnement du système lui-même (enregistrement des informations capteurs, étiquetage des données, indexation, reconfiguration…).

La figure 2 présente un exemple de l'une de nos premières salles d'expérimentation et illustre les équipements que nous pouvons trouver dans ce type d'environnement : des caméras, un ensemble de microphones, un projecteur mobile [41], des dispositifs mobiles (ordinateurs portables, tablettes, téléphones) et des commandes domotiques. Cette salle est également

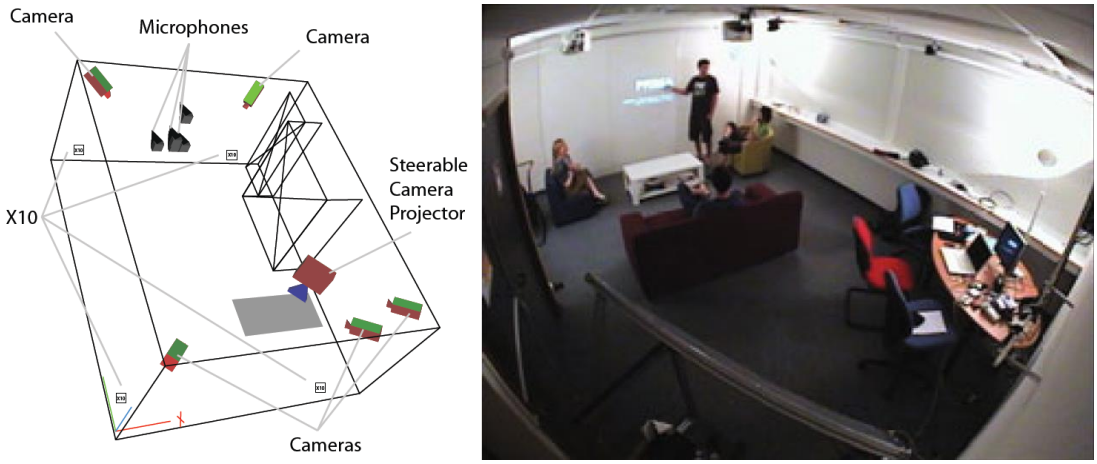


Figure 2 : Espace perceptif expérimental de l'*Inria Rhône-Alpes*.

équipée pour supporter différents réseaux sans fil (Wifi, Bluetooth, X10, *Zigbee*…). Il est donc possible de suivre les activités au sein de cet environnement en utilisant différents outils de perception : détection de parole/de sons, localisation acoustique, reconnaissance de la parole pour la modalité acoustique ; détection de personnes, de postures, de gestes pour la modalité visuelle. Pour compléter l'interaction, il est possible d'agir sur l'environnement par la gestion des lumières, des volets, du vidéoprojecteur ou via les applications installées sur les différents matériels présents dans l'environnement [12].

Au milieu des années 2000, pour faire face au vieillissement de la population, des évolutions de ces espaces perceptifs ont été proposées pour conduire aux maisons intelligentes (« *smart homes* ») [42], des appartements équipés de multiples capteurs. Leur but est de pouvoir maintenir le plus longtemps possible des personnes âgées ou fragiles à domicile. De nombreux projets de recherche [43, 44] ont émergé autour de ce nouveau domaine de recherche qu'est l'Assistance pour la Vie Autonome (AVA), ou « *Ambient Assisted Living* » (AAL). Plusieurs bénéfices sociétaux peuvent en être tirés. Tout d'abord, une qualité de vie meilleure pour les personnes aidées, tout en soulageant la tâche des aidants naturels[5]. Ensuite, le système a la capacité de produire des indicateurs (vitesse de la marche, activité quotidienne, …) pour les aidants professionnels médicaux, indicateurs utiles dans le suivi personnalisé des personnes aidées.

Dans ce chapitre, nous nous intéressons à nos contributions aux espaces perceptifs multimodaux. Ces travaux ont pour dénominateur commun le besoin de *percevoir*, *localiser l'utilisateur* et *ses activités*. Ce chapitre est présenté selon un axe chronologique. Nous verrons que nos travaux ont d'abord porté sur des espaces perceptifs à l'échelle d'une pièce (section

[5] L'aidant naturel ou aidant familial est une « (…) *personne non professionnelle qui vient en aide à titre principal, pour partie ou totalement, à une personne dépendante de son entourage* (…) ». (source [226]).

II.B), puis d'un appartement (section II.C) et enfin à l'échelle d'un bâtiment (section II.D). Ces contributions s'organisent autour de plusieurs facettes :

- la perception acoustique et multimodale dans les espaces perceptifs ;
- l'orchestration des modules de perception et d'interaction dans les environnements ubiquitaires ;
- le maintien de personnes âgées et/ou fragiles à domicile ;
- la localisation multi-utilisateurs en utilisant des technologies sans fil.

II.B Salle augmentée et perceptive

B.1) **Contexte : projets FAME et CHIL**

L'objectif du projet FAME [45] (*Facilitating Agent in Multicultural Environment*) est de fournir un outil de communication homme-homme assisté par ordinateur, et ce possiblement dans un cadre multiculturel, c'est-à-dire avec des personnes échangeant dans leur langue maternelle. Les échanges peuvent intervenir à la fin d'une exposition, lors d'un événement culturel, etc.

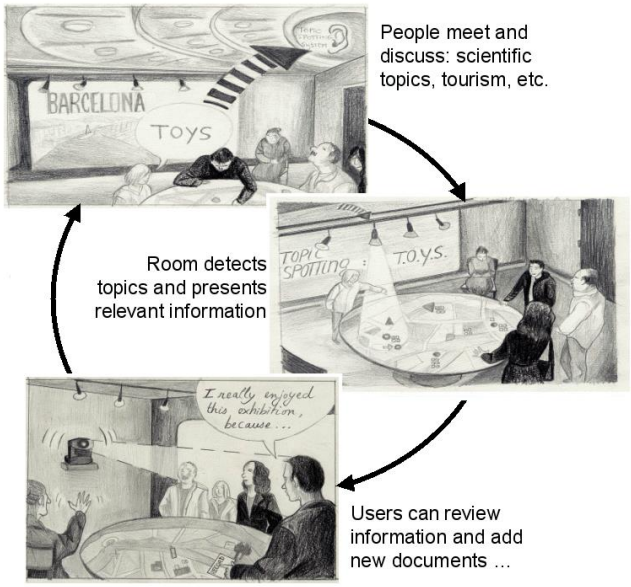


Figure 3 : L'espace interactif du projet FAME. (source [45]).

L'espace perceptif est une table interactive augmentée avec des services perceptifs ubiquitaires (voir fig. 3). L'interaction se déroule autour de cette table augmentée, les utilisateurs échangent autour d'un thème ou de documents disponibles manipulables via la surface tactile de la table. Le système perceptif observe l'interaction humaine plutôt que d'être adressé implicitement. Il scrute les échanges oraux entre les participants, détecte la thématique de ceux-ci et propose dynamiquement de nouveaux documents (photographies, vidéos, etc.). Il peut afficher sur le mur à côté de la table de façon non intrusive des informations connexes.

Les utilisateurs peuvent ajouter des documents au système ou enregistrer une vidéo de témoignage. Ces nouvelles informations sont indexées automatiquement et ajoutées au système pour de futures interactions autour de la table.

Le même type d'approche est employée dans le projet CHIL (*Computer in the Human Interaction Loop*) mais dans un contexte différent [46]. Il s'agit ici d'avoir un majordome ubiquitaire qui assiste les personnes lors d'événements type présentation ou réunion dans des salles augmentées. À la fin de l'évènement, le système indexe celui-ci en utilisant toutes les informations collectées pour enrichir la base de données. Ce majordome a besoin d'une variété de technologies dorénavant matures : la reconnaissance et la synthèse de la parole, l'identification et le suivi des personnes (identification vocale et/ou visuelle), la catégorisation et l'indexation automatique, pour ne citer qu'elles. L'architecture logicielle doit être dynamique et distribuée sur le réseau. La configuration du système d'interaction, de la perception à l'interface homme-machine en passant par le gestionnaire de l'application doit être la plus automatique possible.

Dans le cadre de ces deux projets, nos travaux de recherche ont porté sur les modules de perception acoustique de base pour le français : détection de la parole, localisation acoustique, reconnaissance de la parole, détection du thème du discours. Pour les 3 derniers, nous avons utilisé des approches état-de-l'art [1] à ce moment, nous n'avons donc eu aucune contribution innovante. Dans cette section, nous ne présenterons que le module de détection de parole. Nous présenterons ensuite nos efforts sur l'orchestration des services logiciels pour faciliter le déploiement du système interactif.

B.2) **Détection de parole**

De nombreuses recherches ont été menées sur la détection de parole et différentes approches proposées [47, 48]. Il restait cependant des verrous scientifiques à lever. Le premier concerne les systèmes capables de fonctionner à la fois avec des microphones cravates, des microphones casques ou des microphones distants répartis dans l'environnement. Le second concerne la généricité des systèmes. Au vu de la faible disponibilité de grand corpus parole à ce moment-là, les systèmes état-de-l'art utilisaient des données collectées dans les conditions spécifiques cibles pour l'entraînement des systèmes de détection de parole [46].

L'approche que nous avons proposée pour la détection de la parole répond aux exigences des environnements perceptuels des projets FAME et CHIL, et adresse les verrous précédents. Le but étant d'être le plus générique possible, nous avons fait le choix de n'utiliser les données d'apprentissage des évaluations auxquelles nous avons participé que pour la validation, pas pour l'apprentissage. En outre, nous nous sommes imposés deux contraintes supplémentaires. Notre système devait être autonome et très léger. Autonome, car il devait pouvoir fonctionner sans intervention humaine et ne devait nécessiter aucun réglage spécifique à chaque évolution des conditions expérimentales. Cela renforce la généricité. Léger, car notre environnement

perceptif contient de nombreux microphones : il doit être possible d'en traiter le plus possible en parallèle.

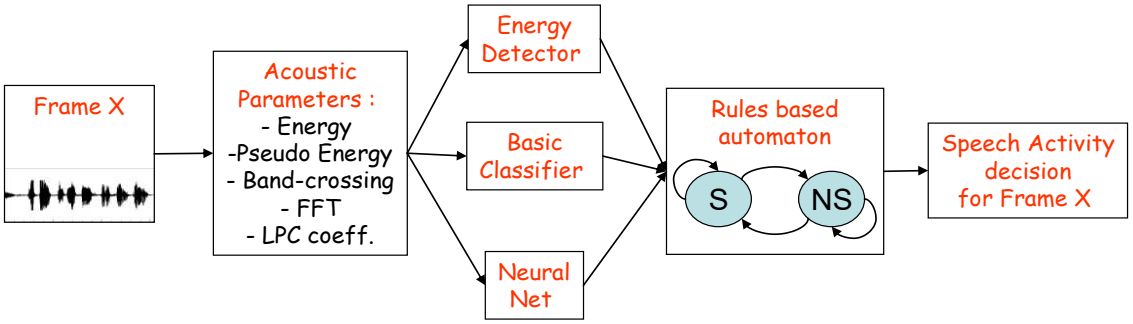


Figure 4 : Diagramme du système de détection de la parole.

Le système de détection de parole est composé de plusieurs sous-systèmes : un détecteur d'énergie, un classificateur spectral et un réseau de neurones entraîné à la reconnaissance de segments voisés, principalement des voyelles. Le détecteur d'énergie qualifie l'évolution de l'énergie dans le signal. Ce module est auto adaptatif et utilise les données collectées sur la dernière minute de signal pour recalculer de nouveaux seuils d'activation. Le classificateur spectral analyse le spectre en bandes de fréquences et tague des classes sonores spécifiques : fricatives ; sons basse fréquence (comme les ventilateurs d'ordinateur ou d'air conditionné) ; autres sons. Le réseau de neurones est un réseau multicouche avec 2 couches cachées de perceptrons. Ce module est le seul sous-système devant être entraîné. Pour rester le plus générique possible, l'entrainement a été réalisé non pas sur des données spécifiques mais sur 1 heure de discours extrait du corpus BREF [49]. Enfin un automate d'état à base de règles détermine le résultat final à partir de la décision des trois sous-systèmes.

Ce système a été évalué dans le cadre du projet FAME et lors de de l'évaluation *NIST Rich Transcription evaluation* en 2006 (RT06s)[6]. La première était l'évaluation du système de détection dans 2 conditions autour de la table augmentée : un microphone cravate, un microphone distant. L'évaluation RT06s consistait en 2 tâches et plusieurs conditions dans des salles de réunion perceptives. La première tâche concernait une présentation orale avec des questions de l'audience, la seconde des réunions impliquant de multiples locuteurs. Plusieurs conditions étaient évaluées : 1 microphone proche, 1 seul microphone distant, plusieurs microphones distants et tous les microphones distants. Les résultats sont résumés dans le tableau suivant et détaillés dans [45, 50].

[6] https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation (dernière visite : 03/2018)

	FAME	RT06s	
		Réunion	Présentation
Microphone proche	5,8%	non pertinent	
1 microphone distant	12,9%	1,63%	4,66%
Multiples microphones distants	-	13,22%	3,73%

Tableau 1 : Taux d'omission moyen des événements parole dans les évaluations FAME et RT06s.

Le principal résultat concernant notre système est son faible taux d'omission d'événements parole : entre 5 % et 15 % d'omissions dans différentes conditions (microphones proches, 1 microphone distant, multiples microphones distants). La limite du système tient dans sa sensibilité à certains types de bruits qu'il détecte comme de la voix et à la détection des frontières de la parole [50]. Pour l'évaluation RT06s, notre système n'a pas été pertinent pour la détection avec un microphone proche. En effet, celui-ci s'adapte automatiquement pour détecter tous les événements parole et détecte les voix des personnes entourant le locuteur principal ce qui est contraire à l'étiquetage des données. Il aurait fallu adapter spécifiquement nos seuils de détection pour cette tâche ce que nous n'avons pas souhaité faire. Ces résultats sont à remettre en perspective des avancées qui se sont déroulées depuis 2006. Les performances actuelles des systèmes de détection de parole sont directement liées à une plus grande disponibilité de données d'apprentissage et aux progrès en apprentissage profond, les meilleurs systèmes obtenant jusqu'à moins de 1,5 % d'erreur [51, 52].

Notre système de détection de la parole a été utilisé dans des travaux permettant de déterminer des sous-groupes d'interaction dans une réunion (thèse d'*Oliver Brdiczka* [53]). Nos travaux de recherche ont aussi bénéficié de ce système de détection de parole : le Cameraman Automatique du projet FAME [34], nos travaux sur le maintien à domicile de personne âgées et nos travaux sur l'interaction multimodale avec des robots compagnons.

B.3) **Perception distribuée dans un environnement perceptif**

a) Prérequis pour la perception distribuée

Dans les applications d'intelligence ambiante développées dans les projets internationaux, de nombreux modules, mis au point par des chercheurs utilisant des techniques et des langages multiples, doivent s'interconnecter dynamiquement afin d'atteindre un objectif commun [40]. Les chercheurs ne sont pas (tous) des architectes logiciels. Chacun a son langage de prédilection soit car c'est celui qu'il maîtrise le plus, soit parce que c'est celui qui est le plus adapté à une tâche particulière. Il faut donc une solution simple à un problème commun : comment découvrir, interconnecter et superviser des services dans le contexte d'applications perceptives multilingues, multiplateformes et distribuées ?

Différentes approches sont possibles. La première approche consiste à convenir, a priori, d'une convention de programmation spécifique : un langage, une plateforme, une technologie, etc. Cela revient à se donner un cadre strict faisant fi des travaux précédemment réalisés et des

évolutions technologiques et algorithmiques. Cette solution n'est envisageable qu'à l'échelle d'une équipe de recherche et à court voire moyen terme. La seconde approche liste tous les services disponibles et cherche s'il existe un intergiciel permettant de tous les interconnecter. Même si cela se révèle compliqué dans les grands projets intégrant plusieurs partenaires travaillant sur des domaines différents, c'est la solution la plus viable. Chaque équipe de recherche, chaque chercheur garde son écosystème de travail. Ils se concentrent sur le cœur de leurs recherches plutôt que d'apprendre un nouveau langage de programmation pour implémenter de manière plus ou moins parfaite leurs travaux.

Dans ce dernier scénario, nous avons identifié plusieurs propriétés que devraient avoir les intergiciels, au moins dans les espaces perceptifs multimodaux qui nous intéressent :

- *Attractif*. Un intergiciel doit être disponible en plusieurs langages de programmation. Nous avons identifié au minimum le C/C++ pour les traitements vidéos/audios lourds, Java pour l'intégration sur smartphone, python pour le *Deep Learning* ou le prototypage rapide aujourd'hui. Cette attractivité sera renforcée s'il fonctionne sur la majorité des systèmes d'exploitation (Windows, Linux, Mac OSX/iOS, Android). Il doit disposer d'une interface de programmation simple orientée utilisateur.
- *Extensible.* L'ajout de nouvelles fonctionnalités à l'intergiciel doit être aussi aisé que possible. L'implémentation de l'intergiciel dans un nouveau langage de programmation doit être facilitée par la documentation.
- *Distribué*. L'intergiciel distribue les services logiciels sur le réseau. Il doit fonctionner sur les réseaux filaires et wifi, ainsi que sur les réseaux *ad-hoc*. Il doit permettre leur découverte et leur interconnexion. Enfin, il doit être capable d'échanger de simples messages textuels ou des structures de données formatées mais aussi des flux de données beaucoup plus conséquents comme les flux audio/vidéo.
- *Maintenable et durable*. La maintenabilité inclut un code source lisible et documenté, la prévisibilité des comportements logiciels et la supervision des services en cours d'exécution sur le réseau. La durabilité est le potentiel de maintenance à long terme et de réutilisation des composants logiciels.

a) *Comparaison des solutions existantes*

En 2005, deux solutions principales remplissaient la plupart des critères énumérés précédemment : *OSGi*[7] [54] et les *Web Services* [55]. D'autres solutions étaient beaucoup plus spécialisées comme *Smart Flow II* [56] du NIST. Pour rapprocher tout cela d'approches plus récentes, ajoutons à notre liste *ROS (Robotic Operating System)* [57] qui est très largement utilisé en robotique pour toutes les tâches de perception, de planification et de contrôle de robot. Le tableau suivant donne une comparaison de ces solutions avec notre intergiciel.

[7] *OSGi Alliance* : http://www.osgi.org/.

Nom	Multi-langage	Multi-plateforme	Messages simples	Flux de données type vidéo	Découverte de service sur le réseau	Outil de supervision
OSGi	Non (Java)	Oui	Oui	Possible	En utilisant *R-OSGi* par exemple	Oui
Web Services	Oui (dépendant de l'implémentation)	Oui	Oui	Design non adapté	En utilisant *WS-Discovery*	Dépendant des solutions
Smart Flow II	C++ mais Java possible	Oui	Design non adapté	Oui	Non	Oui
ROS	C/C++, python	Non	Oui	Oui	Service centralisé	Oui
OMiSCID	C/C++, Python, Java	Oui	Oui	Oui	Oui	Oui

Tableau 2 : Comparaison d'intergiciels disponibles avec OMiSCID, notre proposition.

OSGi permet la construction d'applications Java en recrutant automatiquement des composants. Il valide presque tous nos critères. Cependant certaines fonctions avancées (découverte de composants, …) nécessitent l'ajout de composants spécifiques [58]. Ajoutant à cela sa forte intrication au monde Java, nous ne l'avons pas retenu. Les *Web Services* sont largement utilisés pour les applications distribuées en se basant sur les technologies web. Leur faiblesse pour gérer de gros flux de données, à l'époque, les ont exclus de nos choix. *Smart Flow II* était un intergiciel spécialisé très efficace pour gérer des flux de données provenant de nombreuses sources multimédia disponibles sur le réseau. Sa force était aussi sa faiblesse : il était compliqué de l'utiliser pour échanger des données structurées par simple échange de messages.

Si l'on s'intéresse maintenant à *ROS* qui est plus récent, il est l'un des plus utilisés actuellement dans les domaines de la robotique et quelques domaines associés comme la perception embarquée ou distribuée sur un réseau. Sa plus grande force est la taille de sa communauté. S'il l'on a besoin d'un module d'acquisition vidéo, de calibration de caméra, de détection de personnes, de contrôle ou de navigation pour un robot, de visualisation ou d'un simulateur, plusieurs alternatives s'offrent à nous. Cependant, la forte centralisation de sa solution actuelle (même si cela évolue), la gestion limitée des plateformes (principalement Linux) et l'effort à fournir pour sa première utilisation (l'environnement de compilation avec *catkin* par exemple) sont ses faiblesses. Aurions-nous choisi *ROS* si cela avait été une de nos alternatives en 2005 ? Probablement pas. Nous avions à l'époque, et encore aujourd'hui, un environnement fortement multiplateforme, et le support Java était pour nous indispensable.

Forts de cette analyse, nous avons décidé de développer notre propre intergiciel pour nos besoins dans un premier temps. Nous l'avons ensuite partagé avec nos partenaires puis avec la communauté.

b) OMiSCID, un intergiciel pour les espaces perceptifs

O[3]MiSCID [8] est l'acronyme de *Object Oriented Open-source Middleware for Service Communication Inspection and Discovery*. OMiSCID est un intergiciel dédié aux applications distribuées sur les réseaux avec une approche orientée services (SOA pour *Service Oriented Architecture*). Il offre les briques de base permettant la découverte, l'inspection et la communication multiplateforme et multi-langages entre des modules logiciels distribués sur un réseau. En les illustrant avec des exemples provenant de nos espaces perceptifs, les concepts principaux d'OMiSCID [8] peuvent être décrits comme suit :

- *Services*, le concept fondamental. Un service est un module logiciel qui expose une fonctionnalité sur le réseau. Pour cela, il partage un ensemble de variables et de connecteurs. Par exemple, « *Detect* », un service de détection de personnes, prend en entrée une image provenant d'un service d'acquisition vidéo. Il produit en sortie la liste courante des personnes détectées dans l'image.
- *Variables*. Elles sont simplement des entités nommées associées à des valeurs décrivant un service ou ses valeurs internes. Notre service *Detect* expose des informations sur la localisation et la résolution des images de la caméra.
- *Connecteurs*. Ce sont les points de communication entre les services pour l'échange de messages. OMiSCID fonctionne en mode flux de données. Un message arrive sur un connecteur, il est traité. Le résultat peut être envoyé via un connecteur à un autre service pour continuer le traitement.
- *Connections*. Ce sont les liens entre 2 connecteurs. Cette information permet de superviser, dans le réseau, quels services sont interconnectés et comment.

OMiSCID fonctionne à l'aide de plusieurs couches logicielles comme illustré sur la figure 5

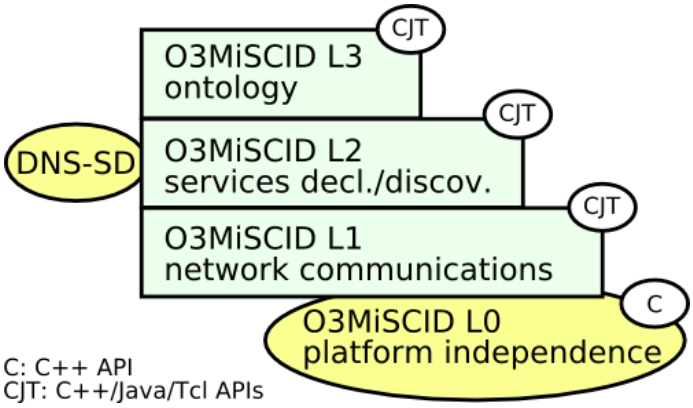


Figure 5 : Couches logicielles de l'intergiciel OMiSCID.

La couche L0 est une couche d'abstraction du système d'exploitation nécessaire pour les programmes en C++. La couche L1 s'occupe de l'échange basique et de messages sur le réseau en utilisant différents protocoles. OMiSCID est capable de transporter aussi bien des messages structurés binaires, textuels (JSON, XML, …) que des flux vidéos lourds. La couche L2 a en charge l'annonce, la découverte et la recherche de services disponibles. La dernière couche

correspond à la partie description et orchestration des services au niveau de l'intelligence ambiante. Cette partie a été traitée dans la thèse de Rémi Emonet [11] (voir section « Orchestration automatique de services »).

OMiSCID a été utilisé dans de nombreux projets (FAME, CHIL, CASPER, PRAMAD, PAL, MobileRGBD, Figurines, CEEGE). Nous l'avons partagé avec nos partenaires et mis à disposition de la communauté. Il bénéficie d'une licence de type MIT et peut donc être librement utilisé[8]. Les travaux autour d'OMiSCID ont été présentés dans plusieurs publications [59, 8, 12, 13, 14].

c) Orchestration automatique de services

Les applications d'intelligence ambiante utilisent de nombreux services à différents niveaux d'abstraction pour fonctionner. Ainsi, un service de localisation 3D dans une salle de réunion augmentée (voir figure 2) devra recruter des services de localisation 2D de plusieurs caméras, un service de localisation acoustique 3D utilisant plusieurs microphones. Les services de localisation 2D ont besoin de services d'acquisition d'images ainsi que de services de projection 2D/3D, le service de localisation acoustique des flux audio en provenance des microphones. Ces services ont besoin d'informations comme la position et l'orientation des caméras et des microphones, les paramètres intrinsèques et extrinsèques des caméras. Tout cela est possible aisément avec un intergiciel comme OMiSCID. Dans une salle équipée de 4 caméras et 8 microphones, cela suppose de lancer 22 services manuellement ou via un script sur plusieurs ordinateurs. L'ajout d'une caméra ou d'un microphone oblige à modifier cette procédure de lancement pour inclure le nouveau dispositif.

Dans les travaux de thèse de *Rémi Emonet* [11], nous nous sommes intéressés à l'orchestration de services, c'est-à-dire au « *processus automatique d'organisation, de coordination, et de gestion de systèmes informatiques complexes, de middleware et de service*s »[9]. Cela correspond à la dernière couche logicielle d'OMiSCID sur la figure 5 (page 37). Si l'on reprend l'exemple précédent, lors du lancement du service de localisation 3D, le système d'orchestration doit être en mesure de lancer automatiquement les services pour chacun des dispositifs matériels disponibles. Il doit en plus instancier tous les services intermédiaires permettant de fournir au service de localisation 3D les paramètres calculés sur les signaux audio et vidéo.

Les principales contributions de cette thèse s'articulent autour de ces besoins et sont résumées dans les sections suivantes.

---

[8] Voir http://omiscid.gforge.inria.fr/ et https://github.com/Vaufreyd/Omiscid/ pour de plus amples informations (dernière visite 04/2018).

[9] Source https://fr.wikipedia.org/wiki/Orchestration_informatique (dernière visite 04/2018)

Notion d'usines à services paramétriques et de composition

Le besoin pour l'orchestrateur de raisonner en intégrant à la fois les services disponibles dans l'environnement à un temps *t* mais aussi les services qu'il pourrait être possible d'instancier a fait émerger une nouvelle notion : l'usine à service (« *Service Factory* »). Une usine à services est un service qui peut, à la demande ou en fonction d'un besoin, instancier une version spécifique d'un autre service. Dans notre approche, il existe principalement 2 types d'usines à services. Les premières sont des usines paramétriques. Elles sont capables d'instancier un service à partir de paramètres, par exemple un service d'acquisition d'images sur une caméra spécifique et/ou à une certaine résolution. Le second type d'usine concerne les usines de composition. Une usine de composition permet de définir qu'un service particulier peut être obtenu en composant plusieurs autres services.

Langage de description de service et d'usine à services (UFCL) et raisonnement

User-oriented Functionality Composition Language (UFCL [10]) est un langage de description de fonctionnalités de service et, pour la première fois, de fonctionnalités d'usines à service. Il permet *via* une syntaxe simple de définir les fonctionnalités d'un service et des usines à services. Il permet de définir des correspondances de fonctionnalités permettant d'exprimer un service à partir d'un autre.

```
composing
    grounding "C(start)"
    format "<run f='?pFreq'/>"
gives a Timer
    with freq = ?f
```

```
a Metronome
    having bpm = ?f
isa Timer
    with freq = ?f / 60
```

Exemple 1 : Exemple UFCL d'une usine à service paramétrique (à gauche) et d'une correspondance de fonctionnalités (à droite) (source [9]).

Dans l'exemple ci-dessus, nous trouvons à droite la description d'une usine à services paramétrique. Elle est capable d'instancier un *Timer* à n'importe quelle fréquence. Pour ce faire, le système d'orchestration lui envoie sur son connecteur *start* (propriété *grounding*) un message XML avec la valeur désirée (*format*). La description UFCL à droite définit la correspondance entre un *Metronome* et un *Timer*. Avec ces informations, l'orchestrateur est capable d'instancier également n'importe quel *Metronome*.

L'expressivité d'UFCL permet de décrire une très grande variété de services. Des exemples et une méthode de design de systèmes avec ce langage sont proposés dans le manuscrit de thèse [11].

Génération de règles et raisonnement

Pour avoir les capacités de planification à partir des descriptions des services et des usines à services, il est nécessaire de disposer de capacités d'inférence. Nous avons fait le choix

d'utiliser un moteur d'inférence existant. Parmi les choix possibles, *Jena* [60], un moteur d'inférence utilisé dans les recherches sur le Web Sémantique, a retenu notre attention. À partir des descriptions *UFCL*, une générateur automatique de règles dans le langage natif de Jena, le *Resource Description Framework* (RDF) [61] a été proposé. L'implémentation autorise à chaque service de se déclarer dynamiquement ainsi que ses usines à services, ce qui met à jour la base de règles. Le système calcule ensuite les orchestrations de services répondant aux requêtes.

Ces travaux ont donné lieu à plusieurs publications [9, 10, 11] en plus des publications précédemment citées pour le middleware *OMiSCID*.

II.C Maintien de personnes âgées à domicile

C.1) **Contexte : projet CASPER**

La fin des années 2000 a vu l'émergence de systèmes perceptifs dont le but était le maintien de personnes âgées à domicile. Dans ce contexte, les données sont multimodales par nature. Des informations visuelles, des informations acoustiques, des informations provenant de capteurs domotiques (températures, luminosité, ouverture de portes, détection de mouvements…) composent le système. De nombreux projets ont vu le jour avec comme objectif l'assistance à domicile[10] (AAL pour « *Ambient Assisted Living* ») : *Aging In Place* [62] de l'Université du Missouri, DESDHIS [63], GERHOME [64] du Centre Scientifique et Technique du bâtiment (CSTB) et de l'*Inria*, *House_n* [65] du *Massachusetts Institute of Technology* (MIT), le projet européen SOPRANO [66], SWEET-HOME [67]…

Notre objectif dans le projet CASPER était d'offrir des services d'assistance contextuelle adaptative à distance et de rupture d'isolement pour les personnes âgées/fragiles et les personnes souffrant de déficits cognitifs légers. L'assistance visée par le projet se fait par un réseau d'aidants distants, qui peuvent être aussi bien des bénévoles que des professionnels. Le système d'assistance à domicile était un environnement perceptif avec des caméras et des microphones disséminés dans l'habitation de la personne aidée. Cette information était couplée à des informations provenant du système domotique (commandes de volets, ouverture de portes, du réfrigérateur, …). Des indicateurs de santé et d'activité étaient remontés aux aidants familiaux pour les rassurer et aux aidants professionnels pour personnaliser la prise en charge de l'aidé(e).

[10] Une liste de projets européens en AAL est disponible sur le site http://www.aal-europe.eu/our-projects/ (dernière visite 04/2018).

La contribution de notre équipe, chercheurs et ingénieurs, a porté sur plusieurs points :

- infrastructure logicielle à l'aide de notre intergiciel OMiSCID, intégration avec les capteurs domotiques et l'infrastructure d'assistance existante chez nos partenaires ;
- perception multimodale (localisation et déplacements dans l'appartement, détection de bruits d'eau dans les tuyaux, …) et détection d'activité (prise de repas, repos, toilette, coucher/lever …).

À partir des informations de perception multimodale, nous avons participé à la mise au point d'un système de suivi par indicateurs qui permettait de mesurer un état de fragilité de la ou des personne(s) suivie(s). Un algorithme de suivi temporel des activités permettant de détecter un décalage dans le mode de vie des personnes, signe d'aggravation de la maladie d'Alzheimer par exemple, a été développé.

Vous ne trouverez cependant aucune publication sur toutes ces thématiques. L'accord de consortium du projet CASPER était très strict et ne nous a pas permis de publier nos travaux.

C.2) **Retour des utilisateurs et évolution de notre approche**

À la fin du projet CASPER, des groupes d'échange autour des propositions que nous avions faites dans le projet ont été organisés. Ces groupes comptaient des personnes potentiellement utilisatrices du système (des aidés), des aidants naturels et professionnels, des personnes intervenant à domicile (aides ménagères, etc.). Le système d'assistance résultant du projet leur a été présenté. Des discussions autour des fonctionnalités, du mode d'assistance et de l'installation du système à domicile ont eu lieu. Les aspects éthiques du système ont été décrits. Par exemple, le système ne stocke aucune image des caméras, juste des paramètres calculés sur celles-ci. Les personnes ont pu exprimer leurs ressentis, leurs inquiétudes et leurs espoirs concernant cette technologie.

Les résultats de ces groupes d'échange furent très intéressants. Ainsi, si l'on s'intéresse à la perception multimodale, la présence de ces technologies n'est pas considérée comme un problème, surtout si cela permet de rester à domicile pour la personne aidée. Ce résultat était surprenant tant le système aurait pu paraître intrusif. Par contre, l'installation du matériel lui-même, c'est-à-dire faire des trous dans les murs pour passer des câbles, fixer les caméras était totalement rédhibitoire. Ce résultat était, par contre, inattendu.

Suite à ces retours, nous avons dû imaginer de nouveaux paradigmes de perception multimodales à domicile. Nous avons dans un premier temps développé des tours de perception multimodale. Celles-ci sont présentées dans la section suivante. Dans un second temps, suivant les progrès de la robotique, nous avons travaillé à doter des compagnons robots de perception leur permettant de remplacer avantageusement l'installation d'un environnement perceptif complet chez des personnes aidées, le robot se déplaçant dans l'habitation. Cette contribution est présentée dans la section « Robots compagnons d'assistance ».

C.3) **Tour multimodale de suivi de personnes**

L'idée fondatrice des tours de perception multimodales est de proposer un suivi visuel et acoustique en un minimum d'encombrement pour répondre à plusieurs préoccupations :

- celles des aidés qui ne souhaitent pas d'installation fixe du matériel de perception ;
- celles du réseau d'aidants qui souhaite un matériel simple à installer et à configurer.

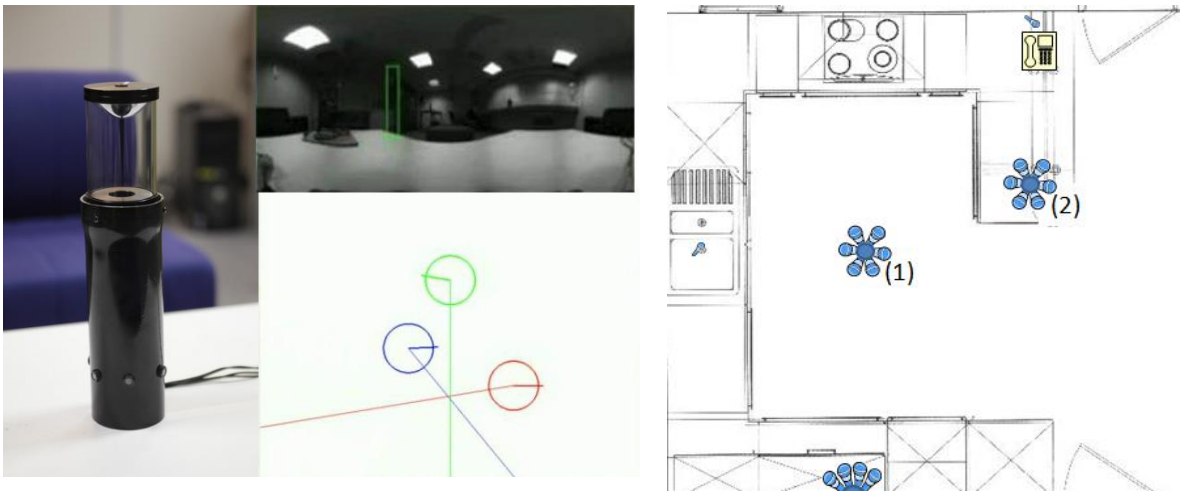


Figure 6 : Tour multimodale de perception avec calibration des positions relatives de 3 tours (à gauche). A droite, installation d'un groupe de tours (en bleu) dans un appartement.

Sur la figure 6, nous voyons à gauche le design de notre prototype de tour multimodale. Cette tour a un encombrement similaire aux assistants vocaux qui se développent depuis quelques années (Alexa, Google Home, …). Celle-ci est équipée d'une caméra panoramique à 360° et d'un ensemble de microphones mobiles à sa base. La caméra 360° permet le suivi des personnes dans l'environnement, les microphones permettent de suivre l'activité sonore (détection de parole ou de bruits) et de localiser la direction d'une source sonore [68]. Ces tours sont prévues pour fonctionner en groupe, comme nous pouvons le voir sur la droite de la figure, pour offrir une couverture optimale de l'espace.

Le principal verrou à leur utilisation est la nécessaire calibration du système : il faut pouvoir connaître la position relative de chacune des tours pour construire de l'information pertinente à partir de toutes. De plus, il est probable que les personnes âgées ou leurs aidants déplacent ces tours, même temporairement, et perturbent le système. Pour la partie acoustique, il faut déterminer la position des microphones sur le pourtour de la tour. Comme on peut le voir sur la figure 6, si une tour est placée près d'un mur, il est possible de déplacer les microphones pour permettre la localisation acoustique.

Dans les travaux de *Wafa Benkaouar-Johal* [26] que nous avons encadrés avec *Amaury Nègre*, nous avons travaillé sur l'auto-calibration d'un groupe de tours multimodales. Cette calibration doit permettre d'avoir un référentiel commun à toutes les tours. Dans un premier temps, pour chaque tour, il faut trouver la position relative des microphones. Pour cela, lorsqu'une cible visuelle est détectée et qu'il y a détection de parole, nous pouvons estimer la

position angulaire des microphones sur la tour. Dans un second temps, il faut calibrer automatiquement la position relative des tours.

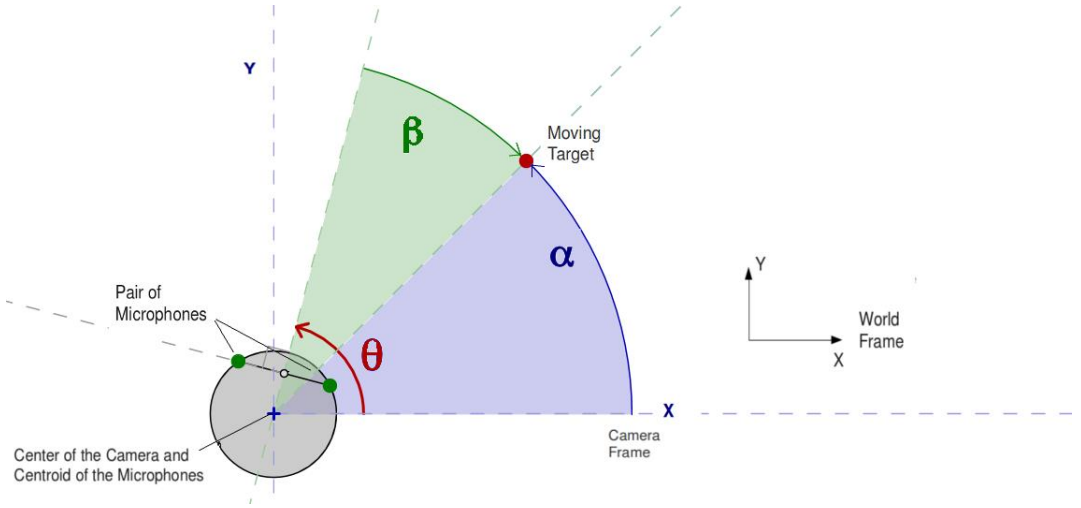


Figure 7 : Illustration dans le plan 2D du suivi par une tour composée d'une caméra omnidirectionnelle et d'une paire de microphones (source [26]).

La figure précédente illustre le fonctionnement de l'auto-calibration pour déterminer la position angulaire des paires de microphones en fonction d'une cible détectée par les modules de perception. En nous limitant à un plan 2D, l'estimation de l'angle θ revient pour chaque paire de microphones n à chaque instant k :

$$\begin{cases} \cos(\alpha^k - \beta_n^k) = \cos\theta_n \\ \sin(\alpha^k - \beta_n^k) = \sin\theta_n \end{cases} \quad (1)$$

En ajoutant un bruit ε à nos mesures, le système d'équations précédent peut se réécrire :

$$\begin{cases} \cos((\alpha^k + \varepsilon_{\alpha^k}) - (\beta_n^k + \varepsilon_{\beta_n^k})) = \cos(\theta_n) + \varepsilon_{\theta_n} \\ \sin((\alpha^k + \varepsilon_{\alpha^k}) - (\beta_n^k + \varepsilon_{\beta_n^k})) = \sin(\theta_n) + \varepsilon_{\theta_n} \end{cases} \quad (2)$$

Afin de trouver une solution optimale pour θ_n, nous utilisons l'historique des dernières valeurs de α et β et nous cherchons la valeur optimale avec la méthode des moindres carrés. Cette approche a été évaluée avec différentes trajectoires sur des données simulées avec différents types de bruits [69]. Nous avons montré qu'avec quelques dizaines de valeurs d'historique, l'erreur angulaire est de moins de 10°. À partir de 300 valeurs, l'erreur d'estimation de θ_n est de moins de 2°.

Sur le même principe et en s'inspirant des travaux de *Ghandi et al.* [70] et de *Meingast et al.* [71], une calibration des positions relatives des tours multimodales a été développée. Sur la figure 6, nous voyons le résultat de cette calibration automatique sur la positions de 3 tours dans l'espace perceptif que nous avons présenté sur la figure 2.

II.D Localisation en intérieur avec des technologies sans fil

D.1) **Contexte**

La fin des années 2000 a connu une autre révolution technologique et sociologique : l'apparition et le développement des smartphones. Avec leur popularité et leur utilisation croissantes dans la vie quotidienne, les capacités de ces dispositifs en terme d'accès réseau, d'interface utilisateur ou de positionnement de leur porteur trouvent de nouvelles applications dans de nombreux domaines tels l'assistance à domicile, la domotique, la gestion énergétique des bâtiments, etc. Leurs capacités de localisation ont attiré l'attention de nombreuses équipes de recherche [72].

Le *Global Positioning System* (GPS) est devenu une norme pour la localisation extérieure avec une précision centimétrique avec des équipements spécifiques [73]. Dans un environnement intérieur où cette modalité n'est pas disponible, il existe plusieurs alternatives pour localiser les utilisateurs. La première, directement corrélée à nos travaux antérieurs, consiste à utiliser des zones de passages spécifiques dans le bâtiment et à y installer des environnements perceptifs. L'avantage principal est la disponibilité de nombreux algorithmes ayant montré leur performance sur cette tâche. Les inconvénients de cette solution concernent l'infrastructure à installer et la couverture faible et discontinue du bâtiment. La seconde alternative est l'utilisation de SLAM visuel (*Simultaneous Localization and Mapping*) pour se repérer et s'orienter à l'intérieur d'un bâtiment même avec un smartphone [74, 75]. Cette approche revêt une contrainte supplémentaire, il faut porter le smartphone ou une caméra de manière à avoir toujours vue sur l'environnement [76]. Si l'on s'intéresse aux technologies sans fil, l'utilisation de capteurs spécifiques *Ultra WideBand* (UWB ou Ultra Large Bande en français, ULB) est la solution la plus précise [77]. Cependant cette performance implique un coût conséquent et nécessite des équipements supplémentaires.

Le développement des réseaux informatiques sans fil (Bluetooth et Wifi) a permis l'émergence d'algorithmes pour localiser les utilisateurs [72]. Ces réseaux sont devenus, en plus d'un vecteur de transport des données, une opportunité avec leur faible coût et leur présence dans la plupart des bâtiments accueillant du public. La performance actuelle de ces systèmes permet d'obtenir une erreur moyenne de positionnement de 5 mètres dans un contexte multi-bâtiments avec plusieurs étages [17].

Dans nos travaux sur la perception et la localisation de personnes, nous nous sommes naturellement intéressés à cette localisation sans fil comme un canal supplémentaire d'information. Dans le projet *PersPos* avec nos partenaires du laboratoire MICA à Hanoï, nous avons imaginé un concept de services ubiquitaires à l'échelle d'un bâtiment ou d'un campus. L'ambition était de de fournir des services contextuels en fonction de la position d'un utilisateur. Cette localisation reposait à la fois sur des technologies sans fil et sur des technologies de suivi disponibles dans des espaces perceptifs à plus petite échelle (salle de

cours, salle de réunion…) bien plus performantes. Ces travaux n'ont pas abouti suite à l'arrêt de la thèse associée à ce projet.

D.2) **Localisation collaborative en contexte multi-utilisateurs**

Pour localiser des utilisateurs en intérieur à l'aide d'une infrastructure réseau sans fil, plusieurs approches sont possibles. Les premières sont les approches géométriques [78]. Celles-ci nécessitent de connaître le positionnement exact des points d'accès réseaux et de pouvoir estimer leurs distances avec le téléphone. Les approches les plus répandues de nos jours emploient une « empreinte digitale » (*fingerprinting*) de la puissance des différents signaux (*Received Signal Strength Index - RSSI*) pour caractériser des points dans l'espace. L'hypothèse sous-jacente est que les signaux seront toujours pratiquement identiques en ces points, ce qui est une hypothèse forte. Pour ces approches, il est donc nécessaire de fournir un corpus de données pour l'estimation de ces empreintes digitales. La littérature reporte différentes techniques d'apprentissage utilisées dans ce contexte : des méthodes probabilistes [79], *K-Nearest Neighbors* (KNN) [80], des réseaux de neurones [81], des *Random Forests* [82] pour ne citer qu'elles.

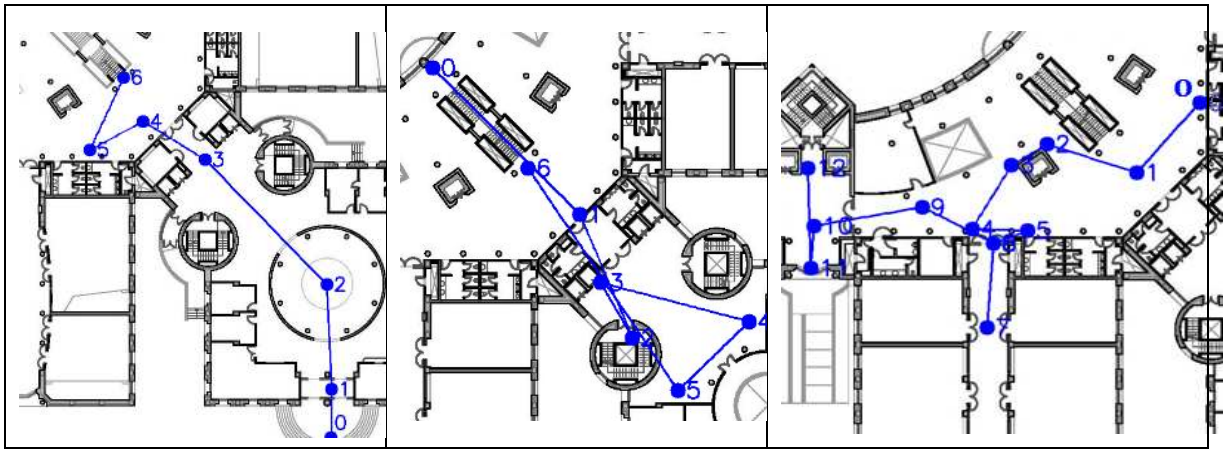


Figure 8 : Exemples de trajectoires dans un bâtiment issue du challenge à la conférence IPIN2016 (source [18]).

La figure précédente illustre des exemples de trajectoires réalisées par des utilisateurs portant des smartphones. Cette figure illustre certaines difficultés pour les algorithmes de localisation. La première est la topologie des lieux. En fonction des murs, des meubles et même des personnes présentes à un temps *t*, la propagation des signaux sans fil est différente. La variété des déplacements, les différences techniques entre les smartphones et la façon de **65**les porter (à la main, dans la poche, dans un sac à dos…) accroissent les erreurs.

Dans les travaux de thèse de *Viet Cuong Ta*, nous avons dans un premier temps étudié les différentes approches pour la localisation intérieure avec les technologies Wifi couplées aux données fournies par les capteurs inertiels des téléphones (accéléromètres, gyroscopes, magnétomètres). Nous avons proposé un algorithme en 2 passes qui commence par détecter l'utilisateur dans le bâtiment et son étage. Ensuite, l'utilisateur est localisé dans cet étage. Pour

valider nos travaux, nous avons participé au challenge international de la conférence *Indoor Positioning and Indoor Navigation* (IPIN2016) [17] qui met en compétition des équipes de recherche sur le thème de la localisation en intérieur. Pour notre première participation à ce challenge, nous avons terminé dernier de la compétition (sur 5 équipes). Suite à cette expérience, nous avons amélioré nos algorithmes pour obtenir 93 % de précision dans l'identification d'étage et une erreur moyenne de 5,12 mètres [18].

La seconde contribution de nos travaux concerne la localisation collaborative. Les deux principes sous-jacents sont d'utiliser le réseau *Bluetooth* comme source additionnelle d'informations et de les partager entre les différents porteurs de smartphones. *Liu et al.* [83] et le système *Social-Loc* [84] en sont des exemples. Ils proposent de détecter la proximité des utilisateurs respectivement avec le réseau *Bluetooth* et le *Wifi Direct*. Dans la littérature, la précision de la localisation *Bluetooth* est donnée avec une erreur moyenne de 2 mètres dans un environnement comptant un grand nombre de capteurs [85].

Dans notre approche, nous avons fait le choix de ne rien rajouter à l'infrastructure réseau mais de profiter des capacités de communication *Bluetooth* des smartphones. À partir des RSSI échangés entre les téléphones, il est possible d'estimer la distance entre ceux-ci. Nous avons proposé 2 approches pour intégrer cette distance inter-smartphones dans nos calculs. Toutes deux associent le positionnement Wifi et le positionnement relatif *Bluetooth* représentés sous la forme d'une distribution gaussienne. La première est dite non temporelle et cherche à minimiser une fonction d'erreur entre les 2 positions. La seconde intègre le temps dans cette fonction d'erreur pour accroitre la précision.

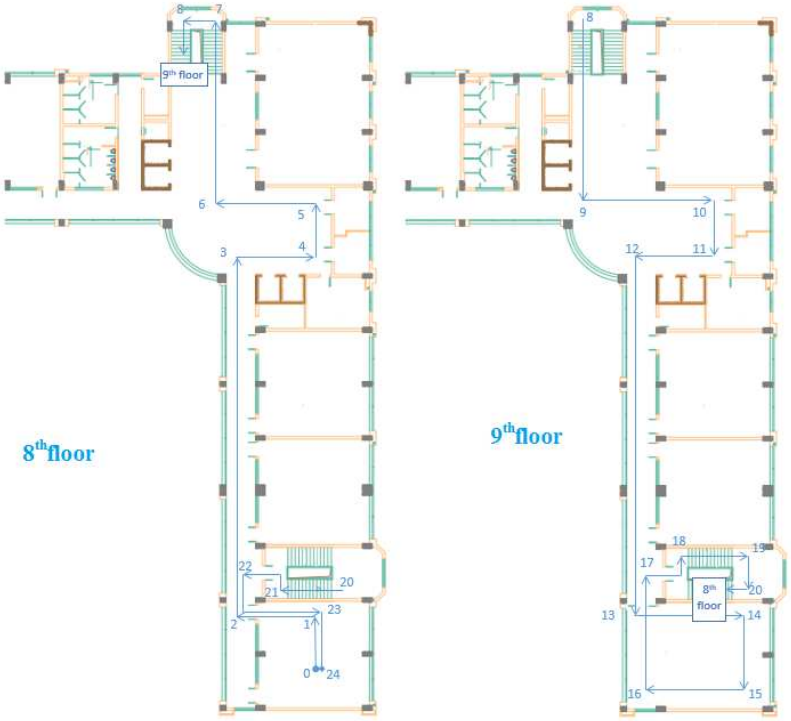


Figure 9 : Trajectoire commune suivie sur 2 étages (en bleu) par les utilisateurs dans notre corpus pour la localisation collaborative *Wifi/Bluetooth* (source [18]).

Pour évaluer les performances de nos approches, nous avons enregistré un corpus de données sur 2 étages du laboratoire MICA à Hanoï (voir figure 9). Tous les enregistrements suivent globalement une trajectoire commune composée de couloirs, de bureaux et d'escaliers. Sa longueur est d'environ 200 m pour un temps de trajet d'environ 300 secondes. Différents scénarios impliquent de 2 à 4 utilisateurs, groupés ou séparés, qui marchent dans le même sens, chaque utilisateur portant un téléphone ou une tablette différents.

Nous avons comparé les performances des approches non temporelle et temporelle avec comme score de référence notre localisation Wifi seule [18]. Sur ces données, la localisation Wifi a une erreur moyenne de 3,68 m, l'approche non temporelle de 3,27 m (-11,27 %), l'approche temporelle de 2,44 m (-37,71 %). Ces résultats montrent un potentiel bénéfice de nos propositions avec un net avantage pour l'approche temporelle.

À l'heure de la rédaction de ce manuscrit, ce résultat doit être validé et confirmé dans d'autres conditions expérimentales : autres bâtiments, autres infrastructures, autres scénarios, plus d'utilisateurs. Cependant, cette amélioration de la précision permet d'envisager, à moyen terme, d'utiliser la localisation collaborative *Wifi/Bluetooth* dans les environnements comme les campus, les gares ou les aéroports.

Chapitre III
Interaction avec des robots

III.A Introduction

La communication humaine est très complexe. Elle n'utilise pas seulement le canal verbal, mais de nombreux canaux pour envoyer et recevoir divers messages tout en interagissant : elle est intrinsèquement multimodale [86]. Dans le livre "*Bodily Communication*" [87], *Argyle* mentionne différents signaux provenant de différentes modalités utilisées pour la communication non verbale. *Bull* [88] propose un modèle où la communication repose sur des signaux et des codes socialement partagés, ce qui implique une composante sociale et culturelle dans la communication.

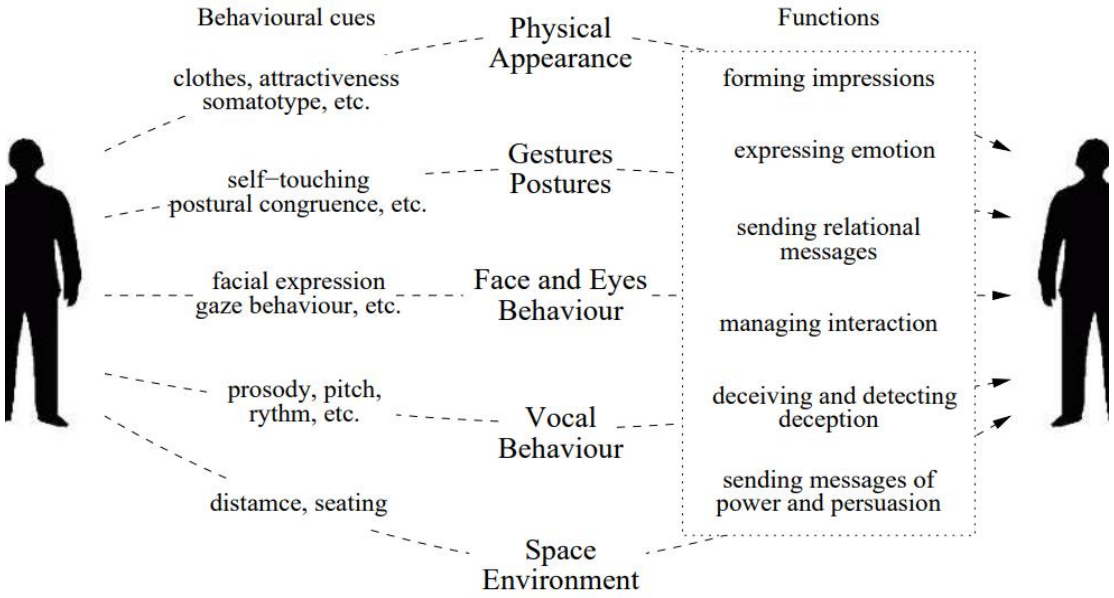


Figure 10 : Comportements, codes sociaux et fonctions sociales associées (source [89])

La figure précédente propose un résumé des modalités présentées dans la littérature. On voit qu'en plus des modalités vocales et faciales, des informations précieuses sont convoyées par l'apparence physique avec une influence de l'habillement ou même de la forme du corps. Les gestes, le toucher et la gestion de l'espace complètent ce tableau. Ces informations sociales sont exprimées de manière intentionnelle ou non intentionnelle. Au début des années 2010, un

nouveau courant de la perception par ordinateur est né pour traiter spécifiquement ces informations : le « *social signal processing* » [90] (traitement de signaux sociaux en français).

Dans nos interactions, nous sommes inconsciemment à la recherche de ces signaux sociaux pour comprendre les attentes et les intentions de notre partenaire. Des phénomènes de mimétisme se mettent également en place entre les partenaires [91]. Dans le cadre de la communication homme-robot, ce dernier doit décoder ces signaux comportementaux et non verbaux afin d'agir en conséquence. Les systèmes capables de s'adapter et de répondre en temps réel à ces signaux sociaux d'une manière polie, non intrusive ou persuasive, sont susceptibles d'être perçus comme étant plus naturels, efficaces et dignes de confiance [92, 93]. Pour le cas particulier que sont les véhicules autonomes, la prédiction des intentions et des déplacements des humains est primordiale, notamment dans les centres-villes [22].

Le fil conducteur de ce chapitre est la prédiction des intentions des humains qui interagissent avec des robots. Nous nous intéressons à la prise en compte des signaux sociaux dans l'interaction avec différents types de robots. Dans le contexte de l'assistance aux personnes âgées et/ou fragiles, les caractéristiques sociales semblent être cruciales pour l'acceptation d'un compagnon robotique dans un environnement domestique [94]. Dans le cadre des véhicules autonomes, nous nous intéresserons à la communication entre ces véhicules et les piétons. Les contributions détaillées dans ce chapitre adressent les thématiques suivantes :

- la détection de l'intention d'interagir avec un robot compagnon d'assistance ;
- l'expressivité des robots compagnons ;
- la modélisation et la prédiction du comportement des piétons pour un robot particulier, le véhicule autonome, en centre-ville.

III.B Robots compagnons d'assistance à domicile

Qu'il s'agisse d'un coach personnel, d'un majordome au bureau ou d'un aidant à domicile, le robot dit compagnon est destiné à être un assistant dans la vie quotidienne. Pour les robots compagnons d'assistance à domicile, le but est principalement le maintien de personnes âgées ou fragiles à domicile. Ce maintien revêt un caractère particulier lorsque l'on sait qu'il a un impact direct sur l'espérance de vie. Parmi les fonctions embarquées sur ces robots pour remplir cette mission, nous trouvons le divertissement et les jeux sérieux, l'assistance physique (aide pour se relever en cas de chute, saisie d'objet), la surveillance d'activités et l'évaluation de l'état de fragilité [95, 96, 97, 98]. Les robots compagnons peuvent également avoir des visées thérapeutiques comme dans les recherches s'intéressant à l'autisme [99].

Comme cela est souligné dans la littérature [100, 101], le principal défi dans la construction de robots compagnons est de fournir une compétence sociale dans la perception, le raisonnement et l'expression des aspects sociaux et affectifs des interactions avec un humain. Cela leur donne ce que l'on peut nommer « présence sociale » [102]. Les robots compagnons sont destinés à interagir avec les humains dans des environnements domestiques. Afin de rester

crédibles, ils sont tenus de se comporter et de réagir selon des modalités prédéfinies correspondant aux instructions et signaux sociaux exprimés par l'utilisateur.

Après avoir déplacé les capacités de perception des espaces perceptifs multimodaux (voir « Maintien de personnes âgées à domicile ») pour les embarquer sur un robot compagnon, nous nous sommes intéressés à deux thématiques que nous développons dans cette section : la détection de la volonté d'interaction des humains et l'expressivité de notre robot compagnon via le rendu d'émotions sur une tête robotisée.

B.1) **Détection d'intention d'interagir avec des robots**

Les humains détectent instinctivement l'intention des personnes présentes autour d'eux. Ainsi, si une personne s'approche de nous pour nous poser une question, nous sommes généralement capables de le détecter avant même qu'elle n'ait exprimé le moindre mot. Son regard, son déplacement, son langage corporel ne laissent aucun doute sur le fait qu'il va s'adresser à nous. La reconnaissance de l'intention est une compétence de base acquise très tôt par les nourrissons [92]. La perception de l'attention des autres est cruciale pour la maîtrise des interactions sociales. Dans son étude, *Knight* [103] souligne l'importance pour un robot de détecter les intentions pour anticiper les objectifs de son partenaire humain. Symétriquement, le robot transmet ses intentions comme clarification de son activité. En apprenant de l'engagement des humains, le robot doit être capable d'apprendre le moment adéquat où il peut s'engager dans une interaction. Le rapport final du DARPA/NSF sur l'interaction homme-robot [104] préconise de questionner l'intentionnalité.

Dans nos travaux, nous nous intéressons à la phase d'engagement au cours de laquelle les humains expriment inconsciemment leur intention d'interagir. Notre but est d'étudier les techniques de détection et de reconnaissance des signaux sociaux reflétant l'intention d'un utilisateur d'interagir avec un robot. Pour *Sidner et al.* [105], l'engagement est défini comme le processus conjoint par lequel plusieurs participants établissent, maintiennent et terminent une interaction. Ils proposent un modèle en trois étapes : (1) l'initiation de l'interaction ; (2) le maintien de l'interaction ; (3) le désengagement. Nous nous consacrons à l'initiation de l'interaction que nous qualifions « d'intention d'interagir ». La détection de cette intention d'interagir est un vrai challenge, surtout lorsqu'il s'agit d'environnements comme le lieu de travail ou la maison, où les gens ne sont pas habitués à interagir avec des robots [106]. *Tahboub* [107] voit l'identification de cette intention comme une condition *sine qua non* à la coordination et à la coopération lors d'une interaction.

Dans la littérature, cette intention d'interagir avec un robot est souvent décrite comme une fonction proxémique essentiellement basée sur la distance. Dans [108], il est proposé d'utiliser les mouvements relatifs d'un humain avec le robot comme référentiel. Les intentions telles l'initiation de l'interaction et le désengagement sont décrites en utilisant des Modèles de Markov Cachés (MMC ou *Hidden Markov Models* - HMM) et des modèles dépendants de la position. D'autres études connexes ont étudié la distance interpersonnelle [109, 110]. Ces

travaux ne semblent pas suffisants pour estimer l'engagement dans un domicile. Prenons le cas où l'utilisateur croise le robot dans un couloir, les informations de proximité ou de vitesse ne permettent pas de déduire son désir d'interaction.

Dans nos travaux, nous proposons une approche multimodale pour détecter l'intention d'interagir. La multimodalité et les informations de posture ont donné de bons résultats dans la mesure de la qualité de l'interaction humain-robot et de l'engagement de l'utilisateur pendant l'interaction [111]. Notre première hypothèse est que la multimodalité sera, à l'identique, plus performante que les informations spatiales seules pour la détection d'interagir. Notre seconde hypothèse est que les caractéristiques posturales seront importantes parmi ces caractéristiques multimodales. Notre objectif est d'étudier l'intention d'interaction en utilisant le maximum de codes sociaux (voir fig. 10) et de déterminer les plus pertinents. Cette recherche prospective vise à construire un ensemble de caractéristiques des signaux sociaux utiles pour la description, la reconnaissance et la discrimination de l'intention d'interagir.

Figure 11 : Kompaï de la société Robosoft[11], notre robot d'étude. Ce robot est équipé d'un télémètre laser, de capteurs de distance infrarouges et ultrasoniques, d'une tablette tactile et d'une webcam. Dans le cadre de nos travaux, un capteur RGBD (Kinect) a été ajouté.

a) Traitement des signaux sociaux

Parmi les signaux sociaux évoqués en introduction de ce chapitre, nous avons cherché une couverture maximum des codes sociaux (fig. 10). S'inspirant des travaux antérieurs, notre

[11] La société s'appelle maintenant *Kompaï Robotics* (voir https://kompai.com/, dernière visite 04/2018).

objectif est de sélectionner des caractéristiques calculables sur notre plateforme robotique mais également réutilisables dans d'autres situations et/ou avec d'autres capteurs semblables. Notre robot, le Kompaï, est décrit sur la figure 11. Il dispose de capteurs permettant de calculer des caractéristiques spatiales, faciales et vocales. L'ajout d'un capteur RGBD a étendu ses capacités de perception aux caractéristiques posturales. Rappelons ici que les dernières avancées en apprentissage profond sur la détection de visage, d'émotion et de personnes n'étaient pas disponible lors de notre étude.

Figure 12 : Exemple de perception embarquée sur notre robot. À gauche, la vue du télémètre laser avec la détection des 3 piétons entourant le robot (ellipses). Au centre la vue RGB, à droite la caméra de profondeur.

Caractéristiques spatiales, faciales et vocales

Les caractéristiques spatiales que nous calculons sont conformes à ce que nous avons trouvé dans la littérature [108]. Nous avons cependant apporté une contribution à la détection de personnes[12] en utilisant le télémètre laser (LIDAR) [29] d'un robot compagnon. À partir de ces informations, le détecteur identifie les probables jambes dans l'environnement. Fortement bruitée en environnement intérieur[13], cette détection sert d'entrée au détecteur à proprement dit. Employant la vitesse et, de manière plus originale la distance inter-jambe, celui-ci réalise l'appariement des pied deux-à-deux pour créer puis suivre les personnes autour du robot. Cette distance inter-jambe est pratiquement constante lors de la marche ce qui augmente la robustesse de notre algorithme. Inspiré des travaux de [112], nous avons créé un modèle de marche intégrant les occlusions temporaires de pied. Si l'on regarde la figure 12, à gauche nous trouvons 3 détections (ellipses) qui représentent les 3 personnes entourant le robot. L'ellipse verte indique celui que le robot considère comme son partenaire d'interaction. Bien que 2 personnes ne soient pas visibles dans le champ des caméras embarquées et que l'une d'elle (à droite) soit de profil et ne laisse paraitre qu'une seule jambe, leur détection est correcte. Après avoir détecté les personnes entourant le robot, nous calculons pour chacun d'eux leur position et leur vitesse dans le référentiel du robot. Les caractéristiques spatiales sont complétées par la localisation acoustique (azimut) des sons dans l'environnement.

[12] Une vidéo est disponible https://www.youtube.com/watch?v=wEGekBQe0cg (dernière visite 04/2018).

[13] De nombreux éléments sont détectés comme des pieds de table ou de chaise par exemple.

Concernant les caractéristiques faciales, nous sommes confrontés aux limites de nos capteurs. Ainsi, même s'il est possible d'estimer la direction du regard avec une caméra [113], la distance des utilisateurs autour du robot est trop importante pour avoir une mesure fiable. Notre système réalise une détection du visage dans l'image en considérant également sa taille comme un indicateur de proximité. Pour les caractéristiques vocales, nous nous sommes limités au calcul de la détection de parole (voir section « Détection de parole »).

Caractéristiques de postures

La principale originalité de notre contribution à la détection d'engagement réside dans notre usage de caractéristiques de posture, indices sur l'intention d'interaction. Ces caractéristiques ont montré leur capacité à mesurer le niveau d'engagement d'un utilisateur à réaliser une tâche comme dans [114] avec une mesure de l'angle d'inclinaison du corps. Les psychologues ont proposé de nombreux modèles pour décrire les postures et leur signification [115]. Dans [116], les auteurs proposent un modèle spatial couplant l'estimation de l'attention et les métriques de distance pour un robot réceptionniste afin de déduire les intentions de l'humain en face à face, ce qui est la limite de leur approche. L'étude des relations spatiales [117] dans l'interaction entre robots et humains a conclu que les mesures proxémiques entre humains et les arrangements spatiaux tels que le système de distance interpersonnelle ne suffisent pas pour obtenir un comportement socialement approprié. Des psychologues tels que *Hall* [118] et *Schegloff* [27] ont proposé des mesures de posture mais il n'y a pas de consensus sur un modèle particulier en perception par ordinateur. La posture était difficile à mesurer et à évaluer à l'aide de la vision par ordinateur. Néanmoins, avec l'arrivée des caméras de profondeur à bas coût comme la *Kinect* sortie en 2010, ces mesures s'en sont trouvées facilitées.

Nous avons choisi les caractéristiques posturales proposées par Schegloff dans son livre [27]. Ce choix est guidé par le fait que celles-ci n'avaient pas été encore utilisées à notre connaissance dans ce contexte et par notre capacité à les calculer à partir des capteurs de notre robot Kompaï (voir fig. 11). Ces caractéristiques ont pour but de représenter la pose du corps. L'accent est mis sur les positions et les rotations relatives des pieds, des hanches, du torse et sur l'orientation des épaules vers le partenaire d'interaction.

Pour remettre en perspective ces caractéristiques, nous pourrions aujourd'hui bénéficier d'avancées dans la détection de personnes [119, 120] (voir la section « Détection de personnes ») et même de reconstruction de poses 3D sans avoir recours à des capteurs de profondeur [121, 122, 123] pour les estimer.

b) Expérience et résultat

Pour questionner nos hypothèses, nous avons recueilli un corpus dans un appartement expérimental. Il est conçu autour de 3 espaces : une partie dinatoire, un salon et un espace cuisine. Comme le montre la figure suivante, cet appartement a une forme de L.

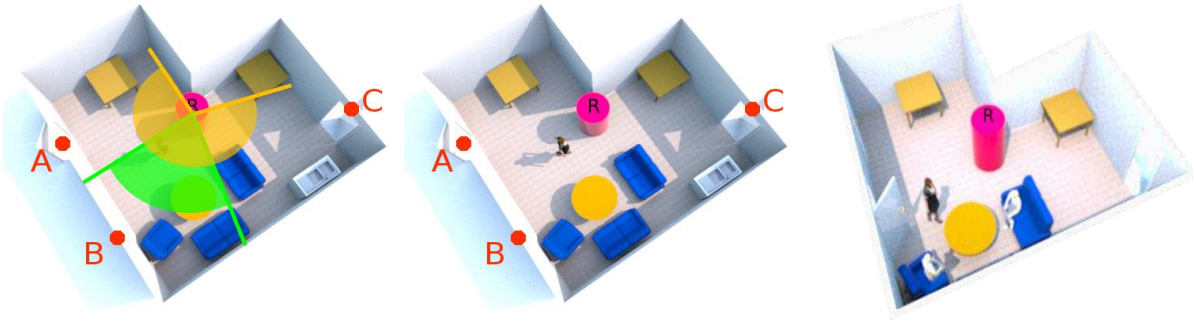


Figure 13 : Présentation de l'environnement de test (à gauche) et des 2 scénarios de nos expérimentations. Le robot est représenté par le cylindre R. Le champ de vision du télémètre laser est présenté en orange, celui de sa Kinect en vert.

Deux scénarios différents sont réalisés par un ou plusieurs participants dans cet espace où le robot *Kompaï* est présent. Le premier scénario est dédié à l'expérience mono-utilisateur. Une seule personne est présente dans la pièce. Le scénario multi-utilisateurs s'intéresse à une condition plus difficile. Trois personnes sont déjà dans la pièce et interagissent entre elles (jeux de cartes, discussion informelle, …). L'idée derrière ce second scénario est de vérifier si nous pouvons distinguer l'engagement envers le robot des interactions sociales entre les participants. Chaque participant reçoit au hasard une ou plusieurs actions à exécuter dans la salle. L'une des actions consiste à interagir avec le robot. Les autres actions consistent à traverser la pièce d'une porte à l'autre, à s'asseoir, à jouer aux cartes ou à se verser de l'eau à l'évier par exemple.

Classe	Précision	Rappel
NO_ONE	0,93	1,00
SOMEONE_AROUND	0,99	0,84
WILL_INTERACT	0,95	0,93

Tableau 3 : Performance de notre système de détection d'engagement avec des SVM et des données multimodales.

Le corpus est annoté en 3 classes dans la version finale de notre système [29] : NO_ONE SOMEONE_AROUND, WILL_INTERACT (des exemples de ce que perçoit le robot sont disponibles en annexe). Celles-ci représentent respectivement 3 phases : personne n'est présent, il y a un ou plusieurs humains dans l'appartement, un humain s'approche du robot avec l'intention d'interagir. Nous avons entraîné des réseaux de neurones et des *support vector machines* avec nos données. Les résultats présentés dans le tableau 3 correspondent à une classification via SVM plus performantes. Au niveau global, le système est capable de distinguer nos classes avec de très bonnes performances. Le résultat le plus intéressant est la performance dans la détection d'intention d'interagir (WILL_INTERACT) avec une précision de 95 % et un rappel de 93 %.

Nous avons également cherché à savoir parmi toutes nos caractéristiques quelles étaient les plus pertinentes. À partir de nos données, nous avons utilisé plusieurs méthodes de sélection de caractéristiques. Nous avons expérimenté, pour la première fois dans ce contexte à notre

connaissance, une méthode provenant des recherches en génétique : *Minimum Redundancy Maximum Relevance* [124] (MRMR). À l'aide de calculs d'information mutuelle, de corrélation et de t-test/F-test, l'algorithme sélectionne un sous-ensemble de caractéristiques minimisant la redondance, maximisant la dissimilarité des caractéristiques et les performances de classifications. Sur notre corpus, MRMR est l'algorithme qui a montré les meilleurs résultats.

Pour évaluer notre ensemble de caractéristiques, nous avons progressivement réduit sa taille et nous avons évalué à chaque fois notre système. Sans perte significative de performance, il est possible de réduire l'espace à 7 caractéristiques seulement, montrant une forte redondance de nos caractéristiques. Ce résultat est intrinsèquement intéressant car il permet de limiter le besoin de puissance de calcul sur les plateformes robotiques. Si l'on s'intéresse à la sémantique des caractéristiques sélectionnées, nous pouvons remarquer que la caractéristique la plus importante pour notre tâche de détection d'engagement est l'angle entre les épaules de l'humain et le robot. Cette métrique est directement inspirée des travaux de *Schegloff* [27], celui-ci mentionnant que l'orientation des épaules est l'un des indices clés de l'engagement. Les informations spatiales provenant des flux audio (localisation de source) et des télémètres sont également importantes. La position et la taille des visages détectés dans la vidéo confirme que faire face au robot est un signe de l'engagement.

Les résultats présentés dans cette section appuient nos hypothèses sur la reconnaissance multimodale de l'intention d'interaction avec un robot :

- il existe des signaux sociaux subconscients exprimés par les humains qui caractérisent leur volonté d'interagir **avec un robot** et ces signaux sont détectables par celui-ci ;
- certaines caractéristiques issues des recherches en sciences sociales et cognitives sont calculables sur un robot compagnon avec des capteurs standards ;
- dans notre expérimentation, 7 caractéristiques sont suffisantes pour la détection d'intention d'interagir ;
- d'autres résultats (non détaillés dans ce manuscrit, voir [29]) confirment que les approches spatiales souvent proposées dans la littérature sont significativement moins performantes en environnement domestique.

B.2) **Retour émotionnel du robot**

Dans la boucle d'interaction entre un robot compagnon et un humain, après nous être intéressés à la perception des intentions de l'humain par le robot (section précédente), nous nous intéressons dans cette section à l'expression d'un rendu « émotionnel » du robot pour augmenter son acceptabilité et l'engagement du partenaire humain.

Dans différents travaux [102, 125, 126], la forme du robot et ses mouvements sont apparus d'une grande influence sur le ressenti des utilisateurs à propos du robot. Plusieurs expériences ont démontré que même des appareils simples, comme les robots aspirateurs, peuvent susciter de l'empathie chez les humains [127]. Cette tendance à "anthropomorphiser" est encore plus

importante avec les robots humanoïdes et augmente les attentes de performance globale [94]. Le principal problème étant que trop d'attentes peut amener à une déception, voire un effet « vallée de l'étrange » (« *Uncanny Valley* » [128]), donc un rejet du robot. Au cours de la dernière décennie, au fur et à mesure des avancées en robotique, un nombre croissant d'études se sont penchées sur l'impact de l'aspect et des attitudes du robot. De telles études utilisent souvent des techniques « magicien d'Oz » [129, 130] (*"Wizard-of-Oz"* - *WoZ*). Elles sont généralement complétées avec des interviews a posteriori des utilisateurs pour questionner des aspects clés de l'expérimentation.

De nombreux robots compagnons sont décrits dans la littérature [102, 126, 131, 132, 133, 134, 135, 20]. Ces robots peuvent être regroupés en 3 catégories. Le premier groupe comprend les robots animaloïdes. Historiquement, l'*Aibo*, qui ressemble à un chien, a été le premier robot animal commercialisé pour le grand public, les premiers modèles arrivant sur le marché en 1999. Des exemples plus récents de cette catégorie comprennent *iCat*, *Paro* (un bébé phoque) et *Pleo* (un dinosaure). La deuxième catégorie comprend les robots utilitaires, généralement dotés de certains attributs anthropomorphes tels que le regard, des mouvements de tête ou des expressions faciales. Dans cette catégorie, on trouve *Sparkly*, *IROMEC*, *Pearl*, *Robotcare*, *Care-o-bot*, *Max* du projet Serroga ainsi que le *Pepper* d'*Aldebaran*. La dernière catégorie comprend les robots humanoïdes tels *NAO* et *ROMEO*, *KASPAR*, un petit robot enfant, et la série *Geminoid*, une série de robots humanoïdes utilisés pour la téléprésence. Des études réalisées avec ces robot indiquent que la tête, le regard, et les mouvements ont un impact sur la présence sociale et sur l'acceptation du robot [133, 136].

Kompaï, notre robot d'expérimentation fait partie de la seconde catégorie (voir fig. 11). S'il ne possède pas de bras ni de jambes, il n'est pas dénué d'attributs anthropomorphiques. Sa tête boule avec des yeux et sa taille (1,44 m) en sont d'importants marqueurs. Dans une moindre mesure, la forme de son corps l'est aussi. Dans le projet PRAMAD, nous nous sommes interrogés sur l'intérêt d'ajouter un retour « émotionnel » à *Kompaï*. Ces émotions s'avèrent utiles dans l'interaction avec les personnes âgées/fragiles. Elles servent de retour périphérique pendant des phases de jeux sérieux pour indiquer le bon déroulement du jeu (joie) ou une réussite sur une difficulté (surprise). La tristesse permet de générer de l'empathie si aucune interaction n'est intervenue récemment. La colère, lors d'interactions avec des personnes souffrant de troubles cognitifs, marque un interdit si par exemple la personne s'en prend physiquement au robot. En collaboration avec nos collègues cliniciens et psychologues, nous avons sélectionné pour notre robot 4 des 6 expressions basiques décrites par *Ekman* [137] : joie, tristesse, surprise, colère. Dans notre contexte, nous n'avons pas considéré la peur et le dégout comme pertinentes.

Après la conception de notre tête robotique et le design des visages associés, nous avons conduit des expérimentations à l'Hôpital *Broca* de Paris pour évaluer cette modalité émotionnelle de notre robot.

a) Conception de la tête robotique

La conception de notre tête robotisée s'est inspirée des robots utilitaires décrits dans la littérature. La première inspiration est venue de *Sparkly* [126], un petit robot mobile avec une tête expressive. En changeant la forme des fils métalliques sur son visage, il peut exprimer plusieurs émotions. Le robot *IROMEC* [138] est similaire mais un écran remplace les fils métalliques. Les conceptions faciales des deux robots sont facilement compréhensibles avec leur conception smiley. *Care-o-bot* est un robot compagnon plus massif comparable au robot *Kompaï*. Il n'est pas anthropomorphe mais il a un torse mobile. Dans les travaux sur *Care-o-bot* [139], l'attention de l'utilisateur était attirée par la rotation de son torse pour simuler la direction du regard. Les résultats ont démontré que les participants appréciaient ce mode de communication.

Le design mécanique de notre tête robotique est inspiré de ces robots et des mouvements dont les humains sont capables. Nous avons construit une structure à 4 degrés de liberté (fig. 14) supportant une tablette. Cette structure est capable de tourner sur elle-même, d'avancer, de reculer en imitant les mouvements du cou humain. Il est possible d'incliner la tablette sur le côté. La tablette est orientée soit horizontalement, soit verticalement et présente un visage animé. L'intérêt du couple cou mobile/tablette est multiple. Premièrement, les capteurs de la tablette sont indispensables dans certaines tâches de perception (suivi de visage, recherche de personnes tombées à terre[14], …). La mobilité du cou permet à cette tête d'aligner

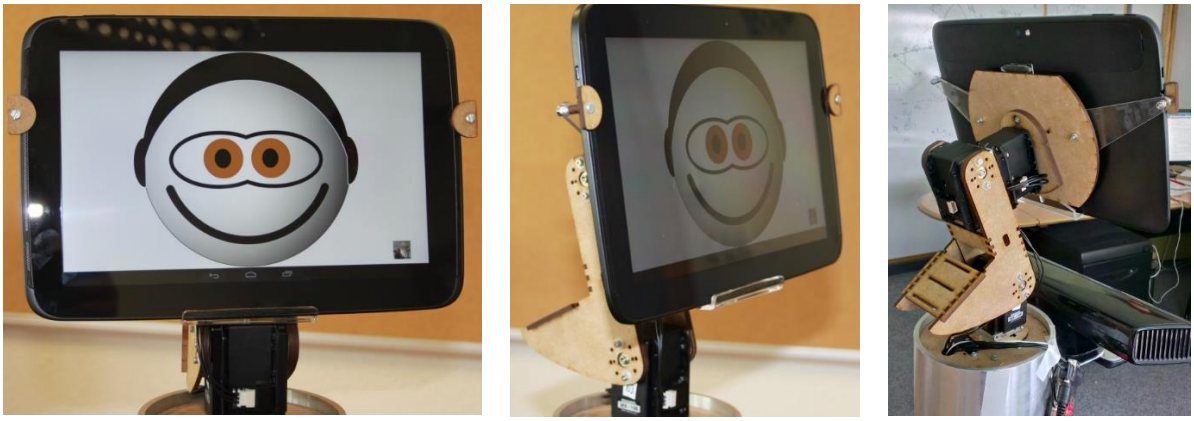


Figure 14 : Design mécanique de notre tête robotique.

[14] Cette tâche de recherche de personnes tombées à terre est très importante dans le cadre du maintien de personnes âgées à domicile. Elle fait partie des prérequis pour la sécurité. Nous travaillons sur cette thématique via le projet Mobile RGBD.

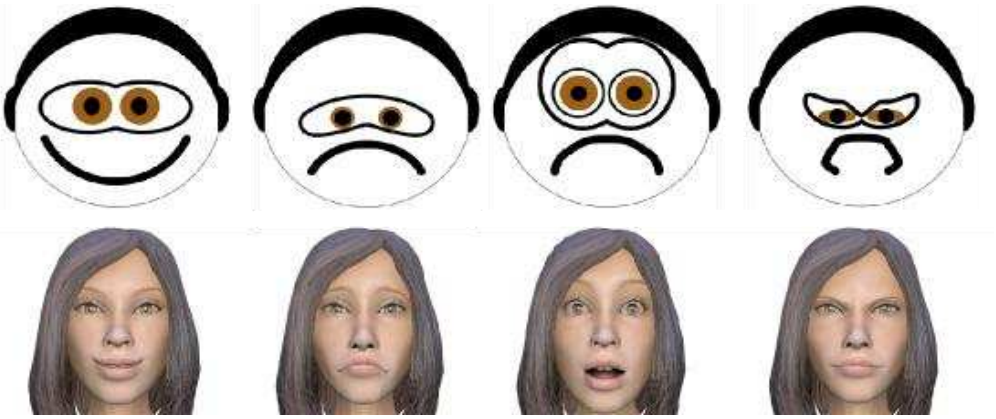


Figure 15 : Design des expressions de nos 2 visages : Kompaï et Louise. De gauche à droite : joie, tristesse, surprise et colère (source [140]).

son visage avec celui de son partenaire humain automatiquement, i.e. de le « *regarder dans les yeux* » quelle que soit sa taille. Enfin, pour exprimer une émotion en fonction de la situation courante, la tête effectue un mouvement tout en jouant une animation sur la tablette.

Pour le design du visage animé sur la tablette, nous avons comparé une approche smiley comme celle utilisée avec *Spackly* et *IMOREC* avec une approche d'agent virtuel. Le design du smiley *Kompaï* s'inspire du design original de la tête boule du robot. *Louise*, l'avatar utilisé dans cette étude, est un personnage virtuel provenant d'un projet d'agent conversationnel d'assistance [140]. La figure 15 montre les expressions faciales de *Kompaï* et de *Louise* pour les expressions joie, tristesse, surprise et colère. Ces visages ont fait l'objet d'une évaluation interne au projet Pramad auprès de 156 personnes qui a montré leur pertinence sur un écran [141].

b) Évaluation et discussion

Les évaluations de la pertinence de ce retour émotionnel ont été réalisées avec des personnes âgées au *Living Lab* de l'Hôpital *Broca* de Paris par deux psychologues, un ergonome et trois chercheurs en informatique [141]. La première phase a consisté en une évaluation collective autour d'une discussion sur le thème « Robots et expressivité ». Lors de cette session, après une présentation du robot aux participants, des échanges ouverts sur les prérequis d'une tête expressive pour l'interaction robot-homme ont été retranscrits. À la fin de la session, les participants évaluaient l'expressivité de la tête robotique avec une échelle de valeur : l'expression du robot est correcte, proche ou elle ne correspond pas. La seconde phase d'évaluation était individuelle. Elle était conduite en magicien d'Oz, un humain contrôlant à distance les mouvements. 28 participants ont noté sur la même échelle de valeur la façon dont ils percevaient chaque émotion dans des interactions avec le robot. Ces participants ont testé différentes conditions : visage Kompaï ou Louise, avec et sans animation mécanique. Ils ont également comparé ces conditions avec la tête boule en plastique d'origine.

De ces expérimentations, nous pouvons tirer différents enseignements concernant l'intérêt des visages et des animations robotiques pour l'expression de retour émotionnel. À plus de 90 %, les participants ont préféré la tête tablette à la tête boule originale, avec une légère préférence pour le visage smiley. D'autres part, la reconnaissance des expressions faciales par les personnes âgées est plus facile avec un visage "smiley" plutôt qu'avec un personnage humain virtuel en 3D. Le principal enseignement de notre étude concerne l'intérêt de l'animation robotique. Celle-ci maintient la pertinence de l'expression mais aucune différence statistiquement significative entre les conditions avec ou sans animation de la tête n'est constatée.

Ces conclusions ont soulevé plusieurs questions après notre étude. Quel mouvement du cou associé à une expression faciale virtuelle améliore l'expressivité du robot ? Y a-t-il des attentes sur l'animation du cou en fonction de la représentation virtuelle du visage (*Kompaï* ou *Louise*) ? La vision anthropomorphique de notre design de tête n'est-elle pas un problème ? Ces questions sont centrales et doivent être prises en compte dans un futur design d'un robot compagnon. Notamment, il faut s'interroger sur la pertinence de l'animation mécanique de la tête même si l'intérêt pour l'alignement de visage ou encore pour la recherche de personnes tombées à terre est indéniable. Les travaux de thèse d'*Étienne Balit* sur l'animation des robots étudient ces questions [19, 20].

III.C Interaction avec des véhicules autonomes

C.1) Espaces partagés en centre-ville

Parmi les robots disponibles de nos jours, les véhicules autonomes se développent rapidement. Malgré la jeunesse de ces technologies, des produits commerciaux destinés au grand public existent déjà. Nous pouvons citer les voitures *Tesla* [7] et leur mode de conduite autonome ou encore les nombreuses technologies d'assistance, comme la détection et l'évitement d'obstacles, présentes dans de plus en plus de véhicules de série. De nombreuses recherches ont été menées dans des environnements de conduite sur piste ou sur autoroute [142]. Leur point commun étant d'être des environnements où l'on ne trouve peu ou pas d'humains avec lesquels ces véhicules interagiraient. De récentes recherches adressent l'usage de véhicules autonomes en condition urbaine, notamment dans les centres-villes [143]. Ces conditions citadines impliquent une interaction importante entre les usagers de la route, notamment entre les piétons et les véhicules.

Figure 16 : Vision future des centres-villes partagés entre les usagers (source [136]).

Des urbanistes envisagent le futur des centres-villes comme proposant des espaces partagés non contraints où les véhicules, les piétons, les cyclistes évolueraient en bonne intelligence. Cette vision est illustrée notamment dans les travaux d'Hamilton (fig. 16). Si l'on pousse cette vision plus loin en intégrant à ce schéma des véhicules autonomes, un nouveau pan de recherche s'ouvre. Le concept de « *situation awareness* » [7] prend ici tout son sens. Le véhicule doit percevoir, modéliser et prédire l'évolution de son environnement et des entités qui le peuplent. Parmi ces entités, les plus fragiles sont les humains, il est donc primordial de s'intéresser à doter les véhicules autonomes de capacités d'anticipation des comportements des piétons en centre-ville.

C.2) **Modélisation et prédiction du comportement des piétons**

Les véhicules autonomes sont dotés de nombreux capteurs leur permettant de percevoir leur environnement : caméras, télémètres laser (LIDAR), Velodyne... Cette perception embarquée, que nous pouvons également appeler perception égocentrique, donne des informations sur la localisation et l'environnement immédiat du véhicule [144, 145]: voies de circulation, panneaux de signalisation, objets mobiles, etc. Cette perception égocentrique souffre de différents problèmes [146]. Le problème des occlusions, problème récurrent en vision par ordinateur, en est l'un des principaux. Les solutions déployées sont identiques à celles que l'on trouve dans la littérature sur les espaces perceptifs multimodaux [147] : fusionner des informations provenant de différentes sources. La communication entre véhicules (autonomes) permet d'échanger des données pour construire une représentation intégrant plusieurs points de vue [148]. Enfin, dans les centres-villes, le développement des caméras de surveillance favorise une large couverture de l'espace public [84]. Cette « exo-perception » complète avantageusement les informations des véhicules grâce à son point de vue plongeant. Elles

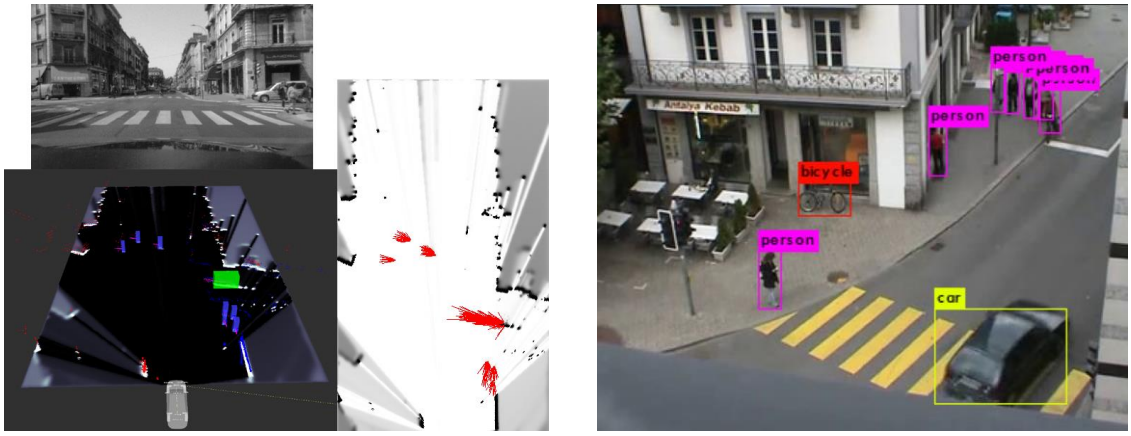


Figure 17 : Perception égocentrique et exo-perception pour les véhicules autonomes en centre-ville.À gauche, la perception embarquée sur un véhicule autonome (source [141]). À droite, la perception distribuée dans l'environnement (source [15]) La détection d'objets est réalisée grâce au système Yolov3 [142].

facilitent l'utilisation d'approches gourmandes en ressources de calcul comme l'apprentissage profond.

La figure 17 présente des exemples de perception égocentrique et d'exo-perception. À gauche, la perception embarquée d'un véhicule autonome avec détection d'obstacles [149]. À droite, une caméra placée au-dessus d'un carrefour [150] fournit des informations sur les piétons et les véhicules présents. Les algorithmes disponibles pour traiter cette vue sont nombreux [151].

La modélisation des comportements et des itinéraires piétonniers a été étudiée à de multiples reprises [152] en utilisant différentes méthodes. Certaines approches s'intéressent à la vitesse ou au regroupement de piétons [153], d'autres modélisent les forces sociales entre les individus [154]. Des approches stochastiques sont souvent employées pour apprendre et prédire des trajectoires [155, 156, 157, 158, 159]. Le point commun de toutes ces méthodes est l'apprentissage sur des données préalablement observées avant de pouvoir réaliser des prédictions.

Dans les travaux que nous développons dans la thèse de Pavan Vasishta, nous nous attelons à fournir un modèle d'analyse de comportement piétonnier ne nécessitant pas de trajectoires d'apprentissage. Il est nécessaire de percevoir l'environnement autour du véhicule (voies de circulation, passages piétons, ...) mais des algorithmes sont disponibles [145, 160, 161]. Ces informations peuvent être complétées par des données géo-localisées disponibles en ligne (magasins, présence de travaux sur la chaussée, etc.). Cela permet une application à des environnements dynamiques, ce que sont les centres-villes et leurs futurs espaces partagés.

Dans la littérature, *Gibson* [21] a décrit le principe de « vision naturelle » (« *Natural Vision* »). Dans ce modèle, les piétons se déplacent suivant la direction de ce qui les intéresse le plus dans leur champ de vision. Cette théorie a été reprise et étendue dans le modèle de

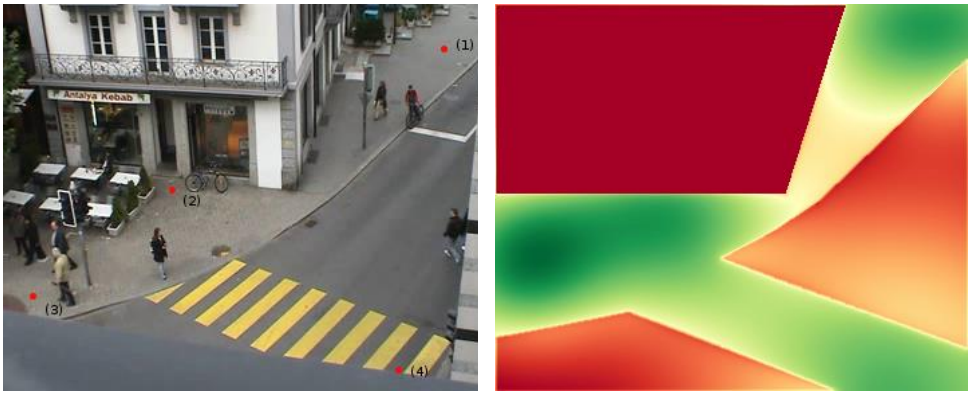


Figure 18 : Vue d'un carrefour et du champ de potentiel correspondant (source [15]).

« mouvement naturel » [162] (« *Natural Movement* »). L'environnement contient des attracteurs et des répulseurs qui régissent les comportements des piétons. Ceux-ci sont dépendants de la sémantique des éléments présents. Des points d'intérêts (*Points Of Interest* - POI) comme par exemple les entrées de magasin, les arrêts de bus ou les monuments sont des attracteurs pour les piétons, des travaux sur un trottoir sont des répulseurs.

Comme dans les travaux de *Khatib* [163] et *Wolf* [164], nous utilisons un champ de potentiel pour modéliser les forces présentes dans l'environnement du véhicule. Le modèle résultant est la somme des potentiels de toutes les composantes de l'environnement : passages piétons, routes, trottoirs, bordures des voies, objets mobiles, points d'intérêts, etc. Les équations ci-dessous décrivent les calculs de potentiel pour les routes et leurs bordures.

$$\mathbf{U}_{Road}^{ij} = \beta_{Road} \exp\left(-\left[\left(\frac{x_{ij}-x_{road}}{\sigma_x}\right)^2 + \left(\frac{y_{ij}-y_{road}}{\sigma_y}\right)^2\right]\right) \quad (3) \qquad \mathbf{U}_{\mathbf{Edge}}^{\mathbf{ij}} = \frac{1}{2}\eta\left(\frac{1}{\rho(x^{ij}, y^{ij})}\right) \quad (4)$$

D'autres équations décrivant l'impact des POI et des passages piétons sont décrites dans notre article [23]. Les obstacles, statiques ou dynamiques, sont modélisés avec un potentiel de *Yukawa* [165].

Sur une vue externe de l'environnement au-dessus d'un carrefour à Martigny en Suisse [150], le résultat de ce calcul de potentiel est présenté sur la figure précédente. Certaines zones sont remarquables. Ainsi le trottoir rétréci à l'angle du bâtiment se retrouve impacté par la présence de la route. Du fait de la somme de toutes les composantes, la route n'a pas un potentiel uniforme, notamment à cause de la présence du trottoir et du passage piéton. Cette proposition de modèle à base d'un champ de potentiel a montré sa capacité à expliquer les trajectoires des piétons [22, 23], sans recourir à un apprentissage de trajectoire.

Disposant d'un modèle de l'environnement, la prédiction de trajectoires des piétons est la dernière tâche à accomplir pour nos véhicules autonomes. La littérature propose différentes approches markoviennes à cette problématique. Des MDP et des MDP partiellement observables (*Partially Observed Markov Decision Processes - POMDP*) sont couramment

employés avec une fonction de coût sur l'environnement [155, 156, 157]. D'autres travaux [153] prédisent la destination des piétons selon le même principe. Des Modèles de Markov Cachés Expansifs (*Growing Hidden Markov Models – GHHM*) sont proposés par [166, 159]. Comme nous l'avons déjà mentionné, ces approches requièrent des trajectoires d'entraînement. Dans le cas de l'utilisation d'une exo-perception, cela ne poserait pas de problème, les données pouvant être collectées au cours du temps. Dans le cas d'une perception égocentrique, cela supposerait, avant que le véhicule ne puisse commencer à prédire le comportement des piétons, qu'il reste suffisamment longtemps au même endroit pour collecter des trajectoires d'apprentissage.

Dans nos travaux, nous proposons une extension des GHMM basée sur notre modélisation de l'environnement. Le champ de potentiel sert d'initiateur pour la construction d'un modèle de trajectoires a priori. Ce modèle est ensuite mis à jour en fonction de ce qui est observé autour du véhicule. De nœuds sont ajoutés ou retirés automatiquement grâce à la mise à jour du champ de potentiel dans différents cas : le véhicule autonome se déplace, les obstacles mobiles se déplacent, des trajectoires de piétons sont observées. Au moment de la rédaction de ce manuscrit, ces travaux sont en cours d'évaluation et de comparaison avec l'état-de-l'art.

En conclusion, notre contribution à l'interaction véhicule autonome/piétons s'est limitée pour l'instant à la partie modélisation et prédiction des intentions des piétons. La perception autour du véhicule autonome est l'une de nos préoccupations actuelles (voir « Détection de personnes »). Enfin, pour compléter la boucle d'interaction, nous verrons dans ce manuscrit que la communication du véhicule autonome à destination des piétons, comme pour le retour émotionnel du robot, est questionnée dans notre projet de recherche.

Chapitre IV
Perception des humains

IV.A Présentation

Le besoin de percevoir des personnes et des informations sur celles-ci n'est pas une thématique nouvelle dans nos recherches. Comme nous l'avons vu dans les chapitres précédents, ce besoin est présent dans différents contextes avec différents capteurs.

Dans le cadre de l'interaction à courte distance, telles les interactions avec des interfaces sur écran, la connaissance que le système peut avoir de son partenaire humain peut être plus grande, du fait de la proximité de ce dernier. Les capacités de perception se font plus proches de l'humain et permettent d'estimer son état mental au sens large du terme : état émotionnel, compréhension, charge cognitive pendant une tâche [167, 168]. Les domaines de recherche associés à ces problématiques se nomment informatique affective (« *affective computing* ») et informatique comportementale (« *behavioral computing* ») [169, 170, 171].

Pour la perception des humains depuis un robot compagnon ou un véhicule autonome, le besoin de prédire les comportements et les activités de ceux-ci passe par la détection de personnes, de postures, de gestes et de comportements à plus grande distance [172, 173, 174].

Ce chapitre présente nos contributions sur ces thématiques. Dans la première section, nous présentons nos travaux sur la perception des humains en champ proche :

- l'enregistrement et l'analyse multimodale de sessions de narration ;
- l'analyse du comportement et des affects de joueurs résolvant des problèmes d'échecs.

Dans la seconde section, nous présentons nos travaux sur la détection de personnes :

- la construction du corpus MobileRGBD pour la détection de personnes autour d'un robot compagnon en environnement domestique ;
- nos travaux préliminaires sur l'apprentissage profond pour la détection de personnes.

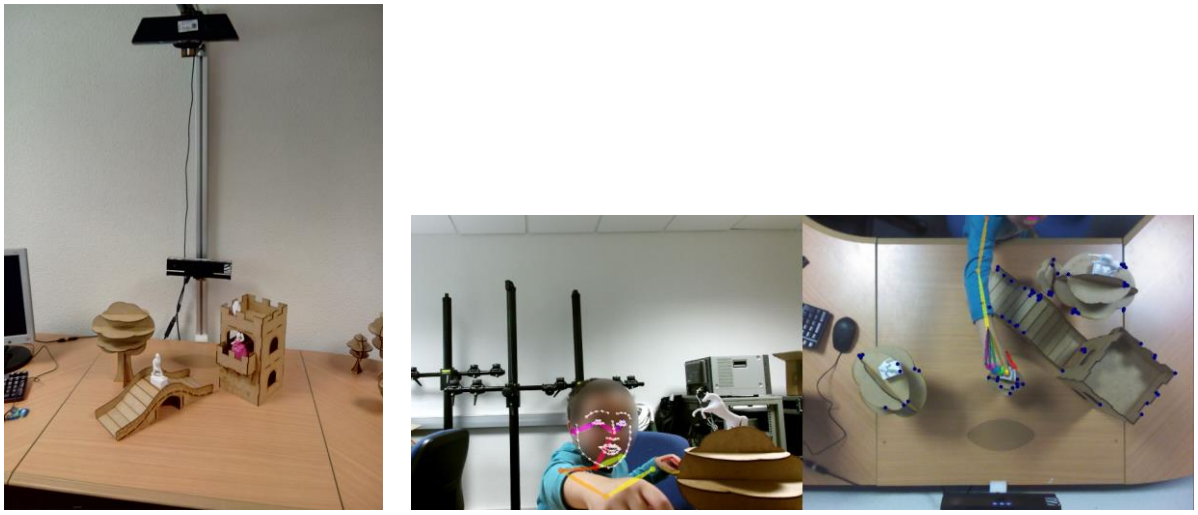


Figure 19 : Setup d'enregistrement du projet Figurines et vue des données acquises et calculées (source [24]). Le setup est composé d'une Kinect enregistrant le(s) conteur(s), d'une Kinect suivant les mouvements des éléments de décors sur la table et des figurines.

IV.B Perception des humains en champ proche

Dans cette section, nous présentons les travaux autours du projet Figurines concernant la perception de conteurs lors de sessions narratives. Nous verrons comment en adaptant ces travaux, nous avons pu travailler sur la perception des comportements et des affects de joueurs d'échecs.

B.1) **Perception de sessions narratives**

La narration (« *storytelling* » en anglais) est l'art de créer ou de partager un récit, une histoire. Il est particulièrement important pour l'éducation des enfants en soutenant et en améliorant l'expression créative et l'apprentissage [175], ainsi qu'en encourageant le travail d'équipe et le partage d'expériences personnelles [176, 177, 178].

Un bon conteur utilise ses talents expressifs pour émettre des signaux sociaux renforçant son récit [179] : posture, déplacements, amplification des gestes, changement de voix, expressions du visage. Dans le projet Figurines, nous nous sommes intéressé aux sessions narratives d'enfants ou d'adultes, le but étant de capturer ces signaux sociaux pour pouvoir produire automatiquement une animation 3D la plus fidèle possible. L'intérêt d'un tel système réside dans sa capacité à être un outil d'analyse des processus créatifs à l'œuvre pendant la narration d'une histoire.

Dans la littérature, de nombreux systèmes ont été développés pour capturer des sessions narratives [180]. Différentes interfaces tangibles existent pour aider les enfants à expérimenter la narration d'histoires. Certains systèmes autorisent la création de l'histoire en utilisant une interface avec des briques que les enfants manipulent [181], parfois sur une table augmentée [182]. Les possibilités de raconter l'histoire sont limitées par l'interaction et les structures

d'histoires disponibles. L'utilisation d'objets tangibles comme marionnettes a amélioré les possibilités d'animation tout en limitant souvent l'usage à cette seule marionnette [183, 184], le corps faisant parfois office d'instrument de manipulation [185, 186]. D'autres systèmes permettent d'utiliser des accéléromètres ou un capteur RGBD pour contrôler les figurines [187, 188].

En prenant inspiration de ces travaux antérieurs, nous avons construit un outil d'acquisition de sessions narratives. Pour accroître son caractère ludique et ne pas perturber les enfants conteurs, nous avons choisi d'être le plus écologique possible. Les sessions doivent ressembler aux activités de jeu habituelles des enfants. Les enfants choisissent les éléments de décor et les figurines avec lesquels il veulent raconter leur histoire. Ils organisent ces éléments comme bon leur semble et racontent leur histoire sans contrainte aussi longtemps que souhaité.

Cet outil est dépeint à gauche sur la figure 19. Les éléments de l'histoire disponibles sont des éléments habituels des contes de fées : princesse, prince, chevalier, animaux, château, pont, arbres, tours, etc. Les figurines sont équipées d'IMU haute fréquences (200 hz). Les objets du décor n'ont pas d'équipement particulier. À droite sur la figure, nous voyons les gestes, les postures et les expressions faciales du conteur ou des conteurs qui sont capturés. La voix l'est aussi. Les éléments du décor ainsi que les figurines sont suivis en 3D grâce au capteur RGBD et aux IMU (lors des occlusions).

Ce système a été utilisé pour l'enregistrement de sessions narratives avec des enfants et des adultes [32]. Cela a permis à nos collègues de l'équipe IMAGINE de l'Inria de produire des animations 3D en mappant les signaux sociaux exprimés par les conteurs sur les alter-egos numériques des figurines [30, 31].

Par la mise au point de ce système et des algorithmes de suivi sous-jacents, nous avons contribué à l'analyse de la narration, et nous l'espérons dans le futur, à l'analyse des processus créatifs des enfants conteurs. Ces travaux sont applicables avec quelques modifications à d'autres contextes. Dans la section suivante, nous présentons les travaux que nous menons avec un outil dérivé dans le cas de parties d'échecs.

B.2) **Perception de l'expertise de joueurs d'échecs**

a) Contexte

En plus de la manifestation d'émotions, l'état cognitif s'exprime également par le biais de signaux sociaux non verbaux. Les états mentaux peuvent être déduits des expressions faciales et des gestes de la tête et du corps [189, 190]. Peu de recherches ont été menées sur la manière dont ces signaux peuvent révéler des processus cognitifs tels que la reconnaissance de la situation, la (non) compréhension d'un problème ou l'engagement dans une tâche. Les signaux sociaux émis sont détectables à courte distance avec des capteurs grand public. Avec une webcam, les émotions se détectent à partir de micro-expressions faciales [191]. La fréquence cardiaque, indicatrice de stress, peut être mesurée à partir des variations de la couleur de la

peau du visage [192]. Le volume occupé par le corps permet d'inférer la dominance ou la soumission [136] en utilisant les données du capteur de profondeur.

D'autres signaux que nous n'avions pas encore exploités dans nos recherches passent par le regard. Les chemins par lesquels le regard passe (« *scan path* ») mais également les fixations, points sur lesquels le regard s'est attardé sont une trace ce que l'on nomme l'« attention visuelle ». Le suivi du regard est réalisé par un oculomètre[15] portable ou fixe [193]. Les informations portées par le regard sont employées à plusieurs fins. La variation de la taille des pupilles est un indicateur de concentration [194]. Les mouvements oculaires sont également révélateurs de processus cognitifs sous-jacents à la reconnaissance d'une situation [195, 196] ou à l'expertise sur une tâche [167, 168].

[15] Voir https://fr.wikipedia.org/wiki/Oculom%C3%A9trie (dernière visite 04/2018).

Figure 20 : Setup d'enregistrement du projet CEEGE (source [16]). Ce setup est composé d'une Kinect v2, d'une webcam, d'un écran tactile et d'un oculomètre fixe (sous l'écran). Deux lampes LED permettent de contrôler les conditions de luminosité.

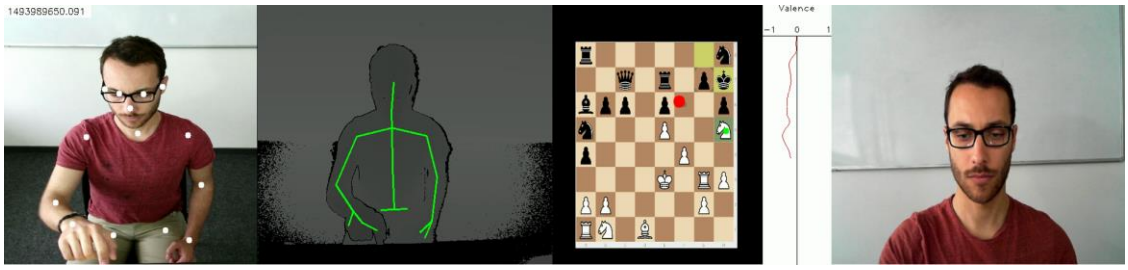


Figure 21 : Enregistrement d'un joueur d'échec lors de la résolution d'un problème (source [16]). De gauche à droite : vue plongeante et détection du corps ; caméra de profondeur et détection de squelette ; suivi du regard sur l'échiquier ; calcul de valence et d'excitation ; vue haute définition de face pour la détection d'émotions.

Dans le cadre du projet CEEGE[16], nous nous intéressons à mesurer l'expertise de joueurs d'échecs et à déduire des modèles de résolution des problèmes de ceux-ci. En quelques mots, nous cherchons à pouvoir répondre à plusieurs questions en collaboration avec nos collègues cogniticiens de Bielefeld (CITEC[17]). Le joueur est-il un expert ou non ? A-t-il correctement analysé la situation courante ? Se trouve-t-il dans une situation qui dépasse ses compétences ? Comment est analysé le jeu par les joueurs ? Quels seront les prochains coups joués ?

Le choix du jeu d'échecs n'est pas anodin. L'analyse de parties d'échecs a longtemps été utilisée en science cognitives comme support au développement de modèles [197]. De plus, le jeu d'échecs est connu et codifié depuis longtemps. Sa résolution complète à l'aide

[16] Voir http://ceege.inria.fr/ (dernière visite 04/2018).

[17] *Cluster of Excellence Cognitive Interaction Technology*. Voir https://www.cit-ec.de/en (dernière visite 04/2018).

d'algorithmes est maintenant ancienne. L'algorithme stockfish[18], permettant de simuler n'importe quel niveau de joueur d'échecs, en est l'illustration.

Dans les travaux de thèse de Thomas Guntz, nous travaillons à la capture et à l'analyse des signaux émis par des joueurs pendant la résolution de problèmes d'échecs. Dans leurs études [168, 198], *Charness et al.* ont montré que l'attention visuelle des joueurs d'échecs caractérise leur engagement et leur compréhension de la situation. Nos premiers travaux visent à reproduire ces résultats en complétant le suivi du regard par la capture de signaux sociaux posturaux, gestuels et faciaux.

b) Caractéristiques (signaux sociaux)

En bénéficiant de l'expérience acquise dans le projet Figurines, nous avons créé un instrument permettant d'enregistrer les signaux multimodaux des joueurs engagés dans la résolution de problèmes d'échecs. Les figures 20 et 21 présentent respectivement notre outil d'expérimentation et les données collectées. À partir de ces données, nous calculons des informations sur le joueur et son état mental. Comme nous l'avons déjà vu, la posture est importante [199] et nous l'intégrons avec le volume occupé par le corps et l'agitation corporelle, i.e. le nombre de parties du corps qui bougent. Nous dénombrons le nombre de *self-touching*[19] car celui-ci dénote un inconfort, une difficulté ou une frustration [200].

Concernant les émotions faciales, *Facereader*, l'outil que nous utilisons, présente une performance d'environ 90 % sur les corpus standards d'évaluation [201]. Cependant, en conditions réelles, ses performances baissent avec notablement des fluctuations dans la détection des émotions. Nous avons considéré trois paramètres : la valence, l'excitation (*arousal*) et le nombre de changements d'émotion détectés. Ce dernier paramètre nous permet d'être sensibles aux variations des micro-expressions du visage.

Concernant le suivi oculaire, nous nous sommes intéressés aux zones d'intérêt dans les problèmes que nous demandons de résoudre, comme par exemple la pièce qui va mettre en échec le roi au prochain coup. À partir du suivi du regard, nous calculons pour ces zones d'intérêt le temps et le nombre de fixations, et le nombre total de visites. *Ehmke et al.* [202] interprètent de longues fixations sur une zone d'intérêt comme une difficulté. *Reingold* and *Charness* [168, 198] montrent une différence entre les novices et les experts dans leur façon de regarder ces zones.

---

[18] Voir https://fr.wikipedia.org/wiki/Stockfish_(programme_d%27%C3%A9checs) (dernière visite 04/2018).

[19] Aucun équivalent satisfaisant français n'a été trouvé, nous avons gardé le mot anglais *self-touching*. Celui-ci indique le fait de se toucher la tête et le cou avec la main.

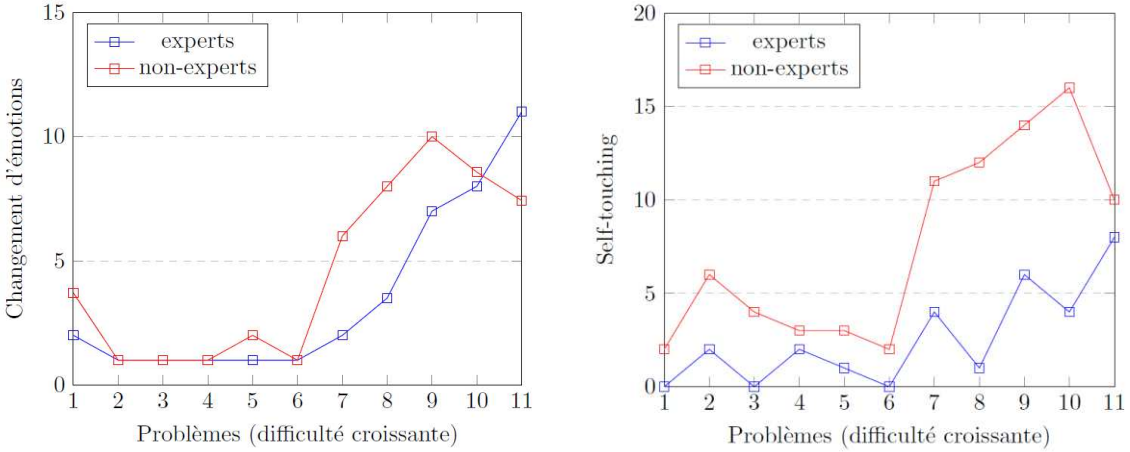


Figure 22 : Évolution du nombre moyen de changements d'émotion et de *self-touching* en fonction de la difficulté du problème de mise en *échec-et-mat* (source [17]).

c) Expérimentations et résultats

Durant nos expérimentations, 23 joueurs d'échecs ont été invités à résoudre des problèmes de difficulté croissante en temps limité. Ils commencent par deux ouvertures de jeu connues[20]. Dans les tâches suivantes, ils doivent faire *échec-et-mat* en *n* coups, *n* connu allant de 1 (facile) à 6 (difficile) [24, 25].

La figure 22 présente l'évolution de deux caractéristiques remarquables en fonction de la difficulté des problèmes soumis aux participants : la variation des émotions et le nombre de *self-touching*. Les courbes sur ces deux graphiques ont le même profil général. À partir du problème numéro 7, la difficulté s'accroit. Le groupe de non-experts éprouve instantanément des difficultés tandis que le groupe d'experts fait face à une difficulté progressant graduellement. Ces caractéristiques semblent, dans notre cas, dénoter de la difficulté qu'éprouvent les participants à résoudre les problèmes.

Nous avons étudié la classification des participants en 2 classes : les experts et les non-experts. Nous avons séparé nos caractéristiques en groupes : caractéristiques corporelles, visuelles et émotionnelles. Ces groupes et toutes leurs combinaisons ont été testés sur notre tâche de classification [24]. Les caractéristiques corporelles seules prédisent l'expertise avec une précision de 90 %, les caractéristiques émotionnelles à 86 % et les caractéristiques visuelles à 62 %. Les meilleures combinaisons obtiennent une précision de 86 %. Ces résultats montrent l'intérêt des caractéristiques corporelles et émotionnelles. Par contre, les caractéristiques visuelles ne comblent pas nos attentes. Une analyse plus complète de ces caractéristiques est disponible dans notre article [25].

[20] *King's Gambit* et une variation de la défense de *Caro-Kann*.

Par ces travaux, nous contribuons à l'analyse multimodale de joueurs d'échecs en phase de réflexion, et plus généralement à l'analyse de personnes faisant appel à leurs facultés cognitives. Les caractéristiques que nous avons utilisées sont pour certaines inédites dans ce contexte. Ces premiers résultats vont nous permettre d'étudier plus en profondeur les mécanismes de réflexion des joueurs, notamment les phases de reconnaissance de situations [203, 204, 205]. Ces travaux sont actuellement utilisés pour analyser le comportement d'apprenants lors du visionnage de cours en ligne.

IV.C Perception des humains à distance

C.1) **Projet MobileRGBD**

À la fin du projet PRAMAD, l'un des besoins exprimés pour notre robot compagnon était lié à la sécurité. Il fallait doter ce robot d'une fonction routine lui permettant de trouver une personne dans l'appartement. Cette routine s'active dans deux cas. Le premier cas, le plus courant, correspond au moment où le robot sort de son mode recharge et part en quête de la personne aidée. Dans le second cas, un événement extérieur, comme un message provenant d'un médaillon d'alarme indiquant une chute [206], oblige le robot à se mettre à la recherche de la personne pour lui porter assistance. Cette personne peut être debout ou tombée à terre suite à un malaise, consciente ou inconsciente. Cette fonction de sécurité est primordiale pour le maintien à domicile des personnes âgées.

De nombreux travaux sur la détection de chutes se trouvent dans la littérature [207, 208, 209, 210, 211, 212], soit avec des caméras, soit avec des capteurs RGBD. Ceux-ci s'avèrent pertinents et efficaces mais ces approches sont insuffisantes dans le cas d'un robot compagnon. Pour que la chute soit détectée, il faut parfois disposer d'une image de fond permettant de distinguer les objets mobiles. La chute doit également avoir lieu dans le champ de vision des capteurs du robot, à une distance raisonnable (généralement moins de 5 m). Même en multipliant les capteurs pour avoir une vue à 360°, pour garantir la sécurité, la détection de chutes ne peut être le seul outil à disposition sur un robot compagnon.

Après avoir évalué les algorithmes disponibles [213, 214, 215], nous avons fait le constat que les performances de ces algorithmes étaient intrinsèquement liées à la posture initiale de la personne (debout ou assise) et/ou au fait qu'elle bouge. En testant ces solutions dans différents cas, nous avons mis en évidence que l'angle d'arrivée du robot vers une personne à terre influence fortement le résultat de détection. Ainsi, si le robot arrive vers les pieds de la personne et que son capteur est orienté vers le bas, alors ces algorithmes fonctionnent. Dans les autres cas, ces algorithmes échouent la plupart du temps.

Figure 23 : Notre plateforme d'enregistrement, un mannequin debout, un mannequin couché [21](à gauche à droite).

En collaboration avec Amaury Nègre, nous avons décidé de nous intéresser à cette détection de personnes avec un robot compagnon équipé de capteurs standards. En cherchant dans la littérature à l'époque, nous n'avons pas trouvé de corpus compatibles avec ces recherches. Nous avons donc lancé le projet MobileRGBD [216], un corpus enregistré en environnement domestique dans des conditions proches du réel. Nous avons décidé de construire un corpus qui puisse également servir de benchmark à différents algorithmes liés aux robots mobiles [217]. Ce corpus a été conçu pour évaluer l'impact de différentes variables sur les performances des algorithmes étudiés :

- le nombre de personnes et leurs positions dans l'environnement, l'orientation du corps relativement au capteur ;
- la position des meubles et les occlusions du corps ;
- la vitesse linéaire et angulaire du robot ;
- la hauteur et l'inclinaison idéales du capteur RGBD.

Pour rendre les différents enregistrements comparables, nous avons eu l'idée d'employer des mannequins[21] qui garderont leur pose et leur position pendant toute une session d'enregistrement. Cela permet de répéter les trajectoires à différentes vitesses, avec différentes hauteurs et inclinaisons du capteur RGBD. La figure 23 montre notre plateforme d'enregistrement et nos mannequins en situation.

[21] Sur notre idée, le mannequin mou pour les poses de personnes tombées à terre a été réalisé par *Laurence Boissieux* du Service Expérimentation et Développement (SED) de l'Inria.

Après avoir réécrit toute la couche de contrôle du robot avec Amaury Nègre, nous avons pu réaliser des enregistrements dans trois différents environnements : un appartement expérimental, des couloirs de notre centre de recherche et un appartement réel du *Living Lab* de l'équipex *Amiqual4Home*[22]. Toutes les données brutes des capteurs embarqués sur le robot ont été capturées, un plan de chaque environnement et un étiquetage des données sont fournis.

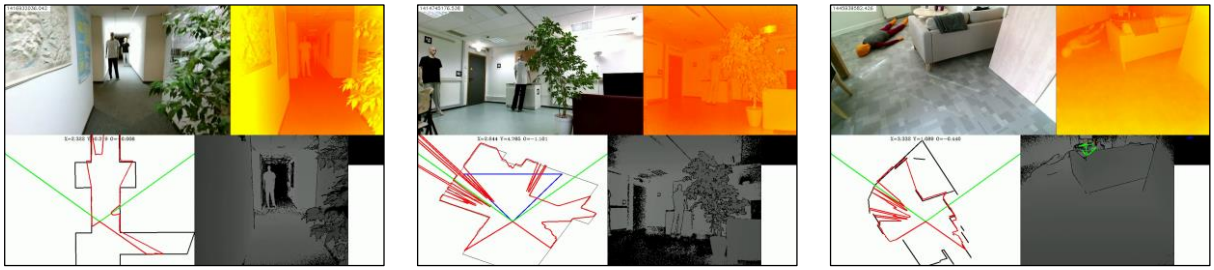


Figure 24 : Exemples de données enregistrées dans nos 3 environnements. De gauche à droite : un couloir, l'appartement expérimental, l'appartement du *Living Lab*.

Comme le montre la figure 24, les environnements présentent des caractéristiques différentes. Le plus réaliste est l'appartement du *Living Lab* car c'est un véritable appartement avec ses contraintes, notamment en terme de variation naturelle de luminosité au cours de la journée.

MobileRGBD, notre contribution à l'étude de la détection de personnes avec un robot compagnon, est un corpus contenant 9h30 d'enregistrement pour un total de 4,52 Tio de données (1,44 Tio compressés). Il est téléchargeable et disponible pour la communauté[23]. Le code source permettant de relire et d'afficher toutes les données de manière synchrone est disponible sur un dépôt *GitHub*[24]. Ce code source permet également de reproduire ces enregistrements sur des plateformes comparables.

C.2) **Détection de personnes**

Les réseaux de neurones sont utilisés depuis plus de 30 ans pour l'apprentissage machine [3]. Il y avait cependant 2 verrous majeurs à lever pour amener à ce que l'on nomme aujourd'hui l'apprentissage profond : la puissance de calcul pour simuler de gros réseaux avec plusieurs millions de neurones et la disponibilité de grand corpus de données annotés, notamment en vision par ordinateur. Concernant le premier point, la mise au point des réseaux convolutifs [218] associée avec l'avènement du calcul sur GPU ont permis de paralléliser massivement les calculs de ces réseaux et de les démocratiser. Concernant le manque de corpus, l'essor des réseaux sociaux avec le partage massif de photographies a contribué à résoudre le problème. Pour faire de ces données massives des corpus utilisables, il fallut avoir recours aux

[22] Voir https://amiqual4home.inria.fr/fr/ (dernière visite 04/2018).

[23] Voir http://mobilergbd.inrialpes.fr/ (dernière visite 04/2018).

[24] Voir https://github.com/Vaufreyd/RGBDSyncSDK (dernière visite 04/2018).

plateformes de myriadisation (« *crowdsourcing* ») comme *Amazon Mechanical Turk* ou obliger les utilisateurs à annoter des images via des captchas[25]. Ces méthodes, même si elles posent des problèmes notamment éthiques [219, 220], ont été à la base de la disponibilité de corpus comme *ImagineNet* [221].

Depuis 2016, nous travaillons sur le corpus MobileRGBD (vois section précédente) pour mettre au point un système de détection de personnes, dont certaines tombées à terre, par un robot mobile. Dans cette tâche, nous nous sommes d'abord intéressés à des approches géométriques [213, 222] utilisant des données de profondeur qui n'ont pas produit les résultats escomptés. Nous nous sommes tournés vers les approches en apprentissage profond. Quatre grands types d'approches pour la détection de personnes sont décrits dans la littérature [223]. Premièrement, les approches à l'échelle du pixel [224] (« *pixel-wise* ») déterminent pour chaque pixel s'il appartient à une personne. D'autres approches recherchent des points clés [225, 120, 226] (« *keypoints* », par exemple nez, coude, genou…) puis les regroupent pour chacune des personnes présentes à l'image. Des approches mixtes recherchent des parties de corps [119] (tête, bras, jambe par exemple) avec une détection au niveau du pixel. Enfin, il est possible de détecter et de classifier des objets dans l'image [161, 227], notamment des personnes. Ces différentes approches sont illustrées sur la figure suivante.

Figure 25 : Différentes approches de détection de personnes dans l'image. De gauche à droite : image originale [228], détection de points clés [225, 120, 226], détection de personne [224] et détection de parties du corps [119] pour chaque pixel, détection d'objets dans l'image [161].

Les approches présentées sur la figure 25 ont bénéficié de corpus comme PASCAL VOC [228] ou MS-COCO [229] avec leurs dizaines de milliers d'images annotées à différents niveaux :

[25] Voir https://fr.wikipedia.org/wiki/CAPTCHA (dernière visite 04/2018).

points clés, masques de pixels pour les personnes détectables, masques de foules pour les groupes de personnes.

Dans notre tâche de détection de personnes debout ou couchées, dans n'importe quelle orientation, nous avons évalué les algorithmes disponibles. Ceux-ci montrent de très bonnes performances mais avec des spécificités propres à leur corpus d'apprentissage : les personnes doivent être plutôt en position debout. La figure suivante présente les résultats d'une évaluation sur la tolérance à la rotation de l'algorithme *OpenPose* [225, 120, 226] sur une image issue du corpus MobileRGBD (d'autres exemples sont disponibles en annexe).

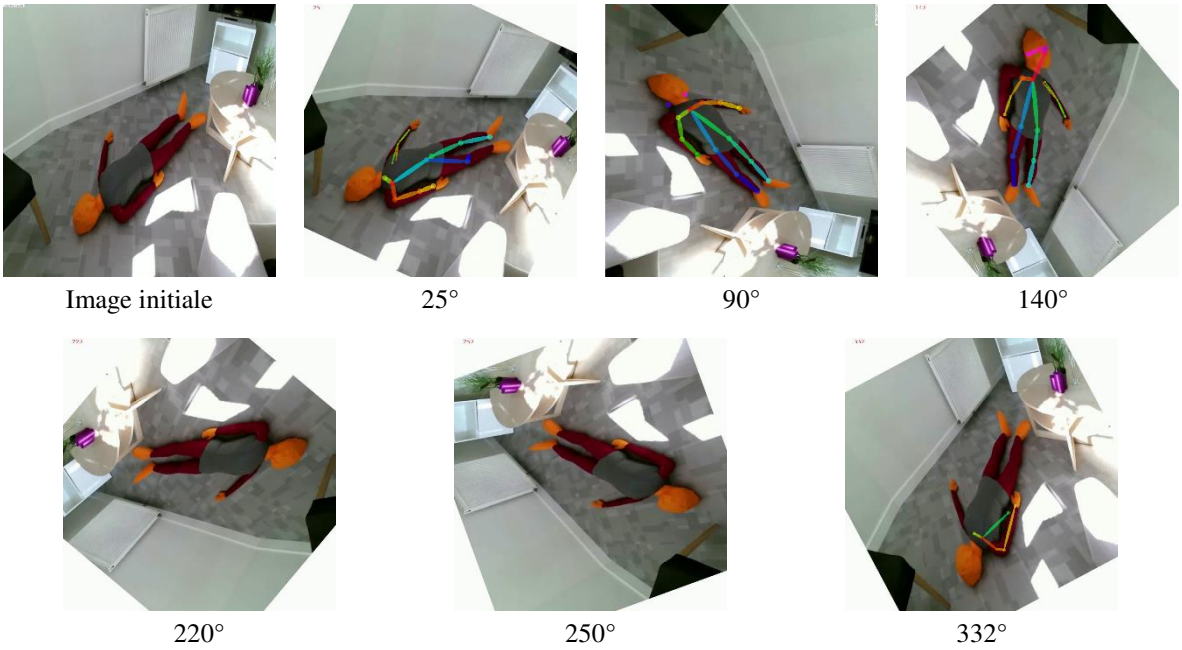


Figure 26 : Détection de *keypoints* avec différentes rotations d'images (valeur en degrés).

Le premier constat que l'on peut faire semble être la tolérance à la couleur de l'algorithme. De manière surprenante, la couleur et l'uniformité de texture d'un mannequin ne pose pas spécifiquement de problèmes tant que celui-ci est dans une position tête en haut dans l'image (rotation de 140° sur la figure précédente), l'algorithme infère la position des yeux et des oreilles alors que notre mannequin n'en possède pas. Notre mannequin a pourtant des spécificités propres qui n'ont jamais été vues dans le corpus d'apprentissage. D'autres expériences sur notre corpus et sur des images modifiées du corpus COCO (non détaillées dans ce manuscrit) ont conclu que l'algorithme était également peu sensible à la luminosité et au flou dans l'image.

La rotation semble être le principal point faible de l'algorithme. Nous pouvons noter qu'en position initiale, il n'y a aucune détection. Autours de 25° de rotation, il y a un début de détection du corps mais pas de la tête. La détection s'améliore ensuite jusqu'à 140° pour se détériorer puis se stopper vers 220°. Une dernière détection d'épaule est présente spécifiquement à 332°. Ces résultats témoignent d'une très forte sensibilité de l'algorithme à

l'orientation avec de faibles changements : même une variation de 1° impacte la détection de *keypoints*.

L'algorithme d'*OpenPose*, comme les autres algorithmes testés [224, 119, 161] ne sont pas ou peu tolérants à la rotation. Pourtant, ceux-ci utilisent des méthodes d'augmentation des données [221] (« *data augmentation* ») lors de leur phase d'apprentissage. Les images présentes dans le corpus sont tournées, recadrées et changées d'échelle pour augmenter synthétiquement la quantité de données disponibles pour l'apprentissage et pour maximiser la capacité de généralisation du réseau de neurones. Dans le cas de l'algorithme *OpenPose*, la première mesure corrective est d'augmenter la rotation des images lors de l'apprentissage jusqu'à ±180° pour refléter toutes les positions possibles des personnes tombées à terre. Cela ne donne malheureusement pas les résultats escomptés. Les performances sur les benchmarks standards [229] diminuent sans augmentation significative des performances sur notre problématique, et ce même en entrainant le système plus longtemps (i.e. plus d'époques) et en ajustant les paramètres d'apprentissage [230].

Une autre solution consiste à réaliser un apprentissage par transfert (« *transfer learning* ») à partir de données spécifiques à notre tâche. Grâce à cela, nous avons mis au point un système détectant les personnes à terre, mais qui n'est plus capable de détecter correctement les personnes debout. Une solution symétrique consiste à traiter les images en entrée à différentes orientations avec un réseau de neurones standard, la problématique étant de choisir ces orientations. Ces solutions restent scientifiquement peu intéressantes. Elles éludent le problème de généralisation pour fournir une solution pour laquelle on ne peut pas garantir la perfection.

Figure 27 : Effet de la rotation sur la détection d'objets avec Yolo v3 [161].

D'autres tâches comme la détection et la classification d'objets souffrent des mêmes affres. Si l'on s'intéresse à la détection de piétons et de véhicules [161] à l'aide de caméras situées au-dessus d'un carrefour (voir section « Interaction avec des véhicules autonomes »), la problématique reste la même. Sur la figure 27, nous voyons que la rotation de l'image perturbe le système. Un résultat étonnant est la non détection du véhicule même si celui-ci parait, pour un humain, très similaire à son image non retournée. Ce problème de détection lors des rotations, même s'il n'est pas présent dans toutes les images, montre encore une fois que les données et la méthode d'apprentissage ne favorisent pas la généralisation à toutes les orientations.

Notre contribution sur la thématique de la détection de personne à base d'apprentissage profond se borne, au moment de la rédaction de ce manuscrit, à une analyse des réseaux disponibles et de leurs faiblesses dans leur capacité de généralisation. Malgré leur performance, les réseaux de neurones actuels ne généralisent pas suffisamment et sont encore fortement conscrits par les données d'apprentissage. Ils ne diffèrent pas, sur ce point, d'autres systèmes mis au point précédemment [231]. La seconde problématique est une tendance de fond dans les recherches en apprentissage profond : visualiser et comprendre ce que le réseau de neurones apprend réellement [232, 233]. Sans cette compréhension, l'amélioration des systèmes de détection de personnes en modifiant la topologie des réseaux ne peut se faire de manière éclairée.

Le perfectionnement des réseaux de neurones, pour la détection de personnes mais pas seulement, est l'une des thématiques présentées dans mon projet de recherche dans le chapitre suivant. Ces travaux serviront de base à nos recherches sur l'interaction sociable avec des robots compagnons ou des véhicules autonomes.

Chapitre V
Perception multimodale et interaction sociable

V.A Perspectives

Au travers de mes contributions présentées dans ce mémoire, j'ai abordé différents systèmes interactifs. Mon projet de recherche se situe dans la continuité de ces travaux et vise à faire de ces systèmes interactifs une réalité dans la vie quotidienne comme l'est devenue la reconnaissance de la parole aujourd'hui. Les maisons connectées et les robots compagnons ont un grand rôle à jouer dans le maintien des personnes âgées et/ou fragiles à domicile. Les véhicules autonomes joueront un rôle certain dans les transports urbains de demain. Il reste cependant des progrès à réaliser, en terme de sécurité d'une part et en terme de qualité d'interaction sociable d'autre part, pour que ces technologies soient disponibles et utilisables par le plus grand nombre.

Mon projet de recherche traite ces thématiques sous deux angles : la robustesse de la perception multimodale et l'interaction sociable. Le but est de rendre ces systèmes interactifs plus sensibles et réactifs aux humains qui les entourent. Cela passe par la création d'outils et de modèles permettant de mieux percevoir, comprendre et anticiper les comportements des humains dans leur environnement.

Pour que cela soit réalisable, il est primordial de pouvoir s'appuyer sur des systèmes perceptifs robustes. Comme nous l'avons vu (voir section « Détection de personnes »), cette robustesse sera évaluée par des mesures sur des benchmarks standards mais également par des analyses de la sensibilité des algorithmes à différentes conditions et dans des évaluations en conditions réelles (« *in the wild* »). Ces connaissances sont nécessaires à la fois pour construire des systèmes perceptifs plus performants mais également pour garantir leur fonctionnement, et par là même, la sécurité des utilisateurs. Cette garantie de fonctionnement est par exemple l'une des préoccupations majeures exprimées par les assureurs concernant les véhicules autonomes comme discuté lors de la conférence « *Les véhicules autonomes. Quelles responsabilités ? Quelle assurance ?* » [234]. Ces recherches sur la robustesse des algorithmes de perception constituent le premier objectif que je traiterai dans mes recherches.

Mon second objectif est de permettre une interaction plus sociable en étudiant de nouvelles formes de perception pour nos robots (tactile, suivi de l'attention visuelle, …) comme nous l'avons fait lorsque nous nous sommes intéressés à la fusion entre la vision, l'acoustique, la proxémique et le langage corporel. Ces nouvelles modalités apporteront une meilleure analyse du comportement des humains et de leurs intentions dans l'environnement de nos systèmes. Elles apporteront également de nouveaux challenges en perception.

Mon troisième objectif consiste à rendre les robots, et plus généralement les systèmes interactifs, plus expressifs en nous intéressant à des formes de communication ne misant pas tout sur l'anthropomorphisme. Ces travaux devront, comme nous l'avons fait par le passé, s'intéresser à d'autres domaines comme la sociologie et les sciences cognitives pour prendre en compte ce que les humains perçoivent et comment ils le perçoivent.

Les recherches menées pour atteindre ces objectifs peuvent s'avérer très intrusives suivant la nature des capteurs employés. Les microphones, les caméras, les capteurs tactiles peuvent potentiellement porter atteinte à la vie privée [235]. Comme il y a souvent un détournement à mauvais escient des technologies, nous ne pourrons pas nous dédouaner d'une réflexion globale sur l'éthique de nos recherches.

Enfin, pour conclure sur des aspects plus pragmatiques, il nous faudra financer ces recherches. Concernant la détection de personnes, nous avons déjà obtenu un financement sur projet. Pour les autres, la tâche consistera à coordonner des actions visant à soutenir nos besoins (recherche de financements de thèses, montages de projets nationaux et internationaux, collaborations industrielles…).

Les premiers jalons pour réaliser nos objectifs concernant la perception multimodale et l'interaction sociable avec des robots sont résumés dans les sections suivantes.

V.B Perception multimodale

B.1) **Robustesse des systèmes de perception**

Les performances des réseaux de neurones profonds dans de nombreuses tâches de perception sont impressionnantes au regard de l'état-de-l'art il y a encore quelques années. Pourtant ceux-ci ne sont pas exempts de problèmes. Certains travaux montrent leur faible résistance à des attaques simples permettant de les tromper tout en obtenant de très bons scores de confiance [236], même en ne modifiant qu'un seul pixel de l'image [237]. Pour la détection de personnes, les systèmes actuels sont très performants [225, 120, 226, 119, 161, 227] mais n'arrivent pas au même niveau de généralisation qu'un humain : identifier une personne de manière robuste, c'est-à-dire dans différentes positions, avec occlusions partielles, etc. Nos travaux ont illustré ces problèmes de généralisation des réseaux de neurones profonds avec une sensibilité forte à l'orientation des images (voir section « Détection de personnes ») mais

parallèlement avec une robustesse aux changements de couleurs, au flou et à la luminosité des images.

Mon projet de recherche vise, à moyen terme, à améliorer la robustesse de réseaux de neurones pour la détection de personnes. La littérature fournit de plus en plus d'outils permettant de visualiser les activations dans les couches cachées des réseaux [233, 238] ou de reconstruire une image synthétique qui maximise la détection du réseau [239]. Ces approches fournissent des indices sur la sensibilité du réseau à différentes entrées.

La nature même des réseaux convolutifs est en partie responsable [232] de ces problèmes de généralisation. L'apprentissage par transfert pose aussi question. En moins d'une décennie, la reconnaissance de formes est passé d'un paradigme où, pour progresser, l'on définissait de nouvelles caractéristiques sur le signal à un apprentissage neuronal profond directement sur les données, les caractéristiques profondes (« *deep features* ») émergeant du signal, ce qui fut un progrès. Quelques années plus tard, nous revenons au paradigme initial par souci d'immédiateté. Ces caractéristiques profondes sont reprises de réseaux connus (*AlexNet*, *VGG*, *GoogleNet*, *ResNet*, *DenseNet*, etc. et leurs variantes [240]) et transférées sur une nouvelle tâche sans toujours mesurer si ce transfert est pertinent [241, 242]. Si nous ne voulons pas tomber dans les travers de nombreux articles en ajoutant/supprimant des couches de neurones, utilisant telle ou telle topologie de réseau sans en mesurer l'intérêt a priori, il nous faut comprendre ce que les réseaux de neurones apprennent, à quoi ils sont sensibles.

Nous devrons réaliser l'analyse des réseaux existants sous le prisme de la généralisation. Dans un premier temps, nous étudierons la généralisation à l'orientation. Nous évaluerons l'intérêt de réseaux spécifiques (*dynamic filters* [243], *spatial transformers* [244], *densecap* [245]…), des méthodes de spécialisations de réseaux [246] ou proposerons de nouvelles topologies de réseaux répondant mieux à notre problématique. À plus long terme, cette compréhension de la généralisation et les nouveaux modèles qui en découleront, seront appliqués à la robustesse de systèmes intégrant des données multimodales fortement hétérogènes comme celles provenant d'un robot compagnon ou d'un véhicule autonome [247].

B.2) **Vers un sens tactile pour les robots**

Jusqu'à présent, les recherches sur l'interaction naturelle et intuitive entre l'homme et le robot se sont principalement concentrées sur les modalités visuelles et sonores. Quelques études ont accordé de l'attention à la modalité tactile de l'interaction. Certains robots comme *Aibo*, *Paro*, *Nao* ou *Reeti* sont équipés de capteurs tactiles. Certains chercheurs ont étudié la détection « cutanée », avec un grand nombre de capteurs répartis sur tout le corps du robot [248, 249, 250]. Même pour les cobots industriels, utiliser le toucher pour partager l'information peut améliorer la coopération entre le robot et son partenaire humain [251, 252].

La reconnaissance des gestes tactiles est un sujet de recherche en plein essor avec des brevets comme celui de *Google* sur une grille de capteurs tactiles [162]. Comme l'ont montré *Cooney et al* [253], le toucher est un vecteur important de communication de leur affect pendant

l'interaction homme-robot. Ces résultats suggèrent de se concentrer davantage sur la reconnaissance des gestes du toucher affectif et l'amélioration de la communication tactile humain-robot. Certains auteurs [254] proposent de catégoriser les interfaces en fonction des couvertures du capteur et en fonction du type d'interaction physique visant à interférer avec le fonctionnement du robot (i.e. saisir pour arrêter le mouvement par exemple), pour faire partie de l'interaction comme moyen de communiquer (i.e. tapoter pour signifier la satisfaction) ou pour apprendre à partir d'entrées (apprentissage par renforcement par exemple).

Lors de nos expérimentations avec des robots, certains utilisateurs se sont parfois servi intuitivement du toucher pour essayer d'interpeller le robot comme on le ferait en tapotant sur l'épaule de quelqu'un pour qu'il se retourne. D'autres tentaient de le guider lors de ses déplacement en l'effleurant. Ce semble confirmer le besoin de sens tactile sur les robots pour augmenter l'intuitivité dans les interactions homme-robot [255].

Partant de ce constat, nous nous sommes déjà intéressés à la reconnaissance de gestes tactiles sociaux lors du *Touch Challenge* à la conférence ICMI2015[26] [15]. Cette expérience nous conduit actuellement au design d'un robot compagnon avec des surfaces tactiles réparties sur le corps. Par l'expérimentation, les travaux que nous projetons de mener doivent faire émerger ce qui permet naturellement et intuitivement à des utilisateurs d'utiliser le toucher comme moyen d'interaction avec le robot. Nous expérimenterons également le toucher sur des véhicules autonomes lors d'interactions à faible vitesse dans des espaces partagés avec des piétons.

V.C Interaction sociable avec des robots

C.1) **Interaction sociable versus sociale**

Contrairement à ce que j'ai pu faire par le passé, je n'utilise plus la dénomination « interaction sociale » pour qualifier les recherches que je mène actuellement. Il y a beaucoup de notions complexes dans les interactions sociales avec des échanges de signaux sociaux synchronisés entre tous les partenaires [256]. Ces échanges supposent une notion de rôle dans un groupe d'entités avec des affects et des enjeux qui ne sont pas imputables intrinsèquement à l'interaction homme/système ou à sa qualité. Dans de nombreux domaines, nous n'avons pas encore atteint une qualité de perception et de modélisation de l'humain qui permette de qualifier les interactions homme-robot de « sociales ».

Dans la littérature, plusieurs auteurs ont proposé des définitions pour l'interaction sociale et/ou l'interaction sociable avec des robots. Dans les travaux de *Barraquand*, les technologies sociables ont une définition précise : « *Sociable technologies refer to the set of things created by a mind as an extension of techniques to improve social cohesion, social interaction and*

[26] *International Conference on Multimodal Interaction*, voir https://icmi.acm.org/ (dernière visite 04/2018).

cooperation » [257]. Nous avons beaucoup échangé avec l'auteur sur cette définition. Les aspects interaction et coopération sont importants. Concernant la cohésion sociale, cela donne un rôle très important au robot qui, à mon sens, n'est pas forcément souhaitable.

Dans les travaux de *Breazeal* [258], les robots sociaux ont jusqu'à 4 facettes :

- *Socially receptive.* Ce type de robot est capable de percevoir les signaux sociaux émis par leurs partenaires humains et de les interpréter. C'est une notion primordiale dans l'interaction homme-robot comme nous l'avons vu dans la section « Interaction avec des robots ».
- *Socially evocative.* Cette facette concerne les robots qui, par leur forme, provoquent de l'anthropomorphisation et un engagement plus profond de l'utilisateur lors des interactions. C'est ce que d'autres auteurs nomme la « glue-socioaffective » [259], glue qui se développe au fur et à mesure des avancées expressives du robot. Le problème sous-jacent de l'anthropomorphisation est l'attente élevée concernant les capacités cognitives du robot que cela amène et la possible désillusion de l'utilisateur [260, 261].
- *Social interface*. Ces robots expriment des signaux sociaux ressemblant à ceux des humains. Dans nos travaux sur les robots compagnons (voir section « Retour émotionnel du robot »), nous nous sommes intéressés à ce type de retour émotionnel des robots compagnons. Nous avons expérimenté la difficulté d'obtenir un retour émotionnel crédible, dans des interactions simples et limitées dans le temps [141]. Cette difficulté s'accroît encore pour les interactions homme/robot au long cours.
- S*ociable*. Pour *Breazeal*, les robots sociables sont motivés par les besoins des utilisateurs mais également par leurs propres buts. Ils ont une conscience d'eux-mêmes et modélisent le comportement humain pour augmenter la performance de l'interaction et leur impact social.

Dautenhahn [94] indique quant à elle : "*A robot companion in a home environment needs to 'do the right things', i.e. it has to be useful and perform tasks around the house, but it also has to 'do the things right', i.e. in a manner that is believable and acceptable to humans*". Cette définition est également valable dans d'autres contextes. Les véhicules autonomes doivent aussi répondre à cette définition dans leurs interactions avec des piétons. La notion de *believable* est intéressante et renvoie, pour moi, à la notion de prédictibilité. Pour que l'on puisse faire confiance à un robot, alors on doit être capable de prédire son comportement.

De mon expérience et de ces définitions, je propose ma définition de ce qui est pour moi une interaction sociable.

Définition : *Une interaction sociable est une interaction naturelle, non intrusive, intuitive, prédictible et expressive.*

Nous pouvons détailler les différentes propriétés d'une interaction sociable :

- Une interaction est naturelle si elle ne provoque ni rejet, ni peur, ni dégout de la part des partenaires humains du robot ;
- La non-intrusivité et l'intuitivité supposent de capturer et d'interpréter finement les signaux sociaux des humains grâce à la perception multimodale distante, domaine de recherche en nette progression suite aux progrès en apprentissage profond. Cependant, ces progrès masquent des problèmes d'explicabilité et de reproductibilité des performances dans des conditions diverses (voir chapitre « Perception des humains ») ;
- La prédictibilité de l'interaction est primordiale. Dans les mêmes conditions, l'interaction doit toujours se dérouler à l'identique, le partenaire humain pouvant alors adapter par anticipation son comportement à celui du robot ;
- Enfin, l'expressivité doit permettre au robot de clairement exprimer son état interne ou ses intentions comme lorsqu'il contourne un groupe de personnes [262]. Plutôt que de mimer les signaux sociaux des humains, il faut, à mon avis, s'atteler au développement de codes sociaux propres aux robots (voir section suivante).

Développer ces caractéristiques des interactions sociales est le préalable à la mise au point de robots performants dans leurs interactions sociables avec des humains. Cela pose néanmoins la question de la forme de ces robots. Quel degré d'anthropomorphisme doit-il avoir ? S'il est humanoïde, quel est l'impact et l'intérêt de cette forme sur l'interaction ? [263]. Si l'on s'intéresse par exemple au robot *Nao*, vu ses piètres qualités de marcheur, notamment pour suivre une trajectoire, il est souvent utilisé dans des interactions où il ne se déplace pas, son torse, ses bras et sa tête servant seuls l'expressivité. Sa forme totalement humanoïde n'a alors pas d'intérêt pour les interactions, un buste pouvant suffire. D'ailleurs, son successeur, le robot *Pepper* ne possède pas de jambe. Il peut y avoir, pour moi, une forme de narcissisme dans le fait de construire des robots sociaux anthropomorphes ou humanoïdes. Il y a bien sûr des défis scientifiques et techniques derrière la forme humanoïde d'un robot, mais ceux-ci doivent être totalement résolus et évalués en terme d'utilisabilité pour les interactions.

À plus long terme, une fois les verrous scientifiques à la mise au point de ces robots sociables levés, nous pourrons définir quelles sont les tâches pertinentes qu'il convient de leur confier et évaluer, dans différents contextes, leur impact sociétal.

C.2) **Vers des codes sociaux pour les robots**

L'expression par le robot de son état interne et de ses intentions est primordiale pour compléter la boucle d'interaction avec leur partenaire humain. Comme nous l'avons évoqué, cette expressivité passe souvent par une forme humanoïde pour tenter de mimer des signaux sociaux humains [264]. Dans nos travaux, nous l'avons-nous même expérimenté avec notre robot compagnon *Kompaï* [141]. Nos résultats ont montré que l'utilisation d'un smiley se montrait plus efficace qu'un avatar humain et que l'intérêt de la tête mobile semblait plus résider dans

les capacités de perception accrues qu'elle donne au robot que dans la compréhension et l'acceptabilité de celui-ci. Mimer de manière imparfaite les codes sociaux humains expose à un plus grand risque de tomber dans la vallée de l'étrange [128].

Par le passé, certains dispositifs ont été doté de codes sociaux. Par exemple, les ordinateurs expriment à l'aide de sons simples les erreurs rencontrées. Ces codes simples sont reconnaissables par toute personne ayant travaillé avec un ordinateur et sont, par là même, devenus universels à l'usage.

Faisant suite à cette habilitation, mes recherches exploreront d'autres voies, plus simples, pour l'expressivité des robots. En s'intéressant aux recherches en sciences cognitives et en sociologie [27, 87], il faudra évaluer de nouvelles formes d'interaction où l'expression des robots leur sera propre. Cette forme pourra être différente selon que le robot soit un robot compagnon, un siège autonome pour personne handicapée ou un véhicule autonome avec plusieurs passagers à l'intérieur par exemple. La communication devra également prendre en compte le contexte au sens large, en intégrant dynamiquement tous les éléments de l'environnement, pour être compréhensible et perceptible pour tous ses destinataires.

Annexes

Informations complémentaires sur le déroulement de carrière

Contrats

Je ne donne ici que des informations concernant les projets pour lesquels j'ai (eu) des responsabilités. Les autres projets auxquels j'ai participé/je participe en tant que chercheur sont listés en introduction.

2008–2011 Projet *PersoPos* (Financement Grenoble INP) avec l'IMU MICA. Il était doté d'un budget de 8k€ ainsi que qu'une allocation de thèse.

09/2011 – 02/2016 Responsable pour *Prima* du FUI PRAMAD avec un budget de 253k€. Celui-ci a été en majeure partie utilisé en embauche d'ingénieurs (4 au total), en achat de matériel robotique et en mission pour les évaluations menées au *Living Lab* de l'Hôpital Broca à Paris.

04/2011 – 10/2015 Responsable pour *Prima* de l'AEN Personally Assisted Living (PAL) de l'*Inria*, le budget de l'équipe consistait principalement en 18 mois d'ingénieur expert pour la mise en place de l'infrastructure d'échange entre les différentes équipes. Le budget matériel, mission et stages était d'environ 15k€.

09/2014 – 09/2015 Projet exploratoire *Figurines* (*Labex Persyval Grenoble*) était lui doté d'un budget de 8k€ à partager avec l'équipe *IMAGINE*.

09/2017- … Responsable scientifique du projet de transfert industriel *Inria-Innovation-Lab* Toutirobo-2, son budget est de 28 homme/mois pour l'embauche d'un ingénieur.

01/2018 – 01/2021 Responsable du Workpackage 5 « *Communication and Negotiation* » du projet ANR HIANIC (*Human Inspired Autonomous Navigation In Crowds*).

Responsabilités collectives

À l'Inria

Comité de Centre (COC) de l'Inria Rhône-Alpes

Je suis membre élu du Comité de Centre de l'Inria Rhône-Alpes en tant que représentant des chercheurs depuis 2007. Je suis très attaché à ce mandat. En effet, avant cela, la représentativité des enseignants-chercheurs n'était pas directe dans cette instance. Nous sommes un public particulier, avec nos spécificités (statut, présence dans les locaux, etc.) qui se doit d'être représenté.

Commission d'attribution des locaux

J'ai été membre de la commission chargée d'affecter les locaux aux équipes de recherches et aux services début 2011. Depuis cette date, j'ai été le responsable des autres commissions d'attribution des bureaux suite à des déménagement d'équipes sur d'autres sites et la construction de nouvelles ailes dans notre bâtiment.

Au Laboratoire d'Informatique de Grenoble (LIG)

Chargé de mission – Animation Scientifique (11/2011-12/2014)

Pour le LIG, j'ai été chargé de mission Animation Scientifique entre 2011 et 2014. À ce titre, j'ai été le co-organisateur des Conférences de Prestige du LIG[27] (*LIG Keynotes speeches*, 10 conférences par an) et responsable des séminaires LIG. Ces séminaires permettent d'inviter un chercheur pour faire un exposé intéressant une ou plusieurs équipes. J'ai mis en place tous les aspects sous-traitance pour les enregistrements et la diffusion en *live* et en *podcast* de ces événements avec l'agence de moyen UMS MI2S. Après 3 ans, notre équipe a décidé de passer la main fin 2014.

À l'Université Pierre Mendès-France (UPMF, avant 2016)

Commissions « Maître de Conférences » et « Dispositifs innovants »

J'ai fait partie en 2011 à l'UPMF des commissions « Maître de Conférences » et « Dispositifs innovants ». La première a abouti à différentes propositions concernant les conditions d'exercice du métier de MCF et à certaines avancées (proposition de crèche pour l'UPMF, dispositif de retour à la recherche pour les MCF non publiant par un contrat tripartite Laboratoire/MCF/Université avec un financement recherche pour le MCF). La seconde a mis en place certains mécanismes d'incitation à la recherche (dispositifs de décharges, …) ainsi

[27] Voir https://mi2s.imag.fr/pm/videos-en-ligne/LIG-keynote (dernière visite (04/2018).

que la mise en place d'une limitation des heures complémentaires possibles conditionnée par l'activité de recherche.

Vice-président du Département de Spécialistes 27ème/71èm sections

En 2010 et 2011, j'ai été vice-président élu du Département mixte 27ème section/71ème section des enseignants-chercheurs de l'UPMF. Dans ce cadre, selon les modalités de recrutement instaurées à l'UPMF, le département de spécialistes participe avec le Conseil Scientifique aux recrutements d'enseignants-chercheurs (ATER/MCF/Professeurs) et j'étais donc amené à participer à ce processus. Je participais également aux commissions interuniversitaires d'affectation pédagogique des étudiants en doctorat (ex-monitorat) du CIES de Grenoble.

Rayonnement

Comités de sélection et jurys de thèse

2005-… Membre des comités de sélections des enseignants temporaires en Informatique à la Faculté d'Économie de Grenoble (25 moniteurs ou doctorants avec charge d'enseignement, 26 ATER, 2 contrats CDD LRU, 1 poste PRAG). Recrutement de plus d'une 50aine d'enseignants vacataires.

Membre du jury de recrutement de 11 ingénieurs de recherche sur projet à l'*Inria*.

2014 Membre du comité de sélection du poste de Maître de Conférences 27 MCF 0327 affecté TC/CITI de l'INSA de Lyon.

2014 Membre du comité de sélection du poste de Maître de Conférences 27 MCF 158 de l'IUT2 d'Informatique de Grenoble.

2015 Examinateur pour la thèse de *Wafa Benkaouar Johal*, « *Companion Robots Behaving with Style: Towards Plasticity in Social Human-Robot Interaction* », à l'Université Grenoble Alpes

2016 Examinateur pour la thèse de *Jonathan Aigrain*, « *Multimodal detection of stress : evaluation of the impact of several assessment strategies* », à l'Université Pierre et Marie Curie de Paris.

Comités de rédaction de revue, comités d'organisation, comités de programme de conférences internationales ou francophones

2000 Comité d'organisation de **RECITAL2000** associé à la conférence Traitement Automatique des Langues Naturelles (TALN2000) à Lausanne

2003 Co-organisateur des **RJC'2003** (Rencontres Jeunes Chercheurs en parole) à Grenoble : 51 communications, 4 intervenants invités, 61 participants.

2009 Comité d'organisation d'Interface Homme Machine (**IHM'09**) à Grenoble. En plus de l'organisation locale, j'ai eu la charge de l'éditions des actes avec Patrick Reignier.

2013 Comité scientifique ainsi que comité d'organisation du **WASSS** (*Workshop on Affective and Social Speech Signal*), événement satellite de la conférence *Interspeech* (>1000 participants tous les 2 ans). Dans l'organisation de ce Workshop, j'ai eu en charge une partie des demandes de subventions ainsi que la gestion du système de soumission *Easy Chair*. Le Workshop a eu une audience internationale de plus d'une cinquantaine de participants.

2017 Co-organisation du BER Workshop (*Behavior, Emotion and Representation: Building Blocks of Interaction*) à la conférence HAI'2017 à Bielefeld en Allemagne. J'ai eu la charge des demandes de subvention et de l'organisation locale. Nous avons attiré un peu moins de 30 participants.

J'ai fait partie de nombreux comités de programmes. Depuis 2008, celui du IEEE International Workshop on Multimedia Technologies for E-Learning (MTEL). J'ai aussi participé aux special issues de ce Workshop. J'ai également servi comme membre du comité de programme pour : UBICOMM (International Conference on Mobile Ubiquitous Computing, Systems) de 2011 à 2016 ; Workshop on Assistance and Service Robotics in a Human Environment at IEEE International Conference on Intelligent Robots and Systems (IROS) en 2013 et 2014 ; Workshop on Speech and Language Processing for Assistive Technologies (2013) ; Workshop on Wireless Communications and User-centered Services in Pervasive Environments (WUSPE2013) ; International on Knowledge and Systems Engineering (KSE 2015/2016) ; International Conference on Affective Computing and Intelligent Interaction (ACII2017) ; International Conference on Multimedia Analysis and Pattern Recognition (MAPR2018).

J'ai été relecteur pour plusieurs journaux ou conférences (liste non exhaustive) : Interface Homme Machine (IHM) ; ACM/IEEE International Conference on Human Robot Interaction (HRI) ; International Conference on Robotics and Automation (ICRA), IEEE Robot and Human Interaction Communication (Ro-Man) ; International Conference on Intelligent Robots

and Systems (IROS) ; ACM International Joint Conference on Pervasive and Ubiquitous Computing ; ACM International Conference on Multimodal Interaction (ICMI) ; International Conference on Social Robotics (ICSR) ; International Journal On Advances in Internet Technology ; Journal of Robotics and Autonomous Systems ; journal of Interactive Technology and Smart Education ; International Journal of Social Robotics.

Présentations invitées et vulgarisation scientifique

2002 Tutorial à ESSLLI'02 (*European Summer School in Logic, Language and Information*, Trento - Italie) en août 2002 sur la reconnaissance de la parole et le passage à l'échelle (« *ASR and scalability* ») devant un public de non spécialistes de la reconnaissance de la parole.

2004-2015 Participation à 5 éditions de la fête de la science à l'Université et à l'*Inria*.

2013 Présentation de mes travaux et participation à table ronde de « *L'Homme 2.0 et son environnement – Entre rejet et fascination* » une journée interprofessionnelle (entrepreneurs, chercheurs, associations et institutionnels) organisée en décembre 2013 par l'Université de Lyon.

2014 « *Nuit des chercheurs* » en 2014 à Lyon. J'y ai présenté les recherches menées dans le projets PAL par les équipes *Inria* et réalisé une interview pour une radio locale (Radio Scoop).

2014 Présentation de mes recherches sur les aspects perception multimodale et maintien de personne fragiles à domicile à des groupes lycéens dans le cadre du programme **MathC+**. Dans le cadre du projet **INS**, je suis allé au lycée du Pont-De-Chéruy pour présenter mes activités de recherche. J'ai également reçu un groupe d'enseignants de l'académie de Grenoble à l'*Inria*.

2015 Présentation de mes travaux devant des enseignants du secondaire de toute l'académie de Grenoble au Lycée Argouges à Grenoble.

2016 Mise au point avec *Étienne Balit* d'une tête robotique qui détecte et suit les personnes du regard. Démonstration pendant l'exposition « *Monstru'Eux, vous trouvez ça normal ?*» pendant plus d'un mois.

2017 Présentation intitulée « *Pedestrian detection and behaviors modelling in Urban environment* » à SMIV 2017 - *Smart Mobility and Intelligent Vehicles*, Novembre 2017, Versailles, France.

2018 Présentation de mes travaux pour le séminaire « Intelligence Artificielle » de l'association BEST-Grenoble (*Board of European Students of Technology*) devant des élèves ingénieurs européens.

Logiciels

J'ai développé de nombreux logiciels aux cours de ces années. Les plus remarquables (ceux qui ont fait l'objet d'un dépôt à l'Agence de Protection des Programmes - APP) sont détaillés dans cette section. Vous trouverez certains de ces logiciels sur ma page GitHub[28].

- ***Online Movie Director*** (*aka* le Cameraman Automatique). Celui permet de réaliser automatiquement un flux vidéo et audio provenant de plusieurs microphones et caméras distribués dans l'environnement, en fonction des activités des personnes. Il peut servir à réaliser des cours ou des conférences tout en les diffusant en direct. Celui-ci fait l'objet d'un transfert industriel vers une startup.
- ***Détection de parole***. J'ai déposé un logiciel de **détection de parole** en environnement bruité qui est utilisé comme source d'information dans le *Online Movie Director*.
- ***OMiSCID***. Je suis le contributeur principal de la partie C++ d'**OMiSCID**, middleware permettant de déployer un ensemble de services dynamiquement et de construire une application de haut niveau en les composant automatiquement. Celui-ci est disponible sous licence MIT pour la communauté.
- ***Smart Servo Framework.*** Nous avons aussi mis à disposition de la communauté en GPL3.0 notre logiciel de contrôle de servomoteurs *Smart Servo Framework*[29] développé dans le projet PRAMAD.
- ***MobileRGBD***. Pour le projet MobileRGBD, j'ai redéveloppé une couche de contrôle de la plateforme du *Kompaï*, le logiciel embarqué n'étant pas suffisamment performant pour nos besoins. Avec Amaury Nègre, nous avons écrit un logiciel de planification de trajectoire. J'ai aussi développé *RGBDSyncSDK* pour l'enregistrement et la lecture synchrones des données en provenance de nombreux capteurs présents sur les robots mobiles (caméras, LIDAR, Kinect, …). Celui-ci est disponible pour la communauté.

[28] Voir https://github.com/Vaufreyd (dernière visite 04/2018).

[29] Voir https://github.com/emericg/SmartServoFramework (dernière visite 04/2018).

Annexes chapitre III

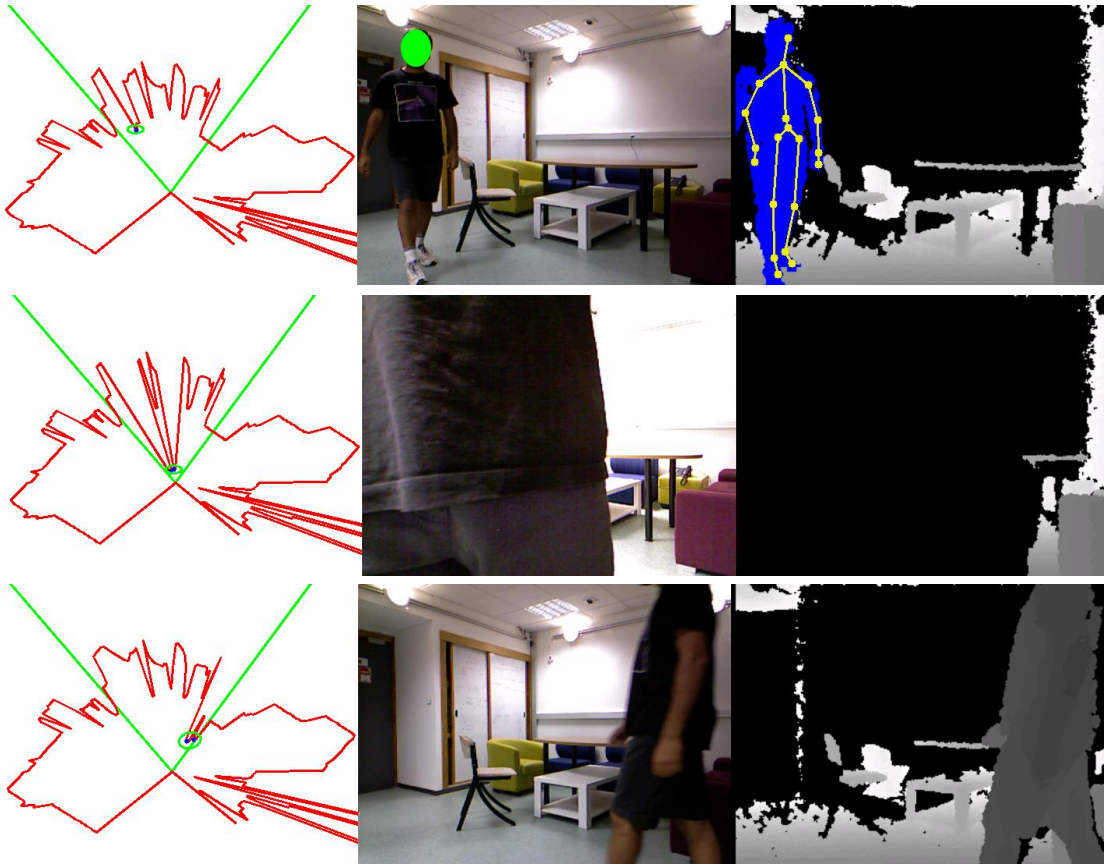


Figure 28 : Exemples de perception tirés de notre corpus de détection d'engagement. En haut, une personne s'approchant du robot avec l'intention d'interagir. Au centre, la personne interagit avec le robot via la tablette de celui-ci. En bas, une personne passe près du robot.

Annexes chapitre IV

Les figures ci-après montrent la sensibilité à la rotation pour la détection de personnes avec *Yolo v3* [161] (fig. 29) et *Mask_RCNN* [265] (fig. 30). Le premier est sensible à l'orientation avec des détections autours de certains angles seulement. *Mask_RCNN* semble sensible aux couleurs/textures n'arrivant à aucune détection sur le mannequin. Dans nos tests sur des images du corpus COCO [229], ce dernier montre comme les autres systèmes une sensibilité forte à l'orientation pour la détection sur des images.

Image initiale 25° 90° 140°

236° 280° 332°

Figure 29 : Détection d'objets avec différentes rotations d'images [161] (valeurs en degrés).

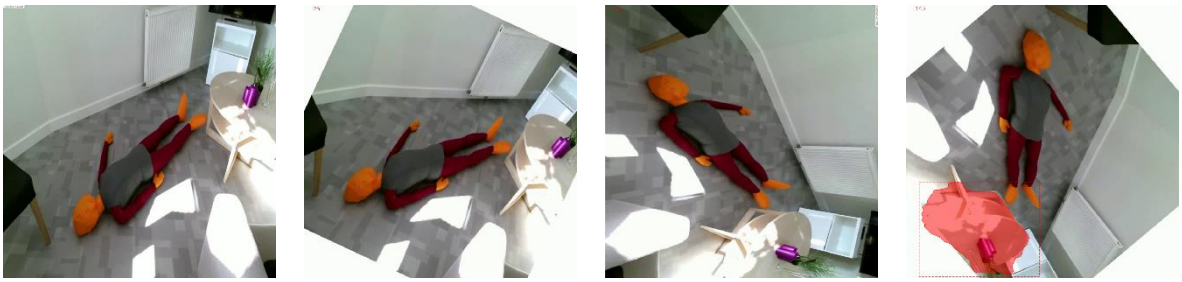


Image initiale 25° 90° 140°

220° 250° 332°

Figure 30 : Détection d'objets « *pixel wise* » avec différentes rotations d'images [265] (valeurs en degrés).

Références

[1] D. Vaufreydaz, «Statistical language modelling using Internet documents for continuous speech recognition,» 2002.

[2] A. Turing, «Computing machinery and intelligence,» vol. 59, 1950, p. 433.

[3] Y. LeCun, Y. Bengio et G. Hinton, «Deep learning,» *nature,* vol. 521, p. 436, 2015.

[4] Y. Taigman, M. Yang, M. Ranzato et L. Wolf, «Deepface: Closing the gap to human-level performance in face verification,» chez *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[5] I. McGraw, R. Prabhavalkar, R. Alvarez, M. G. Arenas, K. Rao, D. Rybach, O. Alsharif, H. Sak, A. Gruenstein, F. Beaufays et others, «Personalized speech recognition on mobile devices,» chez *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016.

[6] F. Ingrand et M. Ghallab, «Deliberation for autonomous robots: A survey,» *Artificial Intelligence,* vol. 247, pp. 10-44, 2017.

[7] M. Endsley, *Toward a theory of situation awarenss in dynamic systems,* vol. 37, 1995, pp. 32-64.

[8] R. Emonet, D. Vaufreydaz, P. Reignier et J. Letessier, «O3MiSCID, a Middleware for Pervasive Environments,» chez *1st IEEE International Workshop on Services Integration in Pervasive Environments*, Lyon, 2006.

[9] R. Emonet et D. Vaufreydaz, «Perceptive Services Composition using semantic language and distributed knowledge,» chez *Common Models and Patterns for Pervasive Computing at the 5th International Conference on Pervasive Computin*, Toronto (Ontario), Canada, 2007.

[10] R. Emonet et D. Vaufreydaz, «Usable developer-oriented Functionality Composition Language (UFCL): a Proposal for Semantic Description and Dynamic Composition of Services and Service Factories,» chez *4th IET International Conference on Intelligent Environments*, Seattle, 2008.

[11] R. Emonet, «Semantic Description of Services and Service Factories for Ambient Intelligence,» 2009.

[12] R. Barraquand, D. Vaufreydaz, R. Emonet et J.-P. Mercier, «Case Study of the OMiSCID Middleware: Wizard of Oz Experiment in Smart Environments,» chez *The Fourth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, Florence, 2010.

[13] R. Barraquand, D. Vaufreydaz, R. Emonet, A. Negre, J.-P. Mercier et P. Reignier, «OMiSCID 2.0, un intergiciel gratuit pour la construction d'applications distribuées,» chez *MajecStic*, Bordeaux, 2010.

[14] R. Barraquand, D. Vaufreydaz, R. Emonet, A. Nègre et P. Reignier, «The OMiSCID 2.0 Middleware: Usage and Experiments in Smart Environments,» *International Journal On Advances in Software,* vol. 4, pp. 231-243, 3 2012.

[15] V. C. Ta, W. Johal, M. Portaz, E. Castelli et D. Vaufreydaz, «The Grenoble System for the Social Touch Challenge at ICMI 2015,» chez *17th ACM International Conference on Multimodal Interaction (ICMI2015)*, Seattle, 2015.

[16] V. C. Ta, D. Vaufreydaz, T.-K. Dao et E. Castelli, «Smartphone-based User Location Tracking in Indoor Environment,» chez *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Madrid, 2016.

[17] J. Torres-Sospedra, A. Jiménez, S. Knauth, A. Moreira, Y. K. Beer, T. Fetzer, V.-C. Ta, R. M. Montoliu, F. Seco, G. M. Mendoza-Silva, O. Belmonte, A. Koukofikis, M. J. Nicolau, A. Costa, F. M. Meneses, F. Ebner, F. Deinzer, D. Vaufreydaz, T.-K. Dao et E. Castelli, «The Smartphone-Based Offline Indoor Location Competition at IPIN 2016: Analysis and Future Work,» *Sensors,* vol. 557, p. 17, 2017.

[18] V. C. Ta, «Smartphone-Based Indoor Positioning Using Wifi, Inertial Sensors And Blue-Tooth,» 2017.

[19] E. Balit, D. Vaufreydaz et P. Reignier, «Integrating Animation Artists into the Animation Design of Social Robots,» chez *ACM/IEEE Human-Robot Interaction 2016*, Christchurch, 2016.

[20] E. Balit, D. Vaufreydaz et P. Reignier, «PEAR: Prototyping Expressive Animated Robots - A framework for social robot prototyping,» chez *HUCAPP 2018 - 2nd International Conference on Human Computer Interaction Theory and Applications* , Funchal, 2018.

[21] J. J. Gibson, «The ecological approach to visual perception.,» 1979.

[22] P. Vasishta, D. Vaufreydaz et A. Spalanzani, «Natural Vision Based Method for Predicting Pedestrian Behaviour in Urban Environments,» chez *IEEE 20th International Conference on Intelligent Transportation Systems*, Yokohama, 2017.

[23] P. Vasishta, D. Vaufreydaz et A. Spalanzani, «Urban Pedestrian Behaviour Modelling using Natural Vision and Potential Fields,» chez *9th Workshop on Planning, Perception*

and Navigation for Intelligent Vehicles at the IEEE International Conference on Intelligent Robots and Systems, Vancouver, 2017.

[24] T. Guntz, R. Balzarini, D. Vaufreydaz et J. L. Crowley, «Multimodal Observation and Interpretation of Subjects Engaged in Problem Solving,» chez *1st Workshop on ``Behavior, Emotion and Representation: Building Blocks of Interaction''*, Bielefeld, 2017.

[25] T. Guntz, R. Balzarini, D. Vaufreydaz et J. Crowley, «Multimodal Observation and Classification of People Engaged in Problem Solving: Application to Chess Players,» *Multimodal Technologies and Interaction,* vol. 2, 2018.

[26] W. Benkaouar, «Autocalibration of Tracking Towers Equipped with an Omnidirectional camera and set of Microphones,» *MoSIG M1, Université Joseph Fourier (Grenoble, France),* 2011.

[27] E. A. Schegloff, «Body torque,» *Social Research,* pp. 535-596, 1998.

[28] W. Benkaouar et D. Vaufreydaz, «Multi-Sensors Engagement Detection with a Robot Companion in a Home Environment,» chez *Workshop on Assistance and Service robotics in a human environment at IEEE International Conference on Intelligent Robots and Systems (IROS2012)*, Vilamoura, 2012.

[29] D. Vaufreydaz, W. Johal et C. Combe, «Starting engagement detection towards a companion robot using multimodal features,» *Robotics and Autonomous Systems,* p. 25, 1 2015.

[30] A. Barbulescu, A. Begault, L. Boissieux, M.-P. Cani, M. Garcia, M. Portaz, A. Viand, P. Heinish, R. Dulery, R. Ronfard et D. Vaufreydaz, «Making Movies from Make-Believe Games,» chez *6th Workshop on Intelligent Cinematography and Editing (WICED 2017)*, Lyon, 2017.

[31] A. Barbulescu, M. Garcia, A. Begault, L. Boissieux, M.-P. Cani, M. Portaz, A. Viand, R. Dulery, P. Heinish, R. Ronfard et D. Vaufreydaz, «A system for creating virtual reality content from make-believe games,» chez *IEEE Virtual Reality 2017*, Los Angeles, United States, 2017.

[32] M. Portaz, M. Garcia, A. Barbulescu, A. Begault, L. Boissieux, M.-P. Cani, R. Ronfard et D. Vaufreydaz, «Figurines, a multimodal framework for tangible storytelling,» chez *WOCCI 2017 - 6th Workshop on Child Computer Interaction at ICMI 2017 - 19th ACM International Conference on Multi-modal Interaction*, Glasgow, 2017.

[33] E. Ferrier-Barbut, D. Vaufreydaz, J.-A. David, J. Lussereau, A. Spalanzani et J. Lussereau, «Personal space of autonomous car's passengers sitting in the driver's seat,» chez *The 2018 IEEE Intelligent Vehicles Symposium (IV'18)*, Changshu, 2018.

[34] P. Reignier, O. Brdiczka, D. Vaufreydaz, J. L. Crowley et J. Maisonnasse, «Context-aware environments: from specification to implementation,» *Expert Systems,* 2007.

[35] M. Danninger et R. Stiefelhagen, «A context-aware virtual secretary in a smart office environment,» chez *Proceedings of the 16th ACM international conference on Multimedia*, 2008.

[36] A. Waibel, T. Schultz, M. Bett, M. Denecke, R. Malkin, I. Rogina, R. Stiefelhagen et J. Yang, «SMaRT: The smart meeting room task at ISL,» chez *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, 2003.

[37] Y. Shi, W. Xie, G. Xu, R. Shi, E. Chen, Y. Mao et F. Liu, «The smart classroom: merging technologies for seamless tele-education,» *IEEE Pervasive Computing,* vol. 2, pp. 47-55, 2003.

[38] M. Weiser, «The Computer for the 21 st Century,» *Scientific american,* vol. 265, pp. 94-105, 1991.

[39] J. C. Augusto, H. Nakashima et H. Aghajan, «Ambient intelligence and smart environments: A state of the art,» chez *Handbook of ambient intelligence and smart environments*, Springer, 2010, pp. 3-31.

[40] J. L. Crowley, J. Coutaz, G. Rey et P. Reignier, «Perceptual components for context aware computing,» chez *International conference on ubiquitous computing*, 2002.

[41] S. Borkowski, «Steerable Interfaces for Interactive Environments,» 2006.

[42] M. Vacher, «Sound and Multimodal Analysis in Ambient Assisted Living,» 2011.

[43] H. A. Yanco et K. Z. Haigh, «Automation as caregiver: A survey of issues and technologies,» *Am. Assoc. Artif. Intell,* vol. 2, pp. 39-53, 2002.

[44] A. Lacombe, F. Rocaries, C. Dietrich, J. L. Baldinger, J. Boudy, F. Delavault, A. Deskata, M. Baer et A. Ozguler, «Open technical platform prototype and validation process model for patient at home medical monitoring system,» *BioMedSim 2005,* 2005.

[45] F. Metze, P. Gieselmann, H. Holzapfel, T. Kluge, I. Rogina, A. Waibel, M. Wölfel, J. L. Crowley, P. Reignier, D. Vaufreydaz, F. Bérard, B. Cohen, J. Coutaz, S. Rouillard, V. Arranz, M. Bertran et H. Rodriguez, «The ``FAME" Interactive Space,» chez *2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Edinburgh, 2005.

[46] D. Macho, J. Padrell, A. Abad, C. Nadeu, J. Hernando, J. McDonough, M. Wolfel, U. Klee, M. Omologo, A. Brutti et others, «Automatic speech activity detection, source localization, and speech recognition on the CHIL seminar corpus,» chez *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, 2005.

[47] J. Ramirez, J. C. Segura, C. Benitez, A. De La Torre et A. Rubio, «Efficient voice activity detection algorithms using long-term speech information,» *Speech communication,* vol. 42, pp. 271-287, 2004.

[48] A. Martin, D. Charlet et L. Mauuary, «Robust speech/non-speech detection using LDA applied to MFCC,» chez *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, 2001.

[49] L. F. Larnel, J.-L. Gauvain et M. Eskenazi, «BREF, a large vocabulary spoken corpus for French,» chez *Second european conference on speech communication and technology*, 1991.

[50] D. Vaufreydaz, R. Emonet et P. Reignier, «A Lightweight Speech Detection System for Perceptive Environments,» chez *3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Washington, 2006.

[51] Q. Wang, J. Du, X. Bao, Z.-R. Wang, L.-R. Dai et C.-H. Lee, «A universal VAD based on jointly trained deep neural networks,» chez *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[52] X.-L. Zhang et J. Wu, «Deep belief networks based voice activity detection,» *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 21, pp. 697-710, 2013.

[53] O. Brdiczka, «Learning Situation Models for Providing Context-Aware Services,» 2007.

[54] T. Gu, H. K. Pung et D. Q. Zhang, «Toward an OSGi-based infrastructure for context-aware applications,» *IEEE Pervasive Computing,* vol. 3, pp. 66-74, 2004.

[55] S. Dustdar et W. Schreiner, «A survey on web services composition,» *International journal of web and grid services,* vol. 1, pp. 1-30, 2005.

[56] A. Fillinger, L. Diduch, I. Hamchi, M. Hoarau, S. Degré et V. Stanford, «The nist data flow system ii: A standardized interface for distributed multimedia applications,» chez *World of Wireless, Mobile and Multimedia Networks, 2008. WoWMoM 2008. 2008 International Symposium on a*, 2008.

[57] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler et A. Y. Ng, «ROS: an open-source Robot Operating System,» chez *ICRA workshop on open source software*, 2009.

[58] J. S. Rellermeyer, G. Alonso et T. Roscoe, «R-OSGi: distributed applications through software modularization,» chez *ACM/IFIP/USENIX International Conference on Distributed Systems Platforms and Open Distributed Processing*, 2007.

[59] P. Reignier, S. Zaidenberg, R. Emonet, D. Vaufreydaz et J. Letesssier, «jOMiSCID, un intergiciel sous OSGi pour l'informatique ubiquitaire,» chez *Atelier de travail OSGi 2006*, Paris, 2006.

[60] J. J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne et K. Wilkinson, «Jena: implementing the semantic web recommendations,» chez *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, 2004.

[61] G. Klyne et J. J. Carroll, «Resource description framework (RDF): Concepts and abstract syntax,» 2006.

[62] M. Skubic, G. Alexander, M. Popescu, M. Rantz et J. Keller, «A smart home application to eldercare: Current status and lessons learned,» *Technology and Health Care,* vol. 17, pp. 183-201, 2009.

[63] M. Vacher, J.-F. Serignat, S. Chaillol, D. Istrate et V. Popescu, «Speech and sound use in a remote monitoring system for health care,» chez *International Conference on Text, Speech and Dialogue*, 2006.

[64] N. Zouba, F. Brémond, M. Thonnat, A. Anfosso, E. Pascual, P. Mallea, V. Mailland et O. Guerin, «A computer system to monitor older adults at home: Preliminary results,» *Gerontechnology Journal,* vol. 8, pp. 129-139, 2009.

[65] S. S. Intille, «Designing a home of the future,» *IEEE pervasive computing,* vol. 1, pp. 76-82, 2002.

[66] M. Klein, A. Schmidt et R. Lauer, «Ontology-centred design of an ambient middleware for assisted living: The case of soprano,» chez *Towards Ambient Intelligence: Methods for Cooperating Ensembles in Ubiquitous Environments (AIM-CU), 30th Annual German Conference on Artificial Intelligence (KI 2007), Osnabrück*, 2007.

[67] P. Chahuara, F. Portet et M. Vacher, «Context-aware decision making under uncertainty for voice-based control of smart home,» *Expert Systems with Applications,* vol. 75, pp. 63-79, 1 2017.

[68] K.-C. Kwak et S.-S. Kim, «Sound source localization with the aid of excitation source information in home robot environments,» *IEEE Transactions on Consumer Electronics,* vol. 54, 2008.

[69] G. E. P. Box, M. E. Muller et others, «A note on the generation of random normal deviates,» *The annals of mathematical statistics,* vol. 29, pp. 610-611, 1958.

[70] T. Gandhi et M. M. Trivedi, «Calibration of a reconfigurable array of omnidirectional cameras using a moving person,» chez *Proceedings of the ACM 2nd international workshop on Video surveillance & sensor networks*, 2004.

[71] M. Meingast, S. Oh et S. Sastry, «Automatic camera network localization using object image tracks,» chez *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007.

[72] P. Davidson et R. Piché, «A survey of selected indoor positioning methods for smartphones,» *IEEE Communications Surveys & Tutorials,* vol. 19, pp. 1347-1370, 2017.

[73] M. Bakula, P. Przestrzelski et R. Kazmierczak, «Reliable technology of centimeter GPS/GLONASS surveying in forest environments,» *IEEE Transactions on Geoscience and Remote Sensing,* vol. 53, pp. 1029-1038, 2015.

[74] T. Taketomi, H. Uchiyama et S. Ikeda, «Visual SLAM algorithms: a survey from 2010 to 2016,» *IPSJ Transactions on Computer Vision and Applications,* vol. 9, p. 16, 2017.

[75] T. Schöps, J. Engel et D. Cremers, «Semi-dense visual odometry for AR on a smartphone,» chez *Mixed and Augmented Reality (ISMAR), 2014 IEEE International Symposium on*, 2014.

[76] C. Marouane, A. Ebert, C. Linnhoff-Popien et M. Christil, «Step and activity detection based on the orientation and scale attributes of the SURF algorithm,» chez *Indoor Positioning and Indoor Navigation (IPIN), 2016 International Conference on*, 2016.

[77] A. R. J. Ruiz et F. S. Granja, «Comparing Ubisense, BeSpoon, and DecaWave UWB location systems: indoor performance analysis,» *IEEE Transactions on instrumentation and Measurement,* vol. 66, pp. 2106-2117, 2017.

[78] C. Gentner et T. Jost, «Indoor positioning using time difference of arrival between multipath components,» chez *Indoor Positioning and Indoor Navigation (IPIN), 2013 International Conference on*, 2013.

[79] P. Kontkanen, P. Myllymaki, T. Roos, H. Tirri, K. Valtonen et H. Wettig, «Topics in probabilistic location estimation in wireless networks,» chez *Personal, Indoor and Mobile Radio Communications, 2004. PIMRC 2004. 15th IEEE International Symposium on*, 2004.

[80] C. Rizos, A. G. Dempster, B. Li et J. Salter, «Indoor positioning techniques based on wireless LAN,» 2007.

[81] N. Dinh-Van, F. Nashashibi, N. Thanh-Huong et E. Castelli, «Indoor Intelligent Vehicle localization using WiFi received signal strength indicator,» chez *Microwaves for Intelligent Mobility (ICMIM), 2017 IEEE MTT-S International Conference on*, 2017.

[82] E. Jedari, Z. Wu, R. Rashidzadeh et M. Saif, «Wi-Fi based indoor location positioning employing random forest classifier,» chez *Indoor Positioning and Indoor Navigation (IPIN), 2015 International Conference on*, 2015.

[83] S. Liu, Y. Jiang et A. Striegel, «Face-to-face proximity estimationusing bluetooth on smartphones,» *IEEE Transactions on Mobile Computing,* vol. 13, pp. 811-823, 2014.

[84] J. Jun, Y. Gu, L. Cheng, B. Lu, J. Sun, T. Zhu et J. Niu, «Social-Loc: Improving indoor localization with social sensing,» chez *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, 2013.

[85] L. Pei, R. Chen, J. Liu, H. Kuusniemi, T. Tenhunen et Y. Chen, «Using inquiry-based Bluetooth RSSI probability distributions for indoor positioning,» *Journal of Global Positioning Systems,* vol. 9, pp. 122-130, 2010.

[86] S. H. Kaminski, Communication Models, 2002.

[87] M. Argyle, Bodily communication, Routledge, 2013.

[88] P. E. Bull, Posture & gesture, vol. 16, Elsevier, 2016.

[89] A. Vinciarelli, M. Pantic et H. Bourlard, «Social signal processing: Survey of an emerging domain,» *Image and vision computing,* vol. 27, pp. 1743-1759, 2009.

[90] M. Pantic, R. Cowie, F. D'Errico, D. Heylen, M. Mehu, C. Pelachaud, I. Poggi, M. Schroeder et A. Vinciarelli, «Social signal processing: the research agenda,» chez *Visual analysis of humans*, Springer, 2011, pp. 511-538.

[91] M. Paetzel, I. Hupont, G. Varni, M. Chetouani, C. Peters et G. Castellano, «Exploring the Link between Self-assessed Mimicry and Embodiment in HRI,» chez *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017.

[92] D. Vernon, C. Von Hofsten et L. Fadiga, A roadmap for cognitive development in humanoid robots, vol. 11, Springer Science & Business Media, 2011.

[93] R. Barraquand, «Designing sociable technologies,» 2012.

[94] K. Dautenhahn, «Human-Robot Interaction,» chez *The Encyclopedia of Human-Computer Interaction, 2nd Ed.*, M. Soegaard et R. F. Dam, Éds., Denmark, : The Interaction Design Foundation, 2014.

[95] C. Schroeter, S. Mueller, M. Volkhardt, E. Einhorn, C. Huijnen, H. Heuvel, A. Berlo, A. Bley et H.-M. Gross, «Realization and user evaluation of a companion robot for people with mild cognitive impairments,» chez *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, 2013.

[96] J. Fasola et M. J. Mataric, «Socially assistive robot exercise coach: Motivating older adults to engage in physical exercise,» chez *Experimental Robotics*, 2013.

[97] D. Fischinger, P. Einramhof, W. Wohlkinger, K. Papoutsakis, P. Mayer, P. Panek, T. Koertner, S. Hofmann, A. Argyros, M. Vincze et others, «Hobbit-the mutual care robot,» chez *Workshop on assistance and service robotics in a human environment*

workshop in conjunction with IEEE/RSJ international conference on intelligent robots and systems, 2013.

[98] D. Feil-Seifer et M. J. Mataric, «Defining socially assistive robotics,» chez *Rehabilitation Robotics, 2005. ICORR 2005. 9th International Conference on*, 2005.

[99] M. Coeckelbergh, C. Pop, R. Simut, A. Peca, S. Pintea, D. David et B. Vanderborght, «A survey of expectations about the role of robots in robot-assisted therapy for children with asd: Ethical acceptability, trust, sociability, appearance, and attachment,» *Science and engineering ethics,* vol. 22, pp. 47-65, 2016.

[100] N. C. Kramer, S. Eimler, A. Von Der Pütten et S. Payr, «Theory of companions: what can theoretical models contribute to applications and understanding of human-robot interaction?,» *Applied Artificial Intelligence,* vol. 25, pp. 474-502, 2011.

[101] S. Pesty et D. Duhaut, «Artificial Companion: building a impacting relation,» chez *Robotics and Biomimetics (ROBIO), 2011 IEEE International Conference on*, 2011.

[102] H.-M. Gross, S. Mueller, C. Schroeter, M. Volkhardt, A. Scheidig, K. Debes, K. Richter et N. Doering, «Robot companion for domestic health assistance: Implementation, test and case study under everyday conditions in private apartments,» chez *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, 2015.

[103] H. Knight, «Eight lessons learned about non-verbal interactions through robot theater,» chez *International Conference on Social Robotics*, 2011.

[104] J. L. Burke, R. R. Murphy, E. Rogers, V. J. Lumelsky et J. Scholtz, «Final report for the DARPA/NSF interdisciplinary study on human-robot interaction,» *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* vol. 34, pp. 103-112, 2004.

[105] C. L. Sidner et C. Lee, «Engagement rules for human-robot collaborative interactions,» chez *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, 2003.

[106] L. Wang, P.-L. P. Rau, V. Evers, B. K. Robinson et P. Hinds, «When in Rome: the role of culture & context in adherence to robot recommendations,» chez *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, 2010.

[107] K. A. Tahboub, «Intelligent human-machine interaction based on dynamic bayesian networks probabilistic intention recognition,» *Journal of Intelligent and Robotic Systems,* vol. 45, pp. 31-52, 2006.

[108] S. Koo et D.-S. Kwon, «Recognizing human intentional actions from the relative movements between human and robot,» chez *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*, 2009.

[109] M. L. Walters, K. Dautenhahn, R. Te Boekhorst, K. L. Koay, D. S. Syrdal et C. L. Nehaniv, «An empirical framework for human-robot proxemics,» *Procs of New Frontiers in Human-Robot Interaction,* 2009.

[110] J. Rios-Martinez, A. Spalanzani et C. Laugier, «From Proxemics Theory to Socially-Aware Navigation: A Survey,» *International Journal of Social Robotics,* 4 2015.

[111] G. Castellano, A. Pereira, I. Leite, A. Paiva et P. W. McOwan, «Detecting user engagement with a robot companion using task and social interaction-based features,» chez *Proceedings of the 2009 international conference on Multimodal interfaces*, 2009.

[112] H. Zhao et R. Shibasaki, «A novel system for tracking pedestrians using multiple single-row laser-range scanners,» *IEEE Transactions on systems, man, and cybernetics-Part A: systems and humans,* vol. 35, pp. 283-291, 2005.

[113] T. Baltrušaitis, P. Robinson et L.-P. Morency, «Openface: an open source facial behavior analysis toolkit,» chez *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, 2016.

[114] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan et A. Paiva, «Automatic analysis of affective postures and body motion to detect engagement with a game companion,» chez *Proceedings of the 6th international conference on Human-robot interaction*, 2011.

[115] R. Mead, A. Atrash et M. J. Mataric, «Proxemic feature recognition for interactive robots: automating metrics from the social sciences,» chez *International conference on social robotics*, 2011.

[116] P. Holthaus, K. Pitsch et S. Wachsmuth, «How can I help?,» *International Journal of Social Robotics,* vol. 3, pp. 383-393, 2011.

[117] H. Hüttenrauch, K. S. Eklundh, A. Green et E. A. Topp, «Investigating spatial relationships in human-robot interaction,» chez *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, 2006.

[118] E. T. Hall, «The hidden dimension,» 1966.

[119] G. L. Oliveira, A. Valada, C. Bollen, W. Burgard et T. Brox, «Deep learning for human part discovery in images,» chez *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, 2016.

[120] Z. Cao, T. Simon, S.-E. Wei et Y. Sheikh, «Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,» chez *CVPR*, 2017.

[121] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas et C. Theobalt, «Vnect: Real-time 3d human pose estimation with a single rgb camera,» *ACM Transactions on Graphics (TOG),* vol. 36, p. 44, 2017.

[122] J. Brauer et M. Arens, «Reconstructing the missing dimension: From 2d to 3d human pose estimation,» chez *Proc. of REACTS workshop, in conj. with Int. Conf. of Computer Analysis of Images and Patterns (CAIP2011)*, 2011.

[123] G. Rogez, P. Weinzaepfel et C. Schmid, «LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images,» *arXiv preprint arXiv:1803.00455,* 2018.

[124] H. Peng, F. Long et C. Ding, «Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,» *IEEE Transactions on pattern analysis and machine intelligence,* vol. 27, pp. 1226-1238, 2005.

[125] B. J. Scholl et P. D. Tremoulet, «Perceptual causality and animacy,» *Trends in cognitive sciences,* vol. 4, pp. 299-309, 2000.

[126] B. Friedman, P. H. Kahn Jr et J. Hagman, «Hardware companions?: What online AIBO discussion forums reveal about the human-robotic relationship,» chez *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2003.

[127] L. D. Riek, T.-C. Rabinowitch, B. Chakrabarti et P. Robinson, «Empathizing with robots: Fellow feeling along the anthropomorphic spectrum,» chez *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, 2009.

[128] M. Mori, «The uncanny valley,» *Energy,* vol. 7, pp. 33-35, 1970.

[129] E. S. Kim, L. D. Berkovits, E. P. Bernier, D. Leyzberg, F. Shic, R. Paul et B. Scassellati, «Social robots as embedded reinforcers of social behavior in children with autism,» *Journal of autism and developmental disorders,* vol. 43, pp. 1038-1049, 2013.

[130] P. Worthy, M. Boden, A. Karimi, J. Weigel, B. Matthews, K. Hensby, S. Heath, P. Pounds, J. Taufatofua, M. Smith et others, «Children's expectations and strategies in interacting with a wizard of oz robot,» chez *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*, 2015.

[131] J. Broekens, M. Heerink, H. Rosendal et others, «Assistive social robots in elderly care: a review,» *Gerontechnology,* vol. 8, pp. 94-103, 2009.

[132] L. Damiano, P. Dumouchel et H. Lehmann, «Towards human--robot affective co-evolution overcoming oppositions in constructing emotions and empathy,» *International Journal of Social Robotics,* vol. 7, pp. 7-18, 2015.

[133] J. Wrobel, M. Pino, P. Wargnier et A.-S. Rigaud, «Robots and virtual agents to assist older adults: A review of present day trends in gerontechnology,» *NPG Neurologie - Psychiatrie - Gériatrie ,* vol. 14, pp. 184-193, 2014.

[134] S. S. Kwak, Y. Kim, E. Kim, C. Shin et K. Cho, «What makes people empathize with an emotional robot?: The impact of agency and physical embodiment on human empathy for a robot,» chez *RO-MAN, 2013 IEEE*, 2013.

[135] M. Watanabe, K. Ogawa et H. Ishiguro, «Field study: can androids be a social entity in the real world?,» chez *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, 2014.

[136] W. B. Johal, «Companion Robots Behaving with Style: Towards Plasticity in Social Human-Robot Interaction,» 2015.

[137] P. Ekman, «An argument for basic emotions,» *Cognition & emotion,* vol. 6, pp. 169-200, 1992.

[138] T. Klein, G. J. Gelderblom, L. Witte et S. Vanstipelen, «Evaluation of short term effects of the IROMEC robotic toy for children with developmental disabilities,» chez *Rehabilitation Robotics (ICORR), 2011 IEEE International Conference on*, 2011.

[139] J. Saez-Pons, H. Lehmann, D. S. Syrdal et K. Dautenhahn, «Development of the sociability of non-anthropomorphic robot home companions,» chez *Development and Learning and Epigenetic Robotics (ICDL-Epirob), 2014 Joint IEEE International Conferences on*, 2014.

[140] P. Wargnier, G. Carletti, Y. Laurent-Corniquet, S. Benveniste, P. Jouvelot et A.-S. Rigaud, «Field Evaluation with Cognitively-Impaired Older Adults of Attention Management in the Embodied Conversational Agent Louise,» chez *4th International Conference on Serious Games and Applications for Health (IEEE SeGAH 2016)*, 2016.

[141] F. Badeig, P. Wargnier, M. Pino, P. De Oliveira Lopes, E. Grange, J. L. Crowley, A.-S. Rigaud et D. Vaufreydaz, «Impact of Head Motion on the Assistive Robot Expressiveness - Evaluation with Elderly Persons,» chez *1st International Workshop on Affective Computing for Social Robotics Workshop at the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, New York, United States, 2016.

[142] M. Aeberhard, S. Rauch, M. Bahram, G. Tanzmeister, J. Thomas, Y. Pilat, F. Homm, W. Huber et N. Kaempchen, «Experience, results and lessons learned from automated driving on Germany's highways,» *IEEE Intelligent Transportation Systems Magazine,* vol. 7, pp. 42-57, 2015.

[143] X. Geng, H. Liang, B. Yu, P. Zhao, L. He et R. Huang, «A Scenario-Adaptive Driving Behavior Prediction Approach to Urban Autonomous Driving,» *Applied Sciences,* vol. 7, 2017.

[144] L. Delobel, R. Aufrere, R. Chapuis, C. Debain et T. Chateau, «Towards automated map updating for mobile robot localization,» chez *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017.

[145] V. Viswanathan et R. Hussein, «Applications of Image Processing and Real-Time embedded Systems in Autonomous Cars: A Short Review,» *International Journal of Image Processing (IJIP),* vol. 11, p. 35, 2017.

[146] J. M. Armingol, J. Alfonso, N. Aliane, M. Clavijo, S. Campos-Cordobés, A. Escalera, J. Ser, J. Fernández, F. García, F. Jiménez et others, «Environmental Perception for Intelligent Vehicles,» chez *Intelligent Vehicles*, 2018.

[147] X. Wang, «Intelligent multi-camera video surveillance: A review,» *Pattern recognition letters,* vol. 34, pp. 3-19, 2013.

[148] R. Prakash, H. Malviya, A. Naudiyal, R. Singh et A. Gehlot, «An Approach to Inter-vehicle and Vehicle-to-Roadside Communication for Safety Measures,» chez *Intelligent Communication, Control and Devices*, Springer, 2018, pp. 1603-1610.

[149] J. Lussereau, P. Stein, J.-A. David, L. Rummelhard, A. Negre, C. Laugier, N. Vignard et G. Othmezouri, «Integration of ADAS algorithm in a Vehicle Prototype,» chez *IEEE International Workshop on Advanced Robotics and its Social Impacts ARSO 2015*, 2015.

[150] J. Varadarajan et J.-M. Odobez, «Topic models for scene analysis and abnormality detection,» chez *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, 2009.

[151] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu et M. S. Lew, «Deep learning for visual understanding: A review,» *Neurocomputing,* vol. 187, pp. 27-48, 2016.

[152] E. Papadimitriou, G. Yannis et J. Golias, «A critical assessment of pedestrian behaviour models,» *Transportation Research Part F: Traffic Psychology and Behaviour,* vol. 12, pp. 242-255, 2009.

[153] K. Yamaguchi, A. C. Berg, L. E. Ortiz et T. L. Berg, «Who are you with and where are you going?,» chez *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011.

[154] S. Pellegrini, A. Ess, K. Schindler et L. Gool, «You'll never walk alone: Modeling social behavior for multi-target tracking,» chez *2009 IEEE 12th International Conference on Computer Vision*, 2009.

[155] B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey et S. Srinivasa, «Planning-based prediction for pedestrians,» chez *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, 2009.

[156] K. M. Kitani, B. D. Ziebart, J. A. Bagnell et M. Hebert, «Activity forecasting,» chez *European Conference on Computer Vision*, 2012.

[157] T. Bandyopadhyay, C. Z. Jie, D. Hsu, H. Marcelo, A. Jr, D. Rus et E. Frazzoli, «Intention-Aware Pedestrian Avoidance,» *The 13th International Symposium on Experimental Robotics,* pp. 963-977, 2013.

[158] D. Vasquez, T. Fraichard et C. Laugier, «Incremental learning of statistical motion patterns with growing hidden markov models,» *IEEE Transactions on Intelligent Transportation Systems,* vol. 10, pp. 403-416, 2009.

[159] I. Pérez-Hurtado, J. Capitán, F. Caballero et L. Merino, «An extension of GHMMs for environments with occlusions and automatic goal discovery for person trajectory prediction,» chez *Mobile Robots (ECMR), 2015 European Conference on*, 2015.

[160] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth et B. Schiele, «The cityscapes dataset for semantic urban scene understanding,» chez *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[161] J. Redmon et A. Farhadi, «YOLOv3: An Incremental Improvement,» *arXiv,* 2018.

[162] B. Hillier, A. Penn, J. Hanson, T. Grajewski et J. Xu, «Natural movement: or, configuration and attraction in urban pedestrian movement,» *Environment and Planning B: planning and design,* vol. 20, pp. 29-66, 1993.

[163] O. Khatib, «Real-time obstacle avoidance for manipulators and mobile robots,» chez *Autonomous robot vehicles*, Springer, 1986, pp. 396-404.

[164] M. T. Wolf et J. W. Burdick, «Artificial potential functions for highway driving with collision avoidance,» chez *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, 2008.

[165] J. S. Rowlinson, «The Yukawa potential,» *Physica A: Statistical Mechanics and its Applications,* vol. 156, pp. 15-34, 1989.

[166] D. Vasquez, T. Fraichard et C. Laugier, «Incremental Learning of Statistical Motion Patterns With Growing Hidden Markov Models,» *IEEE Transactions on Intelligent Transportation Systems,* vol. 10, pp. 403-416, 9 2009.

[167] R. R. Hoffman, «How can expertise be defined? Implications of research from cognitive psychology,» chez *Exploring expertise*, Springer, 1998, pp. 81-100.

[168] N. Charness, E. M. Reingold, M. Pomplun et D. M. Stampe, «The perceptual aspect of skilled performance in chess: Evidence from eye movements,» *Memory & cognition,* vol. 29, pp. 1146-1152, 2001.

[169] R. W. Picard, «Affective computing: challenges,» *International Journal of Human-Computer Studies,* vol. 59, pp. 55-64, 2003.

[170] A. Kleinsmith et N. Bianchi-Berthouze, «Affective body expression perception and recognition: A survey,» *IEEE Transactions on Affective Computing,* vol. 4, pp. 15-33, 2013.

[171] B. Schuller, F. Weninger, Y. Zhang, F. Ringeval, A. Batliner, S. Steidl, F. Eyben, E. Marchi, A. Vinciarelli, K. Scherer et others, «Affective and Behavioural Computing: Lessons Learnt from the First Computational Paralinguistics Challenge,» *Computer Speech & Language,* 2018.

[172] X. Du, M. El-Khamy, J. Lee et L. Davis, «Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection,» chez *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, 2017.

[173] S. Zhang, R. Benenson, M. Omran, J. Hosang et B. Schiele, «Towards reaching human performance in pedestrian detection,» *IEEE transactions on pattern analysis and machine intelligence,* vol. 40, pp. 973-986, 2018.

[174] S. Das, M. Koperski, F. Bremond et G. Francesca, «A Fusion of Appearance based CNNs and Temporal evolution of Skeleton with LSTM for Daily Living Action Recognition,» *arXiv preprint arXiv:1802.00421,* 2018.

[175] M. Fridin, «Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education,» *Computers & education,* vol. 70, pp. 53-64, 2014.

[176] S. Benford, B. B. Bederson, K.-P. AAkesson, V. Bayon, A. Druin, P. Hansson, J. P. Hourcade, R. Ingram, H. Neale, C. O'Malley et others, «Designing storytelling technologies to encouraging collaboration between young children,» chez *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 2000.

[177] J. A. Fails, A. Druin et M. L. Guha, «Interactive storytelling: interacting with people, environment, and technology,» *International Journal of Arts and Technology,* vol. 7, pp. 112-124, 2014.

[178] B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes et L.-P. Morency, «The future belongs to the curious: Towards automatic understanding and recognition of curiosity in children,» chez *Workshop on Child Computer Interaction*, 2016.

[179] R. Miletitch, N. Sabouret et M. Ochs, «Susciter l'émotion dans la narration automatique,» *Technique et Science Informatiques,* vol. 31, pp. 477-501, 2012.

[180] D. Harley, J. H. Chu, J. Kwan et A. Mazalek, «Towards a framework for tangible narratives,» chez *Proceedings of the TEI'16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction*, 2016.

[181] C. Sylla, S. Gonçalves, P. Brito, P. Branco et C. Coutinho, «A tangible platform for mixing and remixing narratives,» chez *Advances in Computer Entertainment*, Springer, 2013, pp. 630-633.

[182] J. H. Chu, P. Clifton, D. Harley, J. Pavao et A. Mazalek, «Mapping place: Supporting cultural learning through a lukasa-inspired tangible tabletop museum exhibit,» chez *Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction*, 2015.

[183] A. Mazalek et M. Nitsche, «Tangible interfaces for real-time 3D virtual environments,» chez *Proceedings of the international conference on Advances in computer entertainment technology*, 2007.

[184] W. Yoshizaki, Y. Sugiura, A. C. Chiou, S. Hashimoto, M. Inami, T. Igarashi, Y. Akazawa, K. Kawachi, S. Kagami et M. Mochimaru, «An actuated physical puppet as an input device for controlling a digital manikin,» chez *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011.

[185] S.-Y. Lin, C.-K. Shie, S.-C. Chen et Y.-P. Hung, «Action recognition for human-marionette interaction,» chez *Proceedings of the 20th ACM international conference on Multimedia*, 2012.

[186] S. Gupta, S. Jang et K. Ramani, «Puppetx: A framework for gestural interactions with user constructed playthings,» chez *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*, 2014.

[187] O. Mayora, C. Costa et A. Papliatseyeu, «ITheater puppets tangible interactions for storytelling,» chez *International Conference on Intelligent Technologies for Interactive Entertainment*, 2009.

[188] F. Lu, F. Tian, Y. Jiang, X. Cao, W. Luo, G. Li, X. Zhang, G. Dai et H. Wang, «ShadowStory: creative and collaborative digital storytelling inspired by cultural heritage,» chez *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011.

[189] R. El Kaliouby et P. Robinson, «Real-time inference of complex mental states from facial expressions and head gestures,» chez *Real-time vision for human-computer interaction*, Springer, 2005, pp. 181-200.

[190] T. Baltrušaitis, D. McDuff, N. Banda, M. Mahmoud, R. El Kaliouby, P. Robinson et R. Picard, «Real-time inference of mental states from facial expressions and upper body gestures,» chez *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011.

[191] P. Ekman et W. V. Friesen, «Nonverbal leakage and clues to deception,» *Psychiatry,* vol. 32, pp. 88-106, 1969.

[192] M.-Z. Poh, D. J. McDuff et R. W. Picard, «Advancements in noncontact, multiparameter physiological measurements using a webcam,» *IEEE transactions on biomedical engineering,* vol. 58, pp. 7-11, 2011.

[193] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook et R. Moore, «Real-time human pose recognition in parts from single depth images,» *Communications of the ACM,* vol. 56, pp. 116-124, 2013.

[194] P. Vassallo, «Thinking, Fast and Slow,» *ETC.: A Review of General Semantics,* vol. 69, pp. 480-481, 2012.

[195] R. Balzarini, A. Dalmasso et M. Murat, «a Study on Mental Representations for Realistic Visualization the Particular Case of Ski Trail Mapping,» *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences,* vol. 40, p. 495, 2015.

[196] L. Paletta, A. Dini, C. Murko, S. Yahyanejad, M. Schwarz, G. Lodron, S. Ladstatter, G. Paar et R. Velik, «Towards real-time probabilistic evaluation of situation awareness from human gaze in human-robot interaction,» chez *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017.

[197] N. Charness, «The impact of chess research on cognitive science,» *Psychological research,* vol. 54, pp. 4-9, 1992.

[198] E. M. Reingold et N. Charness, «Perception in chess: Evidence from eye movements,» *Cognitive processes in eye guidance,* pp. 325-354, 2005.

[199] E. Goeleven, R. De Raedt, L. Leyman et B. Verschuere, «The Karolinska directed emotional faces: a validation study,» *Cognition and emotion,* vol. 22, pp. 1094-1118, 2008.

[200] J. A. Harrigan, «Self-touching as an indicator of underlying affect and language processes,» *Social Science & Medicine,* vol. 20, pp. 1161-1168, 1985.

[201] G. Bijlstra et R. Dotsch, «FaceReader 4 emotion classification performance on images from the Radboud Faces Database,» 2011. [En ligne]. Available: http://gijsbijlstra.nl/wp-content/uploads/2012/02/TechnicalReport_FR4_RaFD.pdf. [Accès le 01 02 2018].

[202] C. Ehmke et S. Wilson, «Identifying web usability problems from eye-tracking data,» chez *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1*, 2007.

[203] D. Wilkins, «Using patterns and plans in chess,» chez *Readings in Artificial Intelligence*, Elsevier, 1981, pp. 390-409.

[204] H. Simon et W. Chase, «Skill in chess,» chez *Computer chess compendium*, Springer, 1988, pp. 175-188.

[205] R. Levinson et R. Snyder, «Adaptive pattern-oriented chess,» chez *Machine Learning Proceedings 1991*, Elsevier, 1991, pp. 85-89.

[206] M. Kangas, R. Korpelainen, I. Vikman, L. Nyberg et T. Jamsa, «Sensitivity and false alarm rate of a fall sensor in long-term fall detection in the elderly,» *Gerontology,* vol. 61, pp. 61-68, 2015.

[207] A. Dubois et F. Charpillet, «Measuring frailty and detecting falls for elderly home care using depth camera,» *Journal of ambient intelligence and smart environments,* vol. 9, pp. 469-481, 6 2017.

[208] S. Hernandez-Mendez, C. Maldonado-Mendez, A. Marin-Hernandez et H. V. Rios-Figueroa, «Detecting falling people by autonomous service robots: A ROS module integration approach,» chez *2017 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, 2017.

[209] Z. Zhang, C. Conly et V. Athitsos, «A Survey on Vision-based Fall Detection,» chez *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, New York, NY, USA, 2015.

[210] V. Bevilacqua, N. Nuzzolese, D. Barone, M. Pantaleo, M. Suma, D. D'Ambruoso, A. Volpe, C. Loconsole et F. Stroppa, «Fall detection in indoor environment with kinect sensor,» chez *Innovations in Intelligent Systems and Applications (INISTA) Proceedings, 2014 IEEE International Symposium on*, 2014.

[211] K. Adhikari, H. Bouchachia et H. Nait-Charif, «Activity recognition for indoor fall detection using convolutional neural network,» chez *Machine Vision Applications (MVA), 2017 Fifteenth IAPR International Conference on*, 2017.

[212] M. Daher, M. E. B. El Najjar, A. Diab, M. Khalil, A. Dib et F. Charpillet, «Ambient assistive living system using RGB-D camera,» chez *Advances in Biomedical Engineering (ICABME), 2017 Fourth International Conference on*, 2017.

[213] R. Girshick, J. Shotton, P. Kohli, A. Criminisi et A. Fitzgibbon, «Efficient regression of general-activity human poses from depth images,» chez *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011.

[214] M. Munaro et E. Menegatti, «Fast RGB-D people tracking for service robots,» *Autonomous Robots,* vol. 37, pp. 227-242, 2014.

[215] M. Mubashir, L. Shao et L. Seed, «A survey on fall detection: Principles and approaches,» *Neurocomputing,* vol. 100, pp. 144-152, 2013.

[216] D. Vaufreydaz et A. Nègre, «MobileRGBD, An Open Benchmark Corpus for mobile RGB-D Related Algorithms,» chez *13th International Conference on Control, Automation, Robotics and Vision*, Singapour, 2014.

[217] K. Berger, «The role of rgb-d benchmark datasets: an overview,» *arXiv preprint arXiv:1310.2053,* 2013.

[218] Y. LeCun, Y. Bengio et others, «Convolutional networks for images, speech, and time series,» *The handbook of brain theory and neural networks,* vol. 3361, p. 1995, 1995.

[219] K. Fort, G. Adda, B. Sagot, J. Mariani et A. Couillault, «Crowdsourcing for Language Resource Development: Criticisms About Amazon Mechanical Turk Overpowering Use,» chez *Human Language Technology Challenges for Computer Science and Linguistics*, vol. 8387, Z. Vetulani et J. Mariani, Éds., Springer International Publishing, 2014, pp. 303-314.

[220] C. G. Harris, «Dirty deeds done dirt cheap: A darker side to crowdsourcing,» chez *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, 2011.

[221] A. Krizhevsky, I. Sutskever et G. E. Hinton, «Imagenet classification with deep convolutional neural networks,» chez *Advances in neural information processing systems*, 2012.

[222] M. Ye, X. Wang, R. Yang, L. Ren et M. Pollefeys, «Accurate 3d pose estimation from a single depth image,» chez *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011.

[223] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler et K. Murphy, «Towards accurate multiperson pose estimation in the wild,» *arXiv preprint arXiv:1701.01779,* vol. 8, 2017.

[224] K. He, G. Gkioxari, P. Dollár et R. Girshick, «Mask r-cnn,» chez *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[225] S.-E. Wei, V. Ramakrishna, T. Kanade et Y. Sheikh, «Convolutional pose machines,» chez *CVPR*, 2016.

[226] T. Simon, H. Joo, I. Matthews et Y. Sheikh, «Hand Keypoint Detection in Single Images using Multiview Bootstrapping,» chez *CVPR*, 2017.

[227] N. Wojke, A. Bewley et D. Paulus, «Simple Online and Realtime Tracking with a Deep Association Metric,» *arXiv preprint arXiv:1703.07402,* 2017.

[228] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun et A. Yuille, «Detect what you can: Detecting and representing objects using holistic models and body parts,» chez

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[229] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár et C. L. Zitnick, «Microsoft coco: Common objects in context,» chez *European conference on computer vision*, 2014.

[230] L. N. Smith, «Cyclical learning rates for training neural networks,» chez *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, 2017.

[231] J. Han, L. Shao, D. Xu et J. Shotton, «Enhanced computer vision with microsoft kinect sensor: A review,» *IEEE transactions on cybernetics,* vol. 43, pp. 1318-1334, 2013.

[232] C. Zhang, S. Bengio, M. Hardt, B. Recht et O. Vinyals, «Understanding deep learning requires rethinking generalization,» *arXiv preprint arXiv:1611.03530,* 2016.

[233] K. Zhou et B. Kainz, «Efficient Image Evidence Analysis of CNN Classification Results,» *arXiv preprint arXiv:1801.01693,* 2018.

[234] P. Pierre, C. Coulon et D. Gardner, *Les véhicules autonomes. Quelles responsabilités ? Quelle assurance ?,* 2017.

[235] M. Zook, S. Barocas, K. Crawford, E. Keller, S. P. Gangadharan, A. Goodman, R. Hollander, B. A. Koenig, J. Metcalf, A. Narayanan et others, «Ten simple rules for responsible big data research,» 2017.

[236] A. Nguyen, J. Yosinski et J. Clune, «Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,» chez *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[237] J. Su, D. V. Vargas et K. Sakurai, «One pixel attack for fooling deep neural networks,» *CoRR,* vol. abs/1710.08864, 2017.

[238] M. D. Zeiler et R. Fergus, «Visualizing and understanding convolutional networks,» chez *European conference on computer vision*, 2014.

[239] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox et J. Clune, «Synthesizing the preferred inputs for neurons in neural networks via deep generator networks,» chez *Advances in Neural Information Processing Systems*, 2016.

[240] V. Sze, Y.-H. Chen, T.-J. Yang et J. S. Emer, «Efficient processing of deep neural networks: A tutorial and survey,» *Proceedings of the IEEE,* vol. 105, pp. 2295-2329, 2017.

[241] J. Yosinski, J. Clune, Y. Bengio et H. Lipson, «How transferable are features in deep neural networks?,» chez *Advances in neural information processing systems*, 2014.

[242] M. Long, Y. Cao, J. Wang et M. I. Jordan, «Learning transferable features with deep adaptation networks,» *arXiv preprint arXiv:1502.02791,* 2015.

[243] X. Jia, B. De Brabandere, T. Tuytelaars et L. V. Gool, «Dynamic filter networks,» chez *Advances in Neural Information Processing Systems*, 2016.

[244] M. Jaderberg, K. Simonyan, A. Zisserman et others, «Spatial transformer networks,» chez *Advances in neural information processing systems*, 2015.

[245] J. Johnson, A. Karpathy et L. Fei-Fei, «Densecap: Fully convolutional localization networks for dense captioning,» chez *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[246] A. Mhalla, T. Chateau, S. Gazzah et N. E. Ben Amara, «Scene-Specific Pedestrian Detector Using Monte Carlo Framework and Faster R-CNN Deep Model: PhD Forum,» chez *Proceedings of the 10th International Conference on Distributed Smart Camera*, New York, NY, USA, 2016.

[247] D. O. Pop, A. Rogozan, F. Nashashibi et A. Bensrhair, «Pedestrian Recognition through Different Cross-Modality Deep Learning Methods,» chez *IEEE International Conference on Vehicular Electronics and Safety*, Vienna, 2017.

[248] H. Knight, R. Toscano, W. D. Stiehl, A. Chang, Y. Wang et C. Breazeal, «Real-time social touch gesture recognition for sensate robots,» chez *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, 2009.

[249] M. D. Cooney, S. Nishio et H. Ishiguro, «Recognizing affection for a touch-based interaction with a humanoid robot,» chez *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, 2012.

[250] A. Billard, A. Bonfiglio, G. Cannata, P. Cosseddu, T. Dahl, K. Dautenhahn, F. Mastrogiovanni, G. Metta, L. Natale, B. Robins et others, «The roboskin project: Challenges and results,» chez *Romansy 19--Robot Design, Dynamics and Control*, Springer, 2013, pp. 351-358.

[251] X. Lamy, F. Colledani, F. Geffard, Y. Measson et G. Morel, «Robotic skin structure and performances for industrial robot comanipulation,» chez *Advanced Intelligent Mechatronics, 2009. AIM 2009. IEEE/ASME International Conference on*, 2009.

[252] A. Cirillo, F. Ficuciello, C. Natale, S. Pirozzi et L. Villani, «A conformable force/tactile skin for physical human--robot interaction,» *IEEE Robotics and Automation Letters,* vol. 1, pp. 41-48, 2016.

[253] M. D. Cooney, S. Nishio et H. Ishiguro, «Importance of touch for conveying affection in a multimodal interaction with a small humanoid robot,» *International Journal of Humanoid Robotics,* vol. 12, p. 1550002, 2015.

[254] B. D. Argall et A. G. Billard, «A survey of tactile human--robot interactions,» *Robotics and autonomous systems,* vol. 58, pp. 1159-1176, 2010.

[255] M. Teyssier, G. Bailly, É. Lecolinet et C. Pelachaud, «Revue et Perspectives du Toucher Social en IHM,» chez *29ème conférence francophone sur l'Interaction Homme-Machine*, 2017.

[256] M. Chetouani, E. Delaherche, G. Dumas et D. Cohen, «15 Interpersonal Synchrony: From Social Perception to Social Interaction,» *Social Signal Processing,* p. 202, 2017.

[257] R. Barraquand, «Designing Sociable Technologies,» 2012.

[258] C. Breazeal, «Toward sociable robots,» *Robotics and autonomous systems,* vol. 42, pp. 167-175, 2003.

[259] Y. Sasa et V. Aubergé, «SASI: perspectives for a socio-affectively intelligent HRI dialog system,» chez *1st Workshop on "Behavior, Emotion and Representation: Building Blocks of Interaction" at Human Agent Interaction Conferences (HAI2017)*, Bielefeld (Germany), 2017.

[260] C. Nass et Y. Moon, «Machines and mindlessness: Social responses to computers,» *Journal of social issues,* vol. 56, pp. 81-103, 2000.

[261] B. R. Duffy, «Anthropomorphism and the social robot,» *Robotics and autonomous systems,* vol. 42, pp. 177-190, 2003.

[262] J. Rios-Martinez, A. Spalanzani et C. Laugier, «From Proxemics Theory to Socially-Aware Navigation: A Survey,» *International Journal of Social Robotics,* 4 2015.

[263] P. Schefler, M. Dickerson et M. Charre, «Les robots sociaux doivent-ils etre humanoides?,» *IC2A,* p. 52, 2014.

[264] A. Specian, N. Eckenstein, M. Yim, R. Mead, B. McDorman, S. Kim et M. Mataric, «Preliminary system and hardware design for Quori, a low-cost, modular, socially interactive robot,» chez *AHRI '18 Workshop on "Social Robots in the Wild"*, Chicago (USA), 2018.

[265] K. He, X. Zhang, S. Ren et J. Sun, «Deep residual learning for image recognition,» chez *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[266] «Charte européenne des droits de l'aidant,» [En ligne]. Available: http://www.aidants.fr/sites/default/files/public/Pages/chartecofacehandicapfr.pdf. [Accès le 15 02 2018].

[267] B. Hamilton-Baillie, «Towards shared space,» *Urban Design International,* vol. 13, pp. 130-138, 2008.

Table des figures

Sélection d'articles

Articles en lien avec le chapitre II

Dominique Vaufreydaz, Rémi Emonet, Patrick Reignier, « *A Lightweight Speech Detection System for Perceptive Environments* », 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, May 2006, Washington, United States.

Rémi Emonet, Dominique Vaufreydaz, Patrick Reignier, Julien Letessier, « *O3MiSCID, a Middleware for Pervasive Environments* », 1st IEEE International Workshop on Services Integration in Pervasive Environments, Jun 2006, Lyon, France.

Viet Cuong Ta, Dominique Vaufreydaz, Trung-Kien Dao, Eric Castelli, « Smartphone-based User Location Tracking in Indoor Environment », International Conference on Indoor Positioning and Indoor Navigation (IPIN), Oct 2016, Madrid, Spain.

Articles en lien avec le chapitre III

Dominique Vaufreydaz, Wafa Johal, Claudine Combe, « *Starting engagement detection towards a companion robot using multimodal features* », Robotics and Autonomous Systems, Elsevier, 2015, Robotics and Autonomous Systems, pp.25.

Fabien Badeig, Pierre Wargnier, Maribel Pino, Philippe De Oliveira Lopes, Emeric Grange et al, « *Impact of Head Motion on the Assistive Robot Expressiveness - Evaluation with Elderly Persons* », 1st International Workshop on Affective Computing for Social Robotics Workshop at the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Aug 2016, New York, United States.

Pavan Vasishta, Dominique Vaufreydaz, Anne Spalanzani. « *Natural Vision Based Method for Predicting Pedestrian Behaviour in Urban Environments* », IEEE 20th International Conference on Intelligent Transportation Systems, Oct 2017, Yokohama, Japan.

Articles en lien avec le chapitre IV

Dominique Vaufreydaz, Amaury Nègre, « *MobileRGBD, An Open Benchmark Corpus for mobile RGB-D Related Algorithms* », 13th International Conference on Control, Automation, Robotics and Vision, Dec 2014, Singapour, Singapore.

Maxime Portaz, Maxime Garcia, Adela Barbulescu, Antoine Begault, Laurence Boissieux et al, « *Figurines, a multimodal framework for tangible storytelling* », WOCCI 2017 - 6th Workshop on Child Computer Interaction at ICMI 2017 - 19th ACM International Conference on Multi-modal Interaction, Nov 2017, Glasgow, United Kingdom.

Thomas Guntz, Raffaella Balzarini, Dominique Vaufreydaz, James L. Crowley, « *Multimodal Observation and Interpretation of Subjects Engaged in Problem Solving* », 1st Workshop on "Behavior, Emotion and Representation: Building Blocks of Interaction", Oct 2017, Bielefeld, Germany.

A Lightweight Speech Detection System for Perceptive Environments

Dominique Vaufreydaz, Rémi Emonet, Patrick Reignier

PRIMA - INRIA Rhône-Alpes, Zirst, 655 avenue de d'Europe,
Montbonnot, 38334 Saint Ismier cedex, France
{Dominique.Vaufreydaz,
Remi.Emonet,Patrick.Reignier}@inrialpes.fr
http://www-prima.inrialpes.fr/

Abstract. In this paper, we address the problem of speech activity detection in multimodal perceptive environments. Such space may contain many different microphones (lapel, distant or table top). Thus, we need a generic speech activity detector in order to cope with different speech conditions (from close-talking to noisy distant speech). Moreover, as the number of microphones in the room can be high, we also need a very light system. The speech activity detector presented in this article works efficiently on dozens of microphones in parallel. We will see that even if its absolute score of the evaluation is not perfect (30% and 40% of error rate respectively on the two tasks), its accuracy is good enough in the context we are using it.

1 Introduction

The base principle of research in ubiquitous computing is to make the computer disappear from the human computer interfaces. Classical input and output devices (keyboard, mouse, screen, etc.) are replaced by other, less intrusive modalities such as voice recognition or computer vision. In order to conduct research in this domain, many research laboratories have equipped dedicated rooms with multiple sensors (cameras, microphones, etc.) and perceptual software (2D and 3D visual tracking systems, speech recognition systems, etc.). The goal is to enable the computer system to understand what the user is saying or doing. The computer system is then able to behave in accordance with the user's intentions. Such highly equipped spaces are often called perceptive environments.

In these environments, all audio sensors, from simplest ones to most complicated ones, have an important role to play. Speech is indeed one of the preferred and most natural communication channels in human to human interactions, and sounds are revealing of human activity. This is why many perceptual environments, such as in the CHIL project [1], are equipped with speech detection, speech recognition and acoustic localization systems. One requirement in such perceptive environments is to be able to process multiple and various microphones in parallel while fitting real time constraints.

Within the CHIL project [1], we are developing a speech detection system that fulfills the requirements of these perceptual environments. Although much research has already been conducted on this point and different approaches have been proposed (such as [2] and [3]), the problem is still open. In addition, we impose two constraints to what is done in most of other systems. Our system must be autonomous. It must be started and then run without human action. It must require neither training nor tuning each time the operating conditions change. We also want to keep our system as light as possible.

In section 2, we first give a description of our speech detection system. Section 3 then presents evaluations that were conducted and the results obtained in the NIST 06s evaluation[1]. Finally we will give a conclusion and some further works to be carried out in order to improve our system.

2 System description

In our perceptive environment, the full SAD (Speech Activity Detection) system can run at the same time over one or many microphones. In this last case, there are two kinds of answer. First, a SAD decision is made at least for each microphone. We can also define a set of microphones in order to get an “ambient" SAD decision using for example multiple microphone arrays. A majority vote is done among all microphones to determine the current state. If speech and non speech votes are equal, the state of the global answer remains the same. This strategy is not optimal when using a large set of different microphones. The design of the system was made in order to run several systems in parallel over multiple groups of microphones.

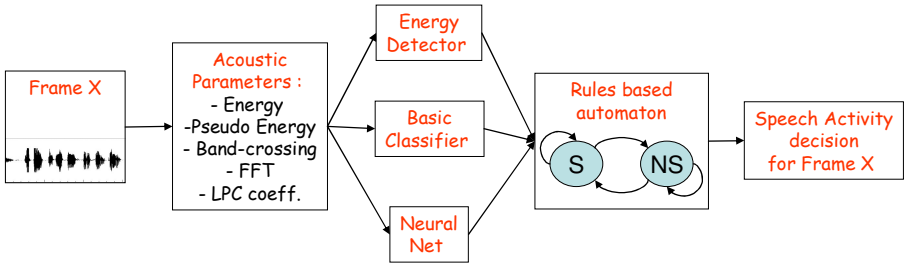


Fig. 1. Design of the SAD System.

On each input, the current version of our SAD system works using several sub-systems: an energy detector, a basic classifier and a neural net trained to recognize voiced segments like vowels for example. At each timestep, i.e. for each frame, each sub-system gives a speech or non-speech answer. Then a hand-made rule-based automaton determines the final result: whether or not there is speech activity. This tools is designed to be enhanced with complementary other subsystems.

[1] See http://www.nist.gov/speech/tests/rt/rt2006/spring/.

2.1. Energy detector

The energy detector uses pseudo energy (we do not sum square values of samples but only absolute values) to determine variation of the input signal energy. It works using two couples of time delay/energy threshold: *TimeOn/EnergyOn* and *TimeOff/EnergyOff*. Simply speaking, if the pseudo energy goes over *EnergyOn* during *TimeOn*, the energy detector emits a *START_SPEAKING* event as a result. In the same way, if the pseudo energy falls under *EnergyOff* during *TimeOff*, the result is *STOP_SPEAKING*. For stable periods, the return values are respectively *STILL_SPEAK* and *NOT_SPEAK*.

As the system is designed to run permanently, it is obvious that these energy thresholds cannot always reflect the current voice/background noise energies. We added to the energy detector the ability to adapt dynamically these thresholds. A sliding window of 50 seconds of previous values for speech portion energy is maintained. In order to smooth threshold changes, the window is filled with the *EnergyOn* value at the initialization time. Then, when the global SAD state changes, for example when final answer of the SAD system goes from speech to non-speech, the training system computes the new threshold value. The system does exactly the same computation on a separate sliding window for *EnergyOff*.

In order to prevent usage of outliers in the online adaptation process, we do not use data from the beginning and the ending of speech or non-speech segments. We privilege inside segments which should be more stable. Thus, we eliminate *TimeOn* data from the beginning and *TimeOff* at the end of speech segments for our online adaptation. Identically, we do not use *TimeOff* and *TimeOn* data respectively from beginning and ending of non-speech segments. We also do not use data from segments that are not long enough. So, if a segment does not contain at least 1 second of interesting data, it is not use to compute new thresholds.

The final computation of new values for *TimeOn* and *TimeOff* do not use a simple average approach but a “median” one. We sort the adaptation data contained in the sliding windows and remove possible outliers, i.e. too low and too high values. At least, we keep only 60% of the data in order to re-estimate our thresholds. Our real adaptation time is thus 30 seconds (60% of 50s) for each threshold.

2.2. Basic Classifier

This classifier is dedicated to recognize and to tag specific sound classes: fricatives, low frequency sounds like computer or air conditioning fans, and other sounds. The first step is a hamming window and a Fast Fourier Transform (FFT) [4] to obtain the spectrum of the signal. The classifier deals with 5 identical sub-frequency bands from 1 to 8000 hertz where it computes energy. This classifier works only on signal recorded at 16000 hertz or higher sampling rates. In this last case, frequencies over 8000 hertz are not used.

With the 5 energy values, the module can classify the audio signal:

- if more than 90% of the total energy is concentrated in the 2 lowest bands, the sound is a *low frequency sound*
- if the energy in the 2 lowest bands is less than in all other bands, the sound is a *fricative*
- in all other cases, the sound remains *unclassified*.

2.3. Neural Network

The neural net is a multi-layer perceptron with 2 hidden layers. It uses as input coefficients computed on the input frames:

- Zero crossing: number of time the signal goes from a negative to a positive value and vice versa. Actually, we use a variant called band-crossing [5] that does not count oscillations in a band around 0.
- Energy: the sum of the square values of samples.
- 16 predictor coefficients: they are extracted from a speech analysis method called Linear Predictive Coding (LPC) [6]. We use the auto-correlation method combined with the Durbin recursion to compute them.

This module is the only sub-system that needs to be trained. The training was made once and for all on 1 hour of French speech extracted from the BREF corpus [7]. The phonetics labels used during the training phase are not the original BREF ones but were computed with RAPHAEL [8], a French recognizer. The training data were almost equilibrated, ~50% of female voice and ~50% of male voice. Result of this module can be *speech* or *non speech*.

2.4. Rules based automaton

This automaton is designed to integrate results from all subsystems to produce a final answer. It consists in 2 states (*speech* and *non speech*) with hand-made rules to change from one state to the other. The rules were defined using knowledge about each subsystem. In all cases, if all subsystems agree on the current state, i.e. when each result is a speech one[2], the system uses it. In the *speech* state, if the energy detector return value and at least one another result are *non speech*, we go into the *non speech* state. We do the same when the basic classifier returns *unclassified* and the neural net *non speech* or when the basic classifier gives *low frequency sound* as answer. Symmetrically, we defined the same type of rules for the *non speech* state.

[2] Speech events are *START_SPEAKING* or *STILL_SPEAK* for the energy detector, *fricative* for the basic classifier and *speech* for the neural net.

3 Evaluation

3.1. Implementation

The implementation of the full system is made in C++ and can run on multiple operating systems. As explained above, the system can handle signal from 16 KHz to 44.1 KHz. During evaluation, the SAD system works on frames of 256 samples on a 16 KHz signal. Thus, the time precision of our system is 16 ms. Another important point: the system remains the same for all evaluation tasks. We do not have specific configuration or training for close talking or far field microphones.

3.2. RT06S evaluation Data

The RT-06S evaluation is focused on the Meeting Domain interaction with two sub-domains (or tasks). The first one consists of ten meetings recorded in a conference room in six different sites: it is called "*confmtg*". The second one, aka "*lectmtg*", is composed of several lectures with lecturer and question/answer speech. Each sub-domain has different sensor setups, different levels of interactions and multiple structures of test excerpts. The reader may refer to [9] to find more detailed information about the evaluation data.

For each task, one may run its speech activity detection system in many different conditions. According to our system characteristics, we decided to evaluate our system on the following subset of conditions:

- Individual head microphone (*ihm*)
- Multiple Distant Microphones (*mdm*)
- Single Distant Microphone (*sdm*)
- All Distant Microphones (*adm*)
- Multiple Source Localization microphone Arrays (*msla*)

The final evaluation contains about 180 minutes of speech for the *confmtg* task and 145 minutes for *lectmtg*.

3.3. Evaluation metrics

In this paper, we use two different metrics in order to test if our system fulfils our needs: a light and accurate system. To check if our SAD system is light enough, we compute the real-time factor as expressed in equation (1).

$$\text{Real - time factor} = \frac{\text{Total processing time}}{\text{Data time}} \quad (\mathbf{1})$$

This factor permits to know easily if the system can be real-time. If the real-time factor is less or equal to 1, so if the processing time is less or equal to the data time,

the system can be considered as real-time. In the next section, we will check how many different inputs we can handle at the same time.

The other evaluation point concerns accuracy. We need to know how efficient our SAD system is. Within the RT06S evaluation, the SAD metric is time based [9]. The scoring system first computes the full speech time over the considered audio signal. Then using output from systems and reference files, manually annotated, we obtain the missed speech and the false alarms times. Doing that on all files from a condition for a given task and accumulating values, we can calculate the overall error rate on the given task using the equation (2):

$$\text{SAD Error} = \frac{\text{Missed speech} + \text{False Alarm time}}{\text{Speech time}} \quad (\mathbf{2})$$

In order to enrich this result, we built tools to compute some extra information. For each task and for each condition, we first extract for each talk, the SAD error rate. Then we decompose the Missed speech time in four time-weighted categories:

- *Full miss*: a complete speech event which is not detected;
- *Miss begin*: the beginning of a speech event is not detected;
- *Miss In*: middle part(s) of a speech event not tagged as speech;
- *Miss end*: the end of a speech event is not found.

Using this new information, we will be able to analyze more precisely the error committed by our SAD system.

3.4. Results

In this section, we detail the results of our system in the RT06s evaluation. First, we will introduce speed measurement and then, the accuracy of our SAD system.

3.4.1 Speed evaluation

As we have already said, speed is a strong constraint for us. We work in interactive spaces and we need low latency application. It is not suitable for us to transmit a lot of data over the network. Thus, we need to be able to process a maximum of microphones on a single computer.

If we measure speed factor on a single 16 KHz audio signal, the computational time for 1 second of speech is 0.0076 second: in theory, the system can handle more than 130 channels at the same time. In practice, operating system scheduling, memory management and multi-channels SAD fusion can alter this result.

We built a test set using 300 seconds segments (containing voice) extracted from the RT06s evaluation database. The full SAD system ran over this set using firstly 1 segment, then 2, etc. Each time, segments were chosen randomly. We stop the test when the real-time constraint was violated, i.e. when the real-time factor goes over 1. The next figure shows the experimental results.

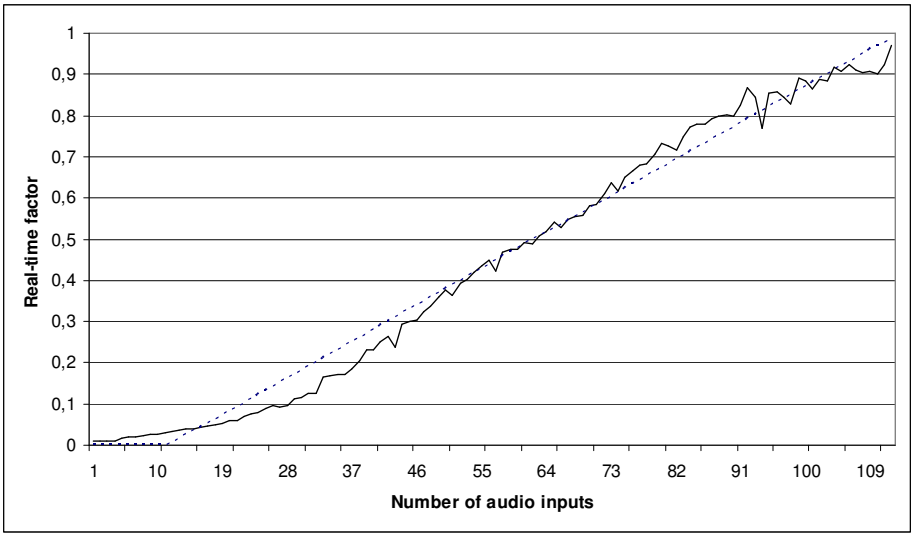


Fig. 2. Speed performance of our SAD system running on a single core unit of a processor and processing multiple inputs.

On this figure, one can found a dashed line obtained by linear regression on data. We also see a curve showing the real-time factor of our system regarding inputs.

We can first see that the real-time factor curve is not always increasing. This surprising phenomenon can be explained by the load of the computer and by the computation time that changes from one file to another (see 2.1). The other result which affords is that our system is linear-like over 50 inputs. It is very important because if we want to add some microphones, we can do it without rethinking the whole computer configuration.

Finally, our SAD system can process 112 streams which is a good score. In real conditions, i.e. when we do not process files, we must also consider that acquisition process will alter this result and certainly slow the system. Nevertheless, we can objectively say that we can process as many microphones as a sound card can record in real time.

3.4.2 Speech Activity Detection accuracy

This section presents experimental results from the RT06s evaluation. In these experiments, starting energy thresholds of the energy detector were values empirically defined during previous research projects (NESPOLE! [10] and FAME [11]). The evaluation metrics of our system are given in the three following tables.

Table 1. Global results over the different tasks.

Task	Condition	Overall Error Rate	Best Error Rate	Worst Error Rate
confmtg	ihm	78,54%	20,31%	1917,99%
	mdm	46,98%	13,12%	80,20%
	sdm	41,26%	18,48%	76,98%
lectmtg	adm	27,81%	3,73%	95,51%
	mdm	32,59%	4,29%	95,47%
	msla	30,61%	3,86%	100,00%
	sdm	33,87%	5,74%	85,29%

The first table above gives us general information about accuracy of our SAD system. The official results of the RT06s evaluation are given in the "Overall Error Rate" column. We can first see that our system is not accurate on the *confmtg-ihm* condition. For this condition, only speech coming from the main speaker must be tagged as speech. As our system is not design to do that (every speech segment can be tagged as speech, even if its energy is low), this result is not really significant: our worst score within this condition is 1918% of error. As we did not understand correctly this task before evaluating, we let the result in the previous table but we will not analyze more precisely this condition. Concerning others conditions of the *confmtg* task, we can see that our average error rate is ~43%. Our best scores, over one seminar, are not good (13% and 18%). We will check in the next section where our system fails. If we look to the *lectmtg* task, we can see that our global results are better: ~30% of error in average over all condition. Moreover, our best scores are good (from 3% to 6%) but our worst score stay high (up to 100%).

We will now trying to understand more precisely where the errors are. The following tables show our additional metrics computed first on the confmtg task, next on *lectmtg*.

Table 2. Detailed results of the *confmtg* task.

Condition	Full Miss	Miss Begin	Miss End	Miss In	False Alarm	Error Rate
mdm	13,22%	3,31%	5,38%	24,97%	0,11%	46,98%
	(28,13%)	(7,04%)	(11,45%)	(53,15%)	(0,23%)	(100,00%)
sdm	1,63%	2,17%	4,39%	26,86%	6,22%	41,26%
	(3,95%)	(5,25%)	(10,63%)	(65,11%)	(15,07%)	(100,00%)

In this table, one can found the official evaluation error rate in the last column. The second sub-row of each entry gives the percentage of each type of error on the overall error rate.

On the multiple distant microphone (*mdm*) condition, our system is not good at all. 13% of the speech segments were entirely missed. 25% of the missed speech is within a speech turn. A good result is the false alarm error rate which is very low. This value is correlated to the 13% of full miss. Our SAD system did not make mistake on non speech segments but was insensible to some speech parts. Our starting thresholds were too high and not adapted to this condition. Moreover the system do not managed to adapt them online. Concerning the single distant microphone (sdm) condition, results are slightly better. In fact, we only have a *full miss* rate of ~2%. We see a rise of the *miss end* and *false alarm* rates. *Miss in* factor stay over 60% of our errors.

If we look globally at the previous table, we can say that ~60% of our errors are due to intra speech undetected portions. Even if we could artificially solve this problem by changing the *TimeOff* delay of our system, we do not want to do it. Field experiments have already shown at our laboratory that adapting system to an evaluation can lead to decrease drastically real world performances. Other metrics can not be averaged because there are too distant between the two conditions.

The table below introduces more accurate results achieved on the *lectmtg* task.

Table 3. Detailed results of the *lectmtg* task.

Condition	Full Miss	Miss Begin	Miss End	Miss In	False Alarm	Error Rate
adm	3,65%	3,06%	4,25%	12,20%	4,66%	27,81%
	(13,12%)	(10,99%)	(15,28%)	(43,87%)	(16,75%)	(100,00%)
mdm	5,83%	3,52%	5,00%	11,59%	6,65%	32,59%
	(17,89%)	(10,81%)	(15,34%)	(35,56%)	(20,40%)	(100,00%)
msla	4,52%	3,65%	4,69%	12,51%	5,24%	30,61%
	(14,75%)	(11,92%)	(15,33%)	(40,88%)	(17,12%)	(100,00%)
sdm	3,73%	4,16%	6,11%	13,93%	5,94%	33,87%
	(11,00%)	(12,29%)	(18,05%)	(41,13%)	(17,53%)	(100,00%)

The results of our SAD system over the *lectmtg* task are better than on the *confmtg*. In absolute, the overall error rate is 10% lower (~30% overall error rate). We can also remark that the percentages for all conditions are similar: in average 4.5% of *full miss*, 3.6% of *miss begin*, 5% of *miss end*, 12.5% of *miss in* and 5.6% of *false alarm*[3]. We can quickly see that the *miss in* errors represents a huge amount of the error rate (40%). *False alarm* and *miss end* follow with almost 20% of the errors. We can analyze these results saying that the data of the *lectmtg* evaluation are closer to the capabilities of our system than *confmtg*. We still see a huge amount of boundaries problems, lost intra speech segments remain our major problem. Finally, the good outcome is that our system can detect ~95% of the speech turns even if the boundaries are not precisely located.

At end, we can draw some conclusions on this evaluation. Our SAD system is not perfect if we look only at the final percentage. Most of the time, problems are boundaries (*miss begin*, *miss in*, *miss end*) and a non negligible part is *false alarm*. After looking at some labels and reference files, we can say that these problems seem to be less present at the end of the seminars. The adaptation process seems to refine the thresholds but with a too long latency. As our system is designed to run permanently, it is usually not a problem. During evaluation, the SAD system starts from scratch for each file. As we already said, we need transitions in order to compute new values. If we do not have enough transitions, it is obvious that we will not have a suitable adaptation process. For the next evaluation, we should consider to use our system in real condition but for this experiment, we decided that grouping seminar by collecting site, thus by similar recording equipments, settings and conditions, can be considered as cheating. We have chosen not to do so.

[3] As a comparison point, the false alarm rate for a system always answering *speech* is 25.44% on this task.

For us, the major result of our system for the RT06s evaluation is the low percentage of *full miss* speech turns (13% of the *mdm* and 1.6% for the *sdm* condition of the *confmtg* task and <5% in average for the *lectmtg* task). This score validates the usability of our SAD system in the context we are using it: detecting speech turns and people interactions for context modeling.

4 Conclusion and future work

Our speech detection system exposes performances that make it suitable for our projects and goals such as CHIL [11] or [12]. Actually, we do not want to use it for automatic speech recognition or diarization but for interaction and context modeling. Thus, we do not need precise speech boundaries but only to be sure to detect interaction between people so at least a part of each speech turn. If we look to the previous section, we can see that this goal is fulfilled: the *full miss* score is low. We think that this result is good for a generic system not trained on twin data of the evaluation that can run in different working environment (smart office, conference room, amphitheater, etc.) without any preliminary training/adaptation. Concerning our speed constraint, we saw that our system is successful because we can process many inputs in parallel. Following proposed improvements will be integrated carefully in order to preserve this capability.

For future work, experiments described in this paper have shown that some improvements of our system still need to be carried out. First of all, the online adaptation of the energy thresholds has to be refined. It could improve the performance of the system but decrease the system accuracy on adaptation failure. Next, we also envision extending the training of our neural network. We still want to keep an *a priori* training for this neural network while including different kind of speech recorded in different environment and different conditions (lapel and distant microphone). We think that doing this will provide us with a more generic speech detection system. We are currently preparing the corpus required by such extended learning.

We also want to improve our fusion scheme. We can substitute our rule based approach by a Bayesian one. The current expert-defined rules are rigid and could be advantageously replaced by a Bayesian fusion process. Moreover, if we plan to add other speech activity detection sub-systems, it will be difficult to rebuild a new set of rules. Doing this Bayesian training will also require having a learning corpus.

The last source of improvement would be to take advantage of all the available data in our perceptive environment. Our system could use visual information (facial features, visual clues of sound events, etc.) to improve the performance of speech detection.

5 References

1. D. Macho, J. Padrell, A. Abad, C. Nadeu, J. Hernando, J. McDonough, M. Wolfel, U. Klee, M. Omologo, A. Brutti, P. Svaizer, G. Potamianos, S.M. Chu. Automatic Speech Activity Detection, Source Localization, and Speech Recognition on the Chil Seminar Corpus. In IEEE International Conference on Multimedia & Expo, January 2005.
2. J. Ramirez, J. Segura, C. Benitez, A. de la Torre, A. Rubio. Efficient voice activity detection algorithms using long-term speech information. Eurospeech'97, pages, 1997.
3. A. Martin, D. Charlet, L. Mauuary. Robust Speech/Non-Speech Detection Using LDA Applied to MFCC. In Proc. ICASSP, vol. 1, 237-240, Salt Lake City, May 2001.
4. M. Frigo, and S.G. Johnson, The Design and Implementation of FFTW3, special issue on "Program Generation, Optimization, and Platform Adaptation", volume 95, pages 216-231, 2005.
5. J. Taboada, S. Feijoo, R. Balsa, C. Hernandez, Explicit estimation of speech boundaries, IEEE Proc. Sci. Meas. Technol., vol. 141, pp. 153-159, 1994
6. L. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall PTR, ISBN 0-130-15157-2, 1993.
7. L. Lamel, J.L. Gauvain, M. Eskenazi, BREF, a large vocabulary spoken corpus for French. In Proc Eurospeech'91, Genova (Italia), 1991.
8. D. Vaufreydaz, Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue, Ph.D. in Computer Science at Joseph Fourier University, Grenoble (France), 226 pages, January 2002
9. Spring 2006 (RT-06S) Rich Transcription Meeting Recognition Evaluation Plan, http://www.nist.gov/speech/tests/rt/rt2006/spring/docs/rt06s-meeting-eval-plan-V2.pdf.
10. F. Metze, J. Mc Donough, H. Soltau, A. Waibel, A. Lavie, S. Burger, C. Langley, L. Levin, T. Schultz, F. Pianesi, R. Cattoni, G. Lazzari, N. Mana, E. Pianta, L. Besacier, H. Blanchon, D. Vaufreydaz, L. Taddei, The Nespole! Speech-to-Speech Translation System, Human Language Technologies 2002, San Diego - California (USA), 6 pages, mars 2002.
11. F. Metze, P. Gieselmann, H. Holzapfel, T. Kluge, I. Rogina, A. Waibel, M. Wolfel J. Crowley, P. Reignier, D. Vaufreydaz F. Bérard, B. Cohen, J. Coutaz, S. Rouillard V. Arranz, M. Bertran,, H. Rodriguez, The "FAME" Interactive Space, 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, Edinburgh - UK, 4 pages, February 2005.
12. O. Brdiczka, J. Maisonnasse, P. Reignier, Automatic Detection of Interaction Groups. In Proc. Int'l Conf. Multimodal Interfaces, October 2005.

O3MiSCID, a Middleware for Pervasive Environments

Rémi Emonet, Dominique Vaufreydaz, Patrick Reignier, Julien Letessier
PRIMA - INRIA Rhône-Alpes
Zirst, 655 avenue de d'Europe, Montbonnot, 38334 Saint Ismier cedex, France
{remi.emonet,dominique.vaufreydaz,patrick.reignier,julien.letessier}@inrialpes.fr

***Abstract*— This paper introduces a new lightweight middleware for pervasive environments. This middleware abstracts network communications and provides service introspection and discovery using DNS-SD (*DNS-based Service Discovery* [1]). Services can declare simplex or duplex communication channels and variables. The middleware supports the low-latency, high-bandwidth communications required in interactive perceptual applications. It has been designed to be easy to learn in order to stimulate software reuse in research teams and is revealing to have a high adoption rate.**

I. Introduction

The fast emergence of pervasive computing is transforming the way the applications are designed. Applications are now commonly distributed on numerous computers with various hardware capabilities (from high-end servers to PDAs). Building mobile, distributed and robust applications is a challenge for system designers and developers. To cope with this new complexity, some pieces of middleware are being designed. They help the designers and developers by abstracting the lower-level details of application development and providing higher-level concepts to manipulate.

The concept of zero-configuration networking proposes the following approach: all application parts are services that are declared over the network. When a service requires another to function, it doesn't need to known its network location (host and port): a service discovery mechanism can provide it, based on its type and its properties (a discovery query). The same approach is used when browsing the yellow pages to find a plumb: you know what service you need but don't know yet the exact coordinates where to contact its provider.

All service-oriented middleware provide service lookup; however, some of them require the application to know where to find a centralized service repository. There are mainly two existing infrastructures for zero-configuration networking that do not rely on a centralized well-located service repository: UPnP and Zeroconf.

UPnP (*Universal Plug and Play* [2]) exposes a large set of features including the SSDP (*Simple Service Discovery Protocol*) protocol that handles service discovery. Although UPnP is widespread, very attractive and fulfills many of our requirements, it has a drawback that is redhibitory for the use we want to make of it. UPnP does not have a mechanism for instant notification when a service becomes inaccessible, and thus, there is no reliable way to have an up-to-date view of the available services.

The second major infrastructure for zero-configuration networking is Zeroconf [3]. Zeroconf leverages DNS-SD [1] combined with Multicast DNS [4] to achieve distributed service discovery with quick service connection and disconnection notifications.

There are also some dedicated middlewares such as [5]–[8], but none is suitable to our requirements. Although some are designed with context-aware applications in mind, which is one of our goals, most of these middlewares are too specialized or Java-centric, or both. None of these middlewares fulfills all our needs; in particular user-friendliness. We choose to propose an easy to learn middleware based on DNS-SD.

In section II we will first identify the requirements for a user-friendly, platform-agnostic middleware for interactive applications. We then present our solution, O3MiSCID (section III), the Object-Oriented Opensource Middleware for Services Communication, Introspection and Discovery. Two examples using this middleware are detailed in section V. Finally we will conclude and present the future work related to this middleware.

II. Objectives and Requirements

In this section we detail the requirements our middleware should meet: attractiveness, network efficiency, robustness, ease of configuration and extensibility.

1) Attractiveness and availability: The main goal of a middleware is to allow (independently developed) software components to be uniformly packaged as services and cooperate. The middleware should be sufficiently attractive: most reaserchers and developers should feel that the benefit of using it will quickly outweight the development overhead. Moreover, developers have different backgrounds, interests and skills. The middleware must be available to any potential user: it must be cross-language, cross-platform, and easy to learn for anyone.

2) Network efficiency: Our middleware needs to be a general purpose middleware, with a specific constraint: it must be useable in the context of interactive applications involving audio and video processing. To match the interactive constraints, it must keep a low latency in the communications. Perceptual processes such as audio-video analysis require high throughput: the middleware should not overload the communication cost between services.

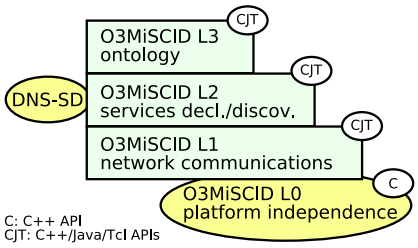


Fig. 1. The layered architecture of the proposed middleware
Each layer is accessible using dedicated APIs currently available in Java, C++ and Tcl.

3) Robustness: We split this requirement in two categories: middleware robustness and service robustness. Regarding the former, we note that if a centralized middleware approach is used, e.g. with a single middleware server offering services to a network of computers, the failure of the server may cause the failure of all client applications. Application reliability requires middleware robustness, which in turn imposes to rely on a decentralized approach.

In pervasive applications, services are often dynamically added or removed. Services should thus be robust to communication loss, and have easy access to up-to-date service availability information. They can then be designed to recover from disconnections by using the list of available services provided by the middleware.

4) Ease of configuration: Distributed software consists in splitting a complex application into multiple processes (services), usually spread on multiple computers. Configuring such software proves challenging, as services may need to find other dependant services to function. In order to easily deploy an application in a new site, we reduce the number of configuration files specifying the interconnections by relying heavily on auto-detection features. It is the task of service designers to ensure strong autonomy in software configuration, however the middleware must make this task as easy as possible.

5) Extensibility: The middleware must be easily extensible: one should be able to design dedicated frameworks featuring higher abstraction level concepts. Such an extension mechanism must fit into the existing middleware. Some of the possible extensions we currently envision are a context awareness extension and an extension to help in designing "autonomic systems" [9].

III. Middleware Architecture

In this section, we present the middleware architecture we propose. It is layered (see fig. 1): layer 1 handles and abstracts communication mechanisms; layer 2 handles services description, discovery and introspection; finally, layer 3 manages ontological manipulations of the services.

A. Layer 1: Network Communications

The lowest layer features a lightweight communication protocol and programmatic facilities to handle communications. We use the BIP communication protocol (*Basic Interconnection Protocol*, [10]), which splits the communication stream into messages. This is achieved by adding a 34-byte header to each message to be sent. This lightweight header contains, among other information, the size of the message payload. This content can contain arbitrary data such as a binary stream, ASCII text, or an XML fragment. Given the small size of the header and the absence of limitation on the message content, this protocol is sufficiently general purpose and can handle high bandwidth messaging and streaming.

Any application using the proposed protocol is called a BIP peer. All services (see next section) are BIP peers but not all BIP peers are services. In particular, any peer can use BIP to access running services without having itself been declared as a service. In other words, layer 1 can be used independently, which allows for easy interoperation.

B. Layer 2: Services

The service layer is built on top of the communication layer and is dedicated to service description, introspection and discovery. Basically, a service is described by its name, its input/output communication channels and other service-specific parameters. This subsection details the service-related concepts manipulated by O3MiSCID.

1) Service Description and Introspection: A service must have a unique identifier: a human-readable name that is unique, network-wide. A service can expose a set of channels. Each channel is a communication "port" that can be mono- or bi-directional channel (*input*, *output*, or *inoutput* channels). In the scope of a particular service, each channel has a unique name, a human-readable description string, and a type string that describes the data format used inside this channel. A service can also export state parameters called variables. Variables can be read and written depending on access rights set to it. Each variable also features a type and a description. A peer can subscribe to variable modifications and receive notifications each time the variable value is changed.

Each service has at least an *inoutput* communication channel called the control channel, dedicated to introspection queries. This allows to list the channels and the variables exported by the service. It is also used to interact with the service: query, update, and subscribe to variable changes.

2) Service Discovery: The middleware allows a client to register new network services and provides two ways to discover available services. A service browser can be notified whenever a service event occurs (service appearance or disappearance). Alternatively, a client can use a wait-for-service functionnality: one provides a service description, and is notified as soon as a matching service is present. Such descriptions are *service filters*: conditions that can be based on any service attribute such as service name or host; channel name, description, format, or supported versions; or service variable.

3) User Point of View: The exact use of the middleware may differ slightly depending on the programming language that is used. However, using this middleware basically consists

in instantiating classes representing a service to register. Then, one can define inputs and outputs (directly in the source code or through an external XML description file, see below), and eventually bind input channels to callback functions. Similar actions are possible for the service variables.

The middleware provide also all the required facilities to inspect and discover remote services. Information about the remote services is accessible programmatically. Filters for the discovery process are manipulated as functors (objects representing a function); in our case, a boolean function taking a service description as a parameter.

C. Layer 3: Ontology

The ontology layer is the highest abstraction layer of O3MiSCID. For the moment, it is only preliminary and the object of current studies. Choices were made in order to check the potentiality of the method. They are not definitive.

There are two requirements driving the ontological layer design. From a system-centric perspective, we need to store information about "the world", i.e. everything related to the perceptive environment, including any hardware and software. This data will be used for monitoring (see section V-A) and service selection purposes for example. The main system issues are information sharing, availability and consistency. From a user's point of view, we aim to provide a means to define a service easily. The end user should only write a terse description of a service to launch it. Another key point is how a service can be found using not only its name as a query. Users will be reluctant to write a programmatic service description to discover a suitable service.

Concerning the user standpoint, we need to provide a means to describe the world. We experimented with two languages: the OWL ontology language and a custom XML language. OWL is currently used in many pervasive ontologies [11] [12]. Even if it is an expressive and powerful language, it is not well known and it requires us to add a reasoner to our middleware. Furthermore, developers are generally proficient in XML-based technologies, and O3MiSCID already uses XML (for the service inspection interface). Our current choice is thus XML.

Using XML, the user can now only write a concise description of the service like in the following example code:

```
<Microphone freqMin='50' freqMax='10000'
    type='long' id='MicroLong3'
    sensibility='5'>
  <position3D X='1.78' Y='5.75' Z='1.38' />
  <orientation horizontal='180' vertical='0'
    axial='0' />
  <genealogy>
    <branch>Hardware::Microphones</branch>
  </genealogy>
</Microphone>
```

This XML code is used to define a microphone service. The end user calls a class constructor with it or a file containing the code to publish the service. Channels and variables are already set with values found in the XML description. By generating multiple descriptions, one can declare multiple services simultaneously and obtain an array of ready-to-use services.

Another application of using such descriptions is the ability of a service to register to several services in order to get a partial vision of the world. For example, a service that needs to use cameras can register for all cameras (current and future) in order to dynamically receive their XML description. It can then build an XML tree using this information and query it using XPath to find the best suitable camera. Creating a partial vision of the world and querying over the resulting XML tree represents only 2 method calls.

IV. O3MiSCID Implementation

In order to cope with our requirements, we need to make several technical choices for the implementation. In this section, we detail and justify these choices, while presenting an overview of O3MiSCID's programmer interface.

A. Conformance to layer requirements

1) Layer 1: The implementation of the network communication layer provide programmatic facilities for communication. Classes are provided to create client, server or bidirectional client-server socket objects that communicate using the BIP protocol. Once instantiated, client objects can be used to send messages; a callback function is called on the server for each message received. Other facilities are provided such as message interpretation (as byte array for binary messages, as string for plain text, as DOM tree for XML documents...) and listing of the clients connected to a server.

To ensure the reliability of the communications, TCP-only connections are used by default by the middleware. However, for some applications, latency is a more important constraint. To cope with this, the middleware allows also mixed TCP/UDP connections to be used. Indeed, tightly coupled interactive applications [13] implemented using this middleware meet the typical human-computer interaction latency constraint (50 ms).

This lightweight lower layer is responsible for meeting the performance requirements.

2) Layer 2: O3MiSCID handles service discovery and registration using a DNS-SD infrastructure. The mainstream DNS-SD implementation (branded *Bonjour*) uses multicast DNS (mDNS) to advertise all the existing services in a non-centralized fashion: there is no central service repository. This allow to meet the robustness requirement. Available mDNS toolkits provide a daemon running on each machine. When an application wants to register a service or browse the available services, it addresses its queries to this daemon, via a dedicated API. An additional advantage provided by DNS-SD is that any system is instantly notified of service failure or disconnections. This property is mandatory for reactive networks to be able to maintain an up-to-date list of available services.

Given the presented concepts of channels, control channel, variables, and introspection, O3MiSCID uses DNS-SD to simplify the implementation of some of the concepts. The

DNS SRV record is used to store the name of the service, the machine it is hosted on and the port of the control channel of the service. The TXT record having a limited size, it is used as as a cache for static service information that is also available through control queries. The main information cached in the TXT record are the list of input, output and inoutput channels, the TCP and/or UDP port number associated to the channels and an information string designing the owner of the service. This permits for performant query-based discovery of services, as it allows to shortcut service inspection for simple queries, and to accelerate more complex queries.

3) Layer 3: In order to cope with the system requirement of the third layer, we need to choose an architecture. Most of the current systems (e.g. [11]) use a centralized approach. On one hand, it allows easy sharing of ontological data with low protocol overhead. On the other, in the event of a server failure, all information about the environment is rendered unavailable. Inconsistencies may also arise because of the network latency between changes in a service's perception of the world and the availability of this information on the server. In other words, asking information about a service at time *t+1* may provide outdated information because the state of that service at time *t* is only updated at time *t+2* on the server. The problem is the same for services that crash, or simply stop.

For all these reasons, using a distributed architecture seems suitable. The first idea that comes to mind is to use several ontology servers. It solves the "server crash" issue but not the data consistency problem: replication and propagation of information is a complex problem we do not want to address. Thus, we decided to delegate information sharing to the services themselves, in a peer-to-peer fashion. If one needs information about a service, it is simply queried. If up-to-date information is required, one may also register to be notified, and receive messages if the service state is modified. The consistency problem then disappears, as any service only sends its current state. Moreover, if a service crashes, DNS-SD notifies all interested peers, and other services will stop querying.

B. Multiplateform cross-language implementation

One of the design goals of this middleware is to be simple enough to be easily implemented. This goal has been achieved, as a few implementations are already available in various languages ranging from an interpreted scripting language (Tcl), to a high performance language (C++), and to a high productivity language (Java).

On top of the basic full featured Java version of the middleware, an OSGI version has also been built. The OSGI packages in the Java version expose a higher abstraction level interface that hides the technical details of the middleware.

The C++ version of the middleware is fully crossplatform and works on Windows, Linux and MacOSX. To achieve this ability to run on any platform, an abstract layer has been added. This layer abstracts the system-level base objects such as threads, mutexes and sockets (see fig. 1).

Depending on the needs, further implementations may be written. By now some interesting languages and implementation we can think of are Python and C# (to cleanly integrate with the .NET framework and all its languages).

V. Example Applications

In this section we give a simplified illustration of how the middleware can be used. The examples we present illustrate the usage of different kinds of services at different levels of abstraction: hardware services (microphones, cameras), perceptual services (speech activity detector, visual trackers) and a few higher-level services like a context modeler or a dialog manager.

Our field tests take place in a smartroom environment. This room is equipped with multiple cameras, microphones and is used to run many O3MiSCID services. We will first show that our middleware can bring us facilities for monitoring hardware, software and even human activities in this room. We will next present a multimodal perceptive application designed to automatically record seminars and conferences for instance.

A. Perceptive environment and services monitoring

When working in a perceptive environment, monitoring issues are critical. There are many things we need to observe. We may need to keep an eye on the perceptive environment itself, on the currently running services and their status, or results from this services, for instance the location objects detected by a tracking service. In our approach, all this information (what we name "the world"), is described by a set of services. Indeed, as seen in section III-C, each service provides information about itself by answering to queries over its control channel.

In the following example, the world is our smartroom with sets of microphones, cameras, beamers, trackers, a speech activity detector, a visual tracker, a movie director, etc. In this case, we can extract much information for distributed data over the O3MiSCID middleware by browsing existing services and instrospecting them (see Section III-B). First, we can dynamically extract information about the room and automatically construct a 3D representation. On figure 2, we can see the actual position and orientation of cameras (around the room ceiling), the array of microphones (on the wall on the top of the schema) and a steerable camera-projector pair [13]. We can also display on-line changes like the movements of mobile pan/tilt cameras, by registering to the orientation variable of the corresponding camera services; or display targets found a visual person tracker, appearing and moving inside the room.

Using the introspection capabilities of O3MiSCID, we can also obtain the interconnection graph of all running services in our environment (Figure 3). In this example, the connectors do not presume of the direction of data between services. We can see different types of connections between services:

- the first type concerns standalone services (BlueTooth-Scanner, AudioRouter...). They do not have any client connected to them at the graph construction time;

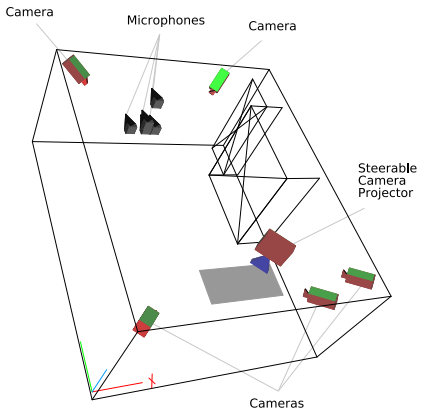


Fig. 2. Gathering and representing data from O3MiSCID

- next, we find a group of interconnected services. All of these connections have been determined fully automatically. This group of services constitutes the Automatic Cameraman and it will be detailed in the next section. It is a typical example of a perceptive application composed of many services, as targeted by our middleware, thus a good case study;
- the last type that we can see is the connection between the `66cd4567` client and the `out3DTracker` service. Indeed, as already mentioned in section III-A, one may connect directly to a service using the BIP protocol [10] using only the first O3MiSCID layer (or even without using O3MiSCID).

To conclude this section, our middleware gives multiple ways to observe the pervasive environment from the system, the application, or the user perspective. The introspection and distributed descriptions mechanisms offer many monitoring facilities. It is easy to envision number of applications still allowing to enrich these capabilities.

B. Automatic cameraman

The automatic cameraman is a multi-services application able to determine what happens in a room equipped with microphones and cameras in order to record the most representative movie. It has been successfully field tested to record a full conference [14].

Our virtual test case takes place in a room containing two cameras and one microphone plugged on two different computers. Services associated to the hardware are started on two machines to handle the data acquisition process. Camera services have a calibration variable (a 4x4 matrix encoding the position, orientation and internal parameters of the camera) and an output channel where each frame is written. The microphone service is quite the same: it has a position variable and an output channel.

Figure 4 shows perceptual services and two higher level services. A visual 2D tracker service is responsible for tracking people into the room using a camera and a homography (this tracker and more generally internal processing of services are not in the scope of this article.). A speech activity detector uses

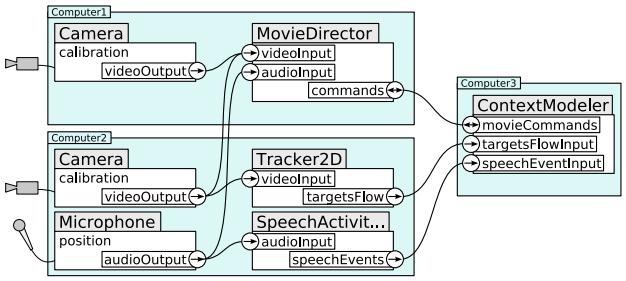


Fig. 4. Example of the "Automatic Cameraman" distributed on 3 different computers. Input, output and inoutput channels of the services are represented by the circles. Curves are representing the connection between the channels. Note that the control channels of the services are not shown in this figure.

a microphone and determines if there is someone speaking in the room. The context modeler takes outputs from the two previous services and performs reasoning to understand what is going on in the room (who is here, where are they, and what are they doing). It outputs commands (change camera view or move mobile camera for example) to the Movie Director service, which is responsible for the production and/or streaming of the final movie.

Actually, the Automatic Cameraman tends to construct this kind of movies using several microphones and cameras. To end up with the connection graph pictured in figure 4, the Movie Director service, for example, first queries a list of all potential services and performs introspection on them using Layer 2 (see III-B). It lists all the cameras and retrieves their position (extracted from their calibration matrix). It does the same with all the microphones. Then it selects the microphone service and the camera service that are closest to each other and then connects to them. We can note that, in this case, the service is doing much more work (positions extractions and distances computations) than just expressing a request describing its needs.

VI. Conclusion and Future Plans

In this paper, we presented O3MiSCID: an Object Oriented Opensource Middleware for Service Connection, Introspection and Discovery. This middleware has a layered architecture. The lowest layer abstracts network communications using BIP/1.0 [10], a performant, low-overhead protocol. The second layer is dedicated to creation, introspection and discovery of BIP services. The last layer, which addresses ontology and more advanced features, is still under development. This middleware is specifically designed for pervasive environments and can maintain low latency and high bandwidth communications, allowing the creation of context-aware interactive applications featuring for instance audio stream and video stream processing.

Its implementation is based on DNS-SD (DNS Service Discovery) over mDNS (multicast DNS) and features a robust decentralized architecture for service advertising and discovery. DNS-SD we favored over UPnP because of the requirement of have highly coupled interaction between services and so require quick and robust service notifications of service status

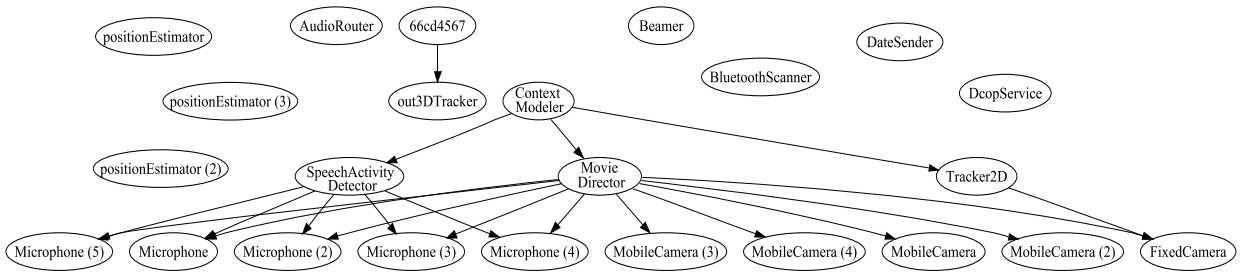


Fig. 3. Interconnection graph of services running in our perceptive environment

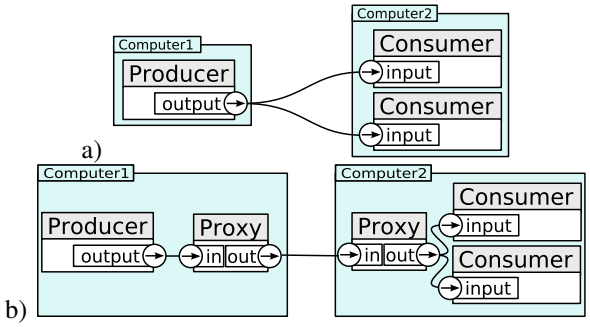


Fig. 5. Proxy Architecture
a) Two clients on a same machine are connected to a same remote service.
b) The addition of a proxy services result in dividing by two the network load. If a proxy is not present or down, the default behavior can still be used as a fallback.

updates, in particular connection and disconnection. DNS-SD also provides a convenient built-in handling of unique names and limited service descriptions.

Implementations already exist in multiple languages such as Java, C++ and Tcl. This middleware is already used by three research teams to deploy interactive applications prototypes, and is revealing to have a high adoption rate.

Since we aim at a high quality, open source release of the O3MiSCID middleware, we still have to achieve cross-implementation code review and interoperation tests to ensure the coherence between all the implementations. New documentation and tutorials also need to be written. We are currently working on this point in order to quickly distribute O3MiSCID on the Sourceforge platform or similar.

A general proxy architecture would be an interesting addition, in order to optimize the network load. Figure VI gives an illustration of the proxy approach. This proxy system is particularly needed for wide streams such as audio and video streams; some middlewares are in fact dedicated to such proxying [15].

The middleware currently satisfies all needs concerning discovery and introspection of the existing services. Through enhancing ontological information about the services and their operations, we envision to tackle with the problem of automatic service configuration and composition.

ACKNOWLEDGMENT

The authors would especially like to thank Sébastien Pesnel, former PRIMA project member, for his considerable development effort concerning the O3MiSCID middleware.

REFERENCES

[1] DNS-SD web site http://www.dns-sd.org/. [Online]. Available: http://www.dns-sd.org/
[2] "UPnP Device Architecture," http://www.upnp.org-/download/UPnPDA10_20000613.htm, 2000.
[3] E. Guttman, "Autoconfiguration for IP Networking: Enabling Local Communication," *IEEE Internet Computing*, pp. 81–86, June 2001.
[4] Multicast DNS web site http://www.multicastdns.org/. [Online]. Available: http://www.multicastdns.org/
[5] T. Gu, H. K. Pung, and D. Q. Zhang, "A service-oriented middleware for building context-aware services," *J. Netw. Comput. Appl.*, vol. 28, no. 1, pp. 1–18, 2005.
[6] H. Chen, T. Finin, and A. Joshi, "Semantic web in the context broker architecture." [Online]. Available: citeseer.ist.psu.edu/646646.html
[7] M.-T. Tran, B. Hirsbrunner, and M. Courant, "A context-aware middleware for multimodal dialogue applications with context tracing," in *MPAC '05: Proceedings of the 3rd international workshop on Middleware for pervasive and ad-hoc computing*. New York, NY, USA: ACM Press, 2005, pp. 1–8.
[8] Jini web site http://wwws.sun.com/software/jini/. [Online]. Available: http://wwws.sun.com/software/jini/
[9] J. O. Kephart and D. M. Chess, "The vision of autonomic computing," *Computer*, vol. 36, no. 1, pp. 41–50, 2003.
[10] J. Letessier and D. Vaufreydaz, "Draft spec : Bip/1.0 – a basic interconnection protocol for event flow services," 2005. [Online]. Available: http://www-prima.imag.fr/prima/pub/Publications/2005/LV05/
[11] A. Paar, J. Reuter, and J. Schaeffer, "A pluggable architectural model and a formally specified programming language independent api for an ontological knowledge base server," in *Australasian Ontology Workshop (AOW 2005)*, Dec. 2005, publikation.
[12] H. Chen, F. Perich, T. Finin, and A. Joshi, "SOUPA: Standard Ontology for Ubiquitous and Pervasive Applications," in *International Conference on Mobile and Ubiquitous Systems: Networking and Services*, Boston, MA, August 2004.
[13] S. Borkowski, J. Letessier, and J. L. Crowley, "Spatial control of interactive surfaces in an augmented environment," in *Engineering Human Computer Interaction and Interactive Systems, Joint Working Conferences EHCI-DSVIS 2004, Revised Selected Papers*, ser. Lecture Notes in Computer Science, R. Bastide, P. A. Palanque, and J. Roth, Eds., vol. 3425. Springer, jul 2004, pp. 228–244, eHCI 04. [Online]. Available: http://www-prima.imag.fr/prima/pub/Publications/2004/BLC04/
[14] F. Metze, P. Gieselmann, H. Holzapfel, T. Kluge, M. Wolfel, J. L. Crowley, P. Reignier, D. Vaufreydaz, F. Bérard, B. Cohen, J. Coutaz, S. Rouillard, V. Arranz, and M. Bertran, "The fame interactive space," in *2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Edinburgh - UK, feb 2005, p. 4. [Online]. Available: http://www-prima.imag.fr/prima/pub/Publications/2005/MGHKWCRVBCCRAB05/
[15] NIST Smart Data Flow System web site http://www.nist.gov/smartspace/nsfs.html. [Online]. Available: http://www.nist.gov/smartspace/nsfs.html

Smartphone-based User Location Tracking in Indoor Environment

Viet-Cuong Ta[1,2], Dominique Vaufreydaz[1], Trung-Kien Dao[2], Eric Castelli[2]

[1] Pervasive Interaction/LIG, CNRS, University of Grenoble-Alpes, Inria, France
[2] MICA Institute (HUST-CNRS/UMI2954-Grenoble INP), Hanoi University of Science and Technology, Vietnam

Author Version

Abstract

This paper introduces our work in the framework of Track 3 of the IPIN 2016 Indoor Localization Competition, which addresses the smartphone-based tracking problem in an offline manner. Our approach splits the path-reconstruction into several smaller tasks, including building identification, floor identification, user direction and speed inference. For each task, a specific set of data from the provided log data is used. Evaluation is carried out using a cross validation scheme. To produce the robustness again noisy data, we combine several approaches into one on the basis of their testing results. By testing on the provided training data, we have a good accuracy on building and floor identification. For the task of tracking the user's position within the floor, the result is 10m at 3rd-quarter distance error after 3 minutes of walking.

1 Introduction

With the widespread of the smartphone and related technologies, tracking users through their phones becomes one of the main research topics of user positioning. Track 3 of the 2016 IPIN Competition addresses the problem in an offline scenario. The data are collected by the organizers in a setup that is similar to the real world situations of daily phone usage. The data thus comes with some noisy and unexpected patterns. Moreover, the required tracking length is at a large scale in term of space and time.

The objective of the competition is to construct as close as possible the path of the users, providing we have full access to data in the smartphone. The number of the users is not specified. Meanwhile, there are four different phone models, which are Samsung Galaxy S3, Samsung Galaxy S3 mini-model, Samsung Galaxy S4 and Google Nexus 5. The collected data contains 12 different types of sensor. Each sensor can be viewed as an independent data stream which have specific update rate (samples per second) and sensor's value output. Due to the hardware dependent properties, each data stream are likely to have some differences for a specific phone model. The collected area involves four different building with multiple floors. The competition then requires to identify the user's position which includes the building, the floor of the building and the latitude/longitude. The evaluation score is a function of building, floor and the distance errors. While the building and floor errors penetrate a wrong prediction with a building cost and floor cost, the distance error is calculated by the 3rd-quarter of the distance errors of all the valid points. A valid point is a point which is predicted with the right building and floor. The position of the user is needed to update every 0.5 second. The training data includes several routes for each building and also has ground truth positions. It is up to contestants to use which types of sensor to track the phone. One important note is that the phone is not required to be handled in a same position throughout the collecting process. There are also supplementary data which include all the building maps and videos of the data collection process.

Intuitively, the process of user tracking starts from identifying building to the floor of the building and then the 2D positions within the floor. We select some specific subset of sensors to carry out these tasks (see Fig. 1):

- Identifying building: using GPS value (`GNSS` tag) and appeared Wi-Fi MAC address data (`WIFI` tag).
- Identifying floor: using Wi-Fi data to build fingerprinting database, then learning floor ID from the database by several models.
- Approximating the 2D path: we assume that for each floor, the user needs to enter and leave at some specific points. The remaining task is to reconstruct the path from these two locations. For solving this, we apply a particle filter. The moving model bases on the inertial data, which are accelerometer, magnetic and gyroscope sensors. Additional adjustment steps are carried out, which could based on Wi-Fi data or map information.

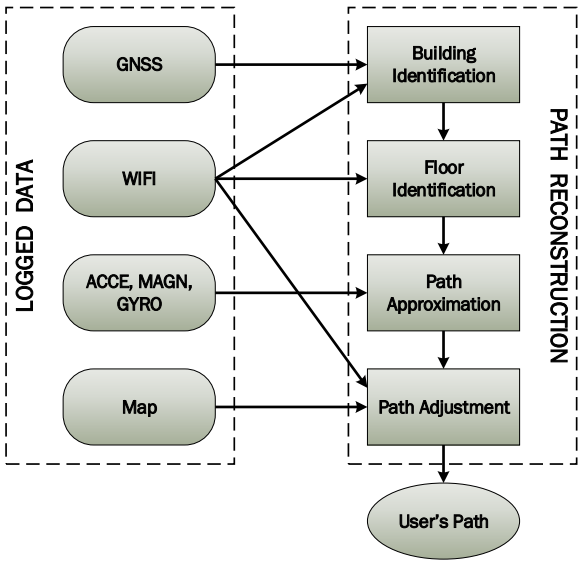


Figure 1: Possible use of sensors type for path reconstruction

While it is straightforward to find stable approaches for identifying building and floor, the path approximation within each floor remain the biggest challenge. The data brings out many practical problems. In term of wireless positioning approach, the system has to handle a spare Wi-Fi fingerprinting database, unreliable scanning periods and multiple devices. For tracking through inertial data, there is no guarantee that the phone's sensors are calibrated. It is therefore difficult to create a robust system that can identify all these types of issue. In our approach, we start to build the path from a simple solution, based on inertial sensors. Then, we try to integrate more sources of information to reduce the drifting errors over time. Several approaches are proposed and tested on a selected path of the provided data. As our best results to now, we can reach around 10m in 3 minutes and 24.5m in 7 minutes, at 3rd-quarter error.

The rest of the paper is arranged as follows. Related works are presented in Section II. Then, we propose our approach for identifying building and floor in Sections III and IV, respectively. In Section V, we try to solve the problem of path approximation with the objective to minimal the distance error. Finally, Section VI provides our summary on the works.

2 Related works

The most straightforward way to track a smartphone is to use the GPS/GNSS data. The GPS/GNSS data is a built-in functionality in most of the smartphone models today. However, its performance in the indoor environment is questionable due to the blocking effect of surrounding environment [12]. Other alternative wireless tracking technologies include RFID, Cellular-Based, UWB, WLAN and Bluetooth [9]. Among them, WLAN has got the most attentions because of its availability in the real world environment. One of most well-known approach for WLAN is fingerprinting, which the traditional positioning problem is transformed to a statistical learning problem. As reported in [15], fingerprinting models can reach the error as low as 2m for indoor environment. In a wild application setup, the approaches stay around 6m with mean squared distance error. These kind of data and results are available in Track 3 of the 2015 IPIN Competition [13]. Compared to the normal database which is used in fingerprinting, the Wi-Fi database extracted from the provided data in the 2016 Competition is relatively small and sparse. There are also four different phone models, which lead to a source of noise on hardware variation [7].

Besides WLAN data, the smartphone can be tracked by signals from its inertial sensors. Three types of sensors, i.e., accelerometer, mangnetometer and gyroscope, provide a good estimation for the phone's movements to some extend. In general, several techniques have been developed for inferring the movements from the signals of inertial sensors [6]. Normally, the task is divided into two parts: one for finding the speed and the other for the direction. One of the principal challenges in user tracking is to handle the drifting effects of the sensors. Kalman and particle filters are two popular choices to work against these problems for short-term errors. For the long-term drifting, it is normally required to have a calibration technique in addition [8]. The direction of using map information is also feasible and is discussed in [14][1].

In terms of finding the direction, there exists some advance filters such as Madwgick filter [10] and Mahony filter [4]. It should be noted that the sensors equipped in smartphones nowaday usually contain much noise. For

Table 1: Number of Wi-Fi scans on UAH Building data

Route	All	Floor0	Floor1	Floor2	Floor3	AvgDistance
R1-S3	80	20	32	19	0	5.77
R1-S4	139	23	54	45	0	4.30
R2-S3	97	64	0	0	19	6.08
R2-S4	149	92	0	0	36	4.82
R4-S3	40	0	18	16	0	4.72
R4-S4	64	0	19	38	0	4.02
Total	569	199	123	118	55	4.96

Android devices, the errors can be as large as 60° for 3 minutes of tracking [16]. On the other hand, tracking the user speed could be more straightforward if we can have the constraint that there are only two movement patterns, namely standing and walking. The number of steps can be counted and then multiplied with step length to find the moving distance [8].

3 Building Identification

In the supplementary data, each building is included with the WGS coordinates of the building. By using these coordinates, the distance between pair of buildings is calculated. There is only the case of UJITI and UJIUB, which have a distance of around 450m. The other pairs are quite far from each other. Beside that, each building comes with a specific set of observed `MAC_BSSID`. There is no `MAC_BSSID` belong to two buildings. We then assume that GNSS data and appeared `MAC_BSSID` could be used to identify the building efficiently.

4 Floor Identification

In this task, we select the UAH building data for validating our model because it has a high number of observed Wi-Fi stations. There are 353 `MAC_BSSID` records appeared in the given data.

4.1 Fixing the POSI data and create the groundtruth data

The training data include `POSI` tag, which is the checkpoints along the user's trajectory. The time period between consecutive `POSI` records varies and is longer than the required sampling time, which is 0.5s. In most cases, the user takes a linear trajectory between two subsequent `POSI` records. However, there are several segments which do not follow the observation. It is thus mandatory to correct those segments. We add some virtual checkpoints along the ambiguous segments. The timestamp together with these virtual checkpoints are calculated from the ratio between the virtual moving distance and the time to complete the real segment. Those virtual checkpoints are put manually on the basis of the provided maps and videos. The small parts where the trajectory crosses the stairs are not corrected.

After fixing the trajectory, the position at a fixed time t will be computed by a linear interpolation between two consecutive checkpoints P_i and P_j, respectively, before and after t. The Wi-Fi fingerprinting database is then created by joining the groundtruth position with the Wi-Fi signals recorded in the log file. Each `WIFI` record is provided with: application timestamp (`AppTimeStamp`), sensor level timestamp (`SensorTimeStamp`), name (`Name_SSID`), MAC address (`MAC_BSSID`) and signal strength (`RSS`). The `SensorTimeStamp` field of the `WIFI` record is used as the time indicator for Wi-Fi.

To produce a completed scan from `WIFI` records, the appeared sensor times are grouped into separated time periods. Each period has the length of 4.5s. Specifically, two `WIFI` records are considered in a same scan if the difference in their `SensorTimeStamp` is less than 4.5s. From the completed scan, we create a feature vector of 353 dimensions, and set the reported RSS value of seen Wi-Fi stations as the feature value. If there is an unseen Wi-Fi station in the completed scan, its value is set to a constant value (`WIFI_ZERO`).

4.2 Wi-Fi fingerprinting results on floor identification

The UAH building comes with four floors (ID from 0 to 3). There are 3 routes in total, with two different devices, namely Samsung Galaxy S3 and Samsung Galaxy S4. Table 1 shows the extracted number of the Wi-Fi fingerprinting data. The `All` column is the number of the `POSI` records appeared in the log file. The `AvgDistance` is the average distance between consecutive completed Wi-Fi scans. The stats indicate that the two phone models Samsung S3 and S4 have some variation within the Wi-Fi scanning data.

Table 2: Floor accuracy on 5 folds cross validation

Id	Model	Raw	2-filter	HLF
1	RF classifer	95.52%	94.28%	92.70%
2	RF regressor	89.80%	91.82%	93.76%
3	KNN classifier	91.47%	91.30%	91.47%
4	KNN regressor	90.60%	90.33%	90.69%
5	XGB classifier	98.24%	97.80%	97.36%
6	XGB regressor	99.38%	98.41%	98.77%

In the task of learning the floor ID, we apply three family models: K-Nearest Neighbour (KNN), Random Forest (RF) [2] and Extreme Gradient Boost (XGB) [3]. While KNN is a popular choice to work with the Wi-Fi fingerprinting data, the others propose learning abilities on a small amount of training data. We also vary the set of features:

- Raw features: The RSS value is used as default value. In case of KNN model, the Euclidean distance is used. We set `WIFI_ZERO` to -150.

  This value is sensitive for the KNN models. The other models, based on decision tree, are unlikely affected.

- Filtering [11]: The set of features is derived from the winner of Track 3 of the 2015 IPIN Competition. In our experiments, instead of splitting into groups like their proposal, we add an addition of D features to represent that. The level of filters is set to two, namely 2-filters. Assume that among D access points, if the i^{th} and j^{th} ones have the highest reported RSS, we set the value of i^{th} and j^{th} in D additional features to an INFINITE value. This was can ensure the robustness of Euclide distance. The total number of features after this operator is $2D$.

- Hyperbolic Location Fingerprinting features [7]: The feature comes from the fact that the data come from two different devices. We simply take the subtraction of the RSS values as the feature between Wi-Fi access points i, j. Normally, using HLF features would results into $D \times (D-1)/2$. Because of the little amount of data we have got, it would be unpractical for training model if we use such an large number of dimension (around 500 samples against over 60000 dimensions). An additional dimensional reduce step thus is necessary. In our experiments, we use the Random Tree Embedding approach [5] to get a 700-dimensional vector approximately.

Each type of model is tested with both options, i.e., classifier and regressor. In the floor learning task, it is straightforward to use a classifier to learn the feature space. In case of using a regressor, the target is a real number indicating the floor, and the output is also a real value x in the range of $[0, 3]$. In order to converting from x to a floor ID, we use a cut vector $C = \{c_1, c_2, c_3\}$. A value x is classified to the Floor i if $x \leq c_{i+1}$, otherwise, x is classified as the Floor 3. The value of C could be computed directly on training data as an sub optimization step. After the regression model is trained, we get the model's prediction values on all the training data. Then, C is selected in a way that maximize the prediction accuracy of the prediction values and the training targets. On the other hand, this method yields a potential overfitting issue.

Table 2 shows the results on 5 folds random cross validation. XGB regression model with Raw feature get the highest results. There is no significant separation between three types of features. The using of regressor with an additional layer of optimization could provide a good solution to the floor identification. It can be explained that the label, which is the number of floor actually, contains some continuous relationship. For example, in the vector space, the samples of Floors 0 and 1 could be overlapped easily than those of Floors 0 and 2.

5 Path Approximating within a Floor

At this point, based on the floor cross validation results, we assume that the floor is able to be predicted correctly. Therefore, the remaining task is to approximate the trajectory within the floor. Moreover, as the users can only use stairs or elevators for moving out or into the floor, the task could be reduced to find an approximate path from a starting point. In other cases, GNSS data and Wi-Fi data could be used to find out the point. As it is simple to produce correct groundtruth path for it, we select Floor 1 from UAH building for testing the idea on path approximation within the floor.

5.1 Using GNSS data

The GNSS data comes with the latitude/longitude of the smartphones. Table 3 provides the details of the GNSS data within Floor 1.

Table 3: GNSS data on the Floor 1 of UAH Building, the errors is reported by 3rd-quarter errors

Route	N	Update rate	Longest missing	Error
R1-S3	44	9.5s	25.0s	466.2m
R1-S4	186	2.3s	127.0s	29.4m
R4-S3	14	7.7s	20.0s	15.6m
R4-S4	0	-	-	-

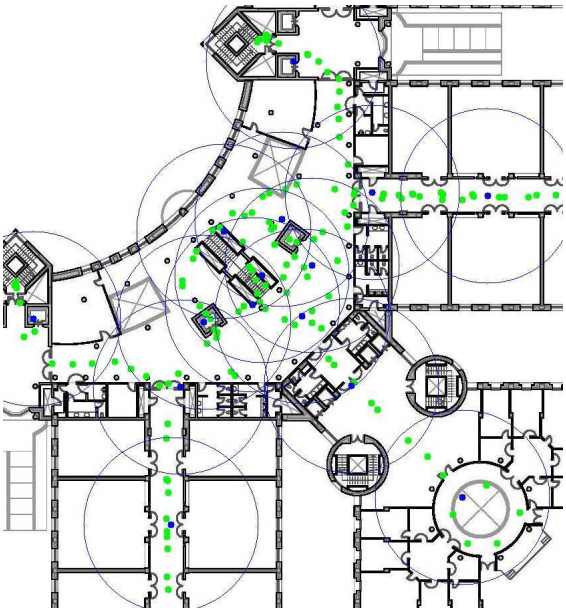


Figure 2: The green dots are the training Wi-Fi points, and the blue dots denote the center of the clusters. The radius of the clusters circle is set to 10m to visualize the quality of the clustering.

There exists a significant difference in the update rate between two phone models, Samsung Galaxy S3 and Samsung Galaxy S4. The S4 phone is updated more frequently than the S3 one. On the other hand, the S4 phone suffers a large missing period. This would make using the GNSS data from the S4 less reliable. The expected errors for using GNSS data is around 30.0m. However, the results is likely unstable.

5.2 Inferring position using Wi-Fi data

The same feature set and models from previous experiments are used. It is straightforward that regressor type models can be applied for learning the position from the Wi-Fi data. For the classifiers, it is possible to transform the 2D coordinates target, including latitude and longitude, to a label. The most popular way is to divide the area into several smaller grids of fixed size. Another approach could involve manually picking the points by using the provided map. In our experiment, we perform a K-means clustering on latitude, longitude of all the groundtruth points. Figure 2 is a result of K-mean clustering with the number of centers fixed to 15.

Table 4 reports the distance errors on 5 folds cross validation. With the total of 123 available points, the effective training samples are around 100 points and 23 points for testing. The best result are around 6.0m by using XGB model with 2-filters feature. In fact, there is a similar performance between three set of features. With three set of features, we prefer to use the HLF feature set because the device noise could be high, especially

Table 4: Errors on Floor 1, UAH building for 5 folds cross validation, report by 3rd-quarter errors

Id	Model	Raw	2-filters	HLF
1	RF classifer	10.6m	11.5m	12.9m
2	RF regressor	13.6m	16.1m	16.4m
3	KNN classifier	10.3m	10.3m	10.3m
4	KNN regressor	9.7m	9.4m	9.1m
5	XGB classifier	6.6m	6.0m	6.2m
6	XGB regressor	9.1m	8.6m	8.7m

we only have a small size of training data. With the tree-based model, classifier approaches outperform the regressor ones. At this step, we can build the user path by interpolating the outputs of the Wi-Fi model. However, like in GNSS data, there is also a long period of time that the phone does not receive any new Wi-Fi data. The errors also are unpredictable if the user moves too far from the appeared area in the training data.

5.3 Inferring direction using magnetic, accelerometer and gyroscope sensors

In order to identify the phone's direction, we look into the data of four types of sensors: magnetic sensor (`MAGN` tag), accelerometers (`ACCE` tag), gyroscope (`GYRO` tag) and ahrs sensors (`AHRS` tag).

The sampling rate is not the same between those types of sensors and between different phone models. In addition to that, the rate can vary in a great range for a walk of the user For example, in the GYRO data of S3 model, it is updated with every 5ms in average. However, at some point, it takes 25ms to get the new data.

Several ways of extracting the phone's direction are tested, including:

- Based on magnetic and accelerometer ($AccMag$): at time t, the data from nearest `MAGN` and `ACCE` read are used for constructing the rotation matrix. Then, the $AccMag$ orientation is devised from the matrix. This is the most standard way for finding the phone orientation. The value of $AccMag$ orientation is supposed to be the low update rate part of the direction. It is then fused with the high update rate part by two other methods, which evolves the integration of the gyroscope data.
- Based on the integration of gyroscope (`GYRO`): the values are used to feed to the system with the $SensorTimestamp$ value. The integration is then calculated on 3-axis, Azimuth, Pitch and Roll. An further step of fusion is added by using a constant α:

$$Gyro = (1 - \alpha) \times IntegratedGyro + \alpha \times AccMag \tag{1}$$

  where α is a threshold weighting contribution of these two factors.
- Use Madgwick filter [10]: the filter is designed to fix the magnetic disortion and gyroscope drifting errors in commercial IMU. It uses a quaternion representation for accelerometer and mangnetometer data. An optimized gradient-descent algorithm is employed to track the gyroscope errors. In our implemenation of Madgwick filter, we downsample both gyroscope and magnetic data to be equal to the update rate of accelerometer data. The starting quaternion is calculated directly from the $AccMag$ above.
- Using `AHRS` data directly: `AHRS` data is received every 0.02 second for both S3 phone and S4 phone, roughly.

From all of the approaches above, we use the value of Azimuth axis (or Yaw) as the user's direction. The direction can have the update rate as high as 0.02 second, which is equal the update rate of accelerometer sensor. Therefore, before intergrating with the speed, it should be downsampled.

5.4 Inferring the moving distance using accelerometer sensor

Figure 3 plots the Z-axis of accelerometer when the user's movement contains both standing and walking movement. There are specific patterns in the signal. For simplicity, we calculate the standard deviation for a fixed window length dt and use a threshold M for splitting between standing and moving. In order to find the suitable value of M, the standard deviation of Z-axis is computed on all the data (Fig. 4). From the figure, there is a point which the standard deviation pattern changes significantly. In this case, the point is around 0.1. Therefore, it is straightforward for choosing $M = 0.1$ in this case.

Whenever the standard deviation of a window length dt is above M, we infer the moving distance as:

$$MovingDistance = AvgSpeed \times dt \tag{2}$$

where $AvgSpeed$ is the average speed over the floor, which could be calculated based on the training data. The local variance of the user's speed is discarded here. We choose to calculate the moving distance based on the $AvgSpeed$ instead of the number of steps because the user's step length is not available.

5.5 Combining the direction and moving distance for the full path approximating

From the calculate direction and moving distance, we downsample by a window of time length $dt = 0.5$ second, which meets the requirements of the competition.

The approximating path will be built by using particle filter, which is a popular choice in this task. As our observations on the data, it is difficult to create an effective observation model from the Wi-Fi scanning results.

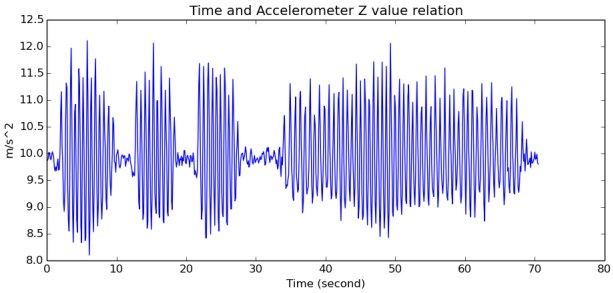


Figure 3: Z-axis plot of walking and standing movements

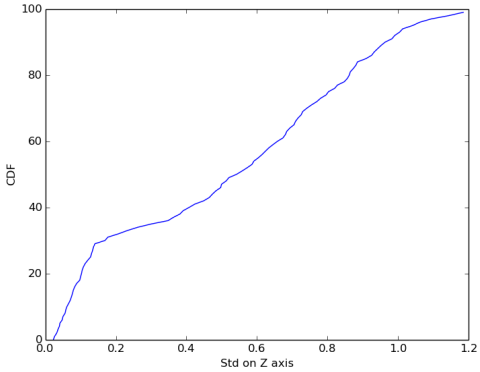


Figure 4: Cumulative density function of standard deviation for each window length of 0.2s, i.e 10 `ACCE` data packages.

There are two main reasons: firstly, there exists long periods of time in the data that Wi-Fi scanning result does not present; secondly, after 4 seconds in general, there could be a huge drifting in the particle's position which can not be adjusted by the Wi-Fi fingerprinting results. Therefore, our particle filter have only the moving model.

To evaluate our approaches, we choose the Route 1 at UAH building. The data is extract from the user's walk with the Samsung Galaxy S3 phone model. The period when the phone entered and left the Floor 1 is from 132.25s to 565.00s, approximately. To prevent the drifting errors, we use only 3 minutes from that period length.

Table 5 shows the results in term of mean squared distance error (MSE), mean error and 3rd quarter error. During a period of 3 minutes, Gyro and Madgwick Filter reach a similar error of 20.0m. The Gyro approach still is more stable than the Madgwick filter as it has both a low MSE and the median error. Meanwhile, the others, $AccMag$ and AHRS, cannot provide a good approximation of the phone's direction.

We plot the output paths in Fig. 5. The turning angles from the Gyro and Madgwick filter are highly correlated with the user's movements. The under-performance of Madqwick filter could come from our specific implementation on the identifying the first quaternion. The technique suggests that a motion stop is needed, instead of calculating it directly as in ours. In addition to that, there is a downsampling step for `GYRO` and `MAGN` data, which could introduce more noise.

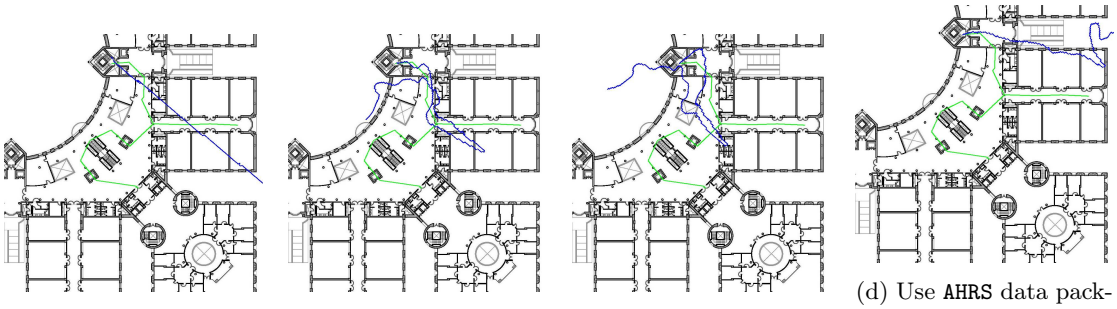
Compare to the performance of Gyro and Madwick filter, the correlation of AHRS and $AccMag$ is quite low. It is reported that the magnetic sensor has a poor performance in the indoor environment. Moreover, there is not guarantee that a calibration process of the magnetic sensor is carried before the data is collected. As the magnetic data are presented in computing the direction of Gyro and Madgwick filter approaches, it could possibly reduce the accuracy of those output directions.

5.6 Path adjusting

In the above setup, the drifting error could be as high as 20m after only 3 minutes approximation. For a longer period, i.e. 7 minutes in this specific case, the error is likely unpredictable. Therefore, additional calibration steps should be employed for controlling the drifting affects. There are several sources of data could be used for this purpose, which includes GNSS, Wi-Fi and map information. Comparing between GNSS and Wi-Fi, the Wi-Fi data has a more stable update rate and also a better accuracy. Therefore, using the Wi-Fi data to adjust

Table 5: MSE distant errors, median and 3rd-quarter errors for in 3 minutes

Technique	MSE	Median	3rd-quarter
AccMag	43.7m	11.7m	65.0m
Gyro	13.12m	6.5m	19.4m
Madgwick Filter	17.2m	12.4m	20.4m
AHRS	39.0m	17.1m	55.5m



(a) AccMag direction (b) Gyro direction (c) Use Madgwick filter (d) Use `AHRS` data package

Figure 5: The generate paths from the 4 techniques. The blue line is the real trajectory, the green path is the generated ones.

the position is quite a popular approach, including [8] or the on-site Track 1 of the 2015 IPIN Competition [13]. On the other hand, integrating the map information for calibrating the errors could possibly make the system less stable. Moreover, the performance highly depends on the environment and on how map is modeled. For example, [1] uses the assumption that the building has a rectangle shape for reducing the direction drifting. In [14], the map information is used at different levels, including using polygons to model the floor. In this section, we are going to explore several directions of using Wi-Fi and map information to get stable results.

5.6.1 Combine Wi-Fi data and the path approximation

It is unlikely that the observation model built on the Wi-Fi fingerprinting results can calibrate the drifting errors generated by the sensors of the smartphone. However, whenever the phone complete a Wi-Fi scan, we can have a small adjustment of the particles at that specific time t. Let the Wi-Fi feature vector at time t is X, and Y is the position estimated by the Wi-Fi fingerprinting model. A particle P then can be pulled to position P_{new} near M with a constant λ:

$$P_{new} = P + \lambda \times (Y - P) \tag{3}$$

In the next section, we extend the above idea by identifying a local value of Y base on the results of classifier model. From the results of Wi-Fi fingerprinting in a cross validation scheme, it can be concluded that the classifier outperforms the regression. In order to employ the classifier output efficiently, we use the prediction probability of each class. Given there are N classes, the output of each model (KNN, RF or XGB) for a sample X can be in the form of:

$$prob_X = \{p_1, p_2, ..., p_N\} \tag{4}$$

where p_i is the probability of X belonging to centers C_i. At the time of adjusting position for a particle P, we already know the its position. It is thus possible to find the three most nearest centers around P. For examples, let them be C_1, C_2 and C_3. The corresponding probability are p_1, p_2 and p_3. Those probability then is normalized to have the sum being equal to 1. The estimated position Y is the weighted sum of the three centers C_1, C_2 and C_3 with the p_1, p_2 and p_3 are the coefficients:

$$Y = p_1 \times C_1 + p_2 \times C_2 + p_3 \times C_3 \tag{5}$$

5.6.2 Adjust the wall-crossing step

This is the most straightforward usage of map information. The movement of each particle is constraint by walls. Those wall is drawn manually from the provided map data. If a particle goes cross a wall, its direction is adjust to go parallel with the wall.

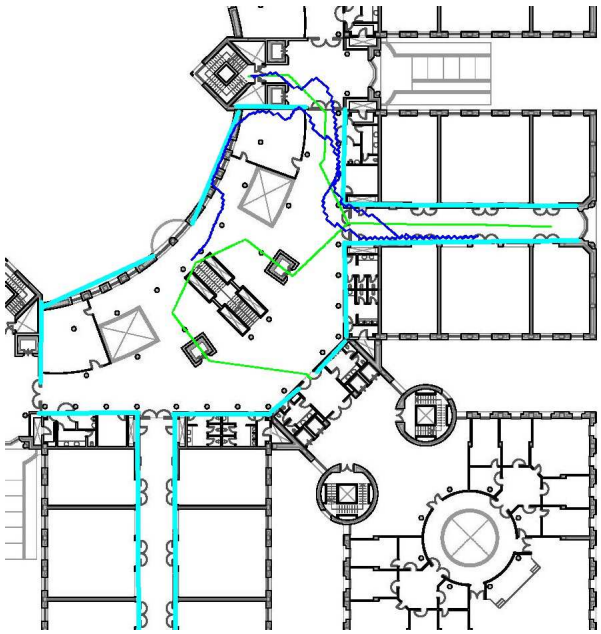


Figure 6: Apply local adjustments for fixing wall crossing. The wall is colored with the cyan

Figure 6 illustrates this idea. For simplicity, some of the door along the walls are ignored. By applying the adjustment on the Gyro approach from above, we can have the 3-quarter error distance stays around 11.2m, which is a significant reduction from 19.4m. This approach could be efficient for stabilizing the errors in a short term. On the other hand, its performance depends on the properties of the area where the user is standing. When the user goes to the large hall, where there is no wall, the accuracy is heavily affected by the drifting in direction.

In the capability to fix long-term drifting, this approach is still not sufficient. The adjustment operators only have local effects. As we can see from Fig. 6, the source of the errors is not at the time when the particle crosses the walls. It come from a changing direction at some place in the previous path. Thus, it leads us to a more complicated approach.

5.6.3 Minimize the number of wall crossing

In this part, we try to define an optimization problem for path adjusting. Firstly, we create two necessary operators for changing the path. Assuming that the errors mainly come from the noisy magnetic data in indoor environment, a rotation operator could be apply to fix. Moreover, the using of *AvgSpeed* is not flexible enough to describe the velocity of the user. In general, the user can move faster or slower than the *AvgSpeed*. Therefore, an additional normalization operator is needed for deriving the local speed from the global *AvgSpeed*.

At this step, it is sufficient to find an optimal set from these two operators with the objective is to minimize the number of wall-crossing. However, to reduce the search space and also make the results more illustrative, we divide the map of each floor into smaller regions. Each region will have several pivot points and its walls. The pivot points are used to signal the entering or leaving the regions. We then define the cost of a path within a region as the number of wall-crossing by following the path.

Beside that, the user's movement pattern can be split into two types, i.e., *move straight* and *turn*, on the basis of the output direction. For example, Fig. 7 plots the standard deviation on the direction for each time window of 2.5s.

A simple threshold of 0.25 is enough for classifying each small window length. Once each small part is assigned with a label, we use those parts as the seeds and extend their two endpoints. The extension process is terminated when a part with a different label is reached. The sequence is divided into a list of sub-parts. Two consecutive sub-parts are two different kinds of movement. We then put more constraints on the operators. The rotation operator can be used only on the *turn* sub-part, and the normalize operator can be used only on the *move straight* sub-part. Moreover, we only allow the starting of a new region after a *turn* sub-part.

To find a proper solution for the optimization problem, we employ a greedy strategy. From the starting configuration, the first collision with walls will be detected. Then the nearest previous *turn* and *move straight* sub-parts are selected to apply the two operators. The searching range of each operator is chosen from a predefined range. The range should reflect the standard deviation errors in the process of inferring movement speed and direction.

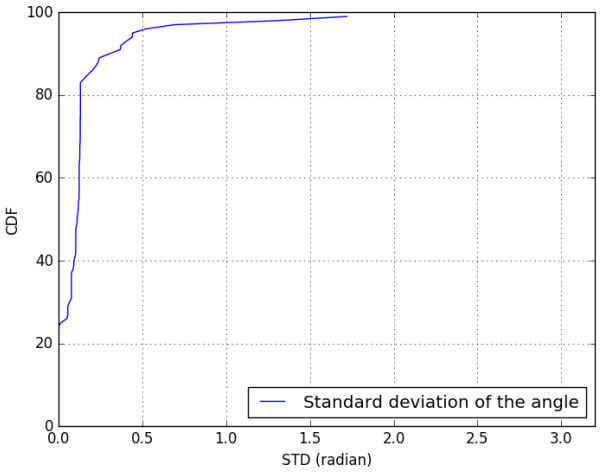


Figure 7: Standard deviation of the inferred angle for each window time 2.5s

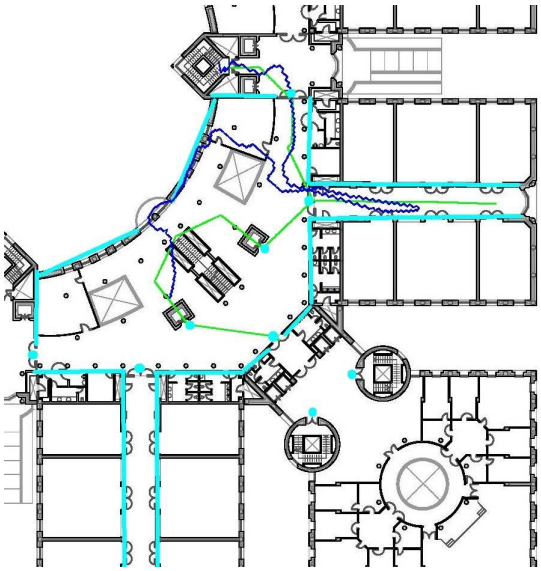


Figure 8: The new path after optimizing the numbers of wall crossing.

Figure 8 presents the results of the optimization process. The map are derived from the wall-crossing step by adding several points for the pivot locations. The regions are split on the basis of its wall characteristics with the purpose is to reduce the complexity of the searching step. Additional pivot points are added at the stairs and elevators.

In our setup, the range of rotate is set within 10°and the range of normalization operator is set to have a maximum absolute value of 0.1m/s. The output approximation path has a 3rd-quarter error distance of 11.2m. Like the wall-crossing adjustment, this approach depends heavily on the map configuration. For example, the role of long horizontal corridor is sensitive to the error reduction process. By applying a fix operator here, the remaining path is rotated toward the real path. On the other hand, in the large hall, where there are only a few walls, it becomes difficult to find the suitable adjustment because there is no error.

A global minimum could be reached by using a dynamic programming approach. However, it is difficult to ensure that the global solution is better than the local one, in term of distance error. Moreover, it would require a more complicated model to represent all the constraints in the problem. For example, the objective function should include a cost between choosing a normalization operators instead of a rotate operators. On the other hand, a more complex map model could lead to overfitting.

5.7 Combining several adjustment strategies

The adjustment steps above are combined in several ways to produce the outputs, which is listed in Table 6. More specifically, the Wi-Fi output only adjust the particles by the output position of Wi-Fi fingerprinting

Table 6: Distance errors at 3rd-quarter for 3 minutes and 7 minutes walking

Technique	3 minutes	7 minutes
Wi-Fi	16.4m	29.8m
Wall + Wi-Fi	14.2m	28.1m
Optimizing + Wall + Wi-Fi	10.1m	24.5m

model. In case of Wall + Wi-Fi, the particles are adjust to avoid hitting the wall first, then use the positions of Wi-Fi as above. The third approach runs the optimize algorithm on the generated directions, then is continued to process as in the second system. In all three approaces, the Gyro direction is preferred to used

The results in Table 6 are reported by running approximation for 3 minutes and 7 minutes, which are a half and full path of the floor, respectively. The best performance is the third system with around 25.0m errors at 3rd-quarter. The high errors of it could be explained by the three walking on long corridors in the Route 1, at Floor 1 of UAH building. If the system miss any of these paths, its distance error will increase greatly.

In term of avoiding overfitting, the Gyro + Wall system seems to be the most robust one. The two other ones' performance could depend on the characteristics of the path and the user's walking route to some extent. It is needed a more complete evaluation to choose the best approach. However, we are unable to make it available at this point because of the time constraint.

6 Conclusion

The data provided at IPIN 2016 Competition is very closed to the real world data. It thus comes with many practical problems. In this paper, we introduce the works for solving the tracking problems. There are some parts of the problems, e.g., building identification and floor identification, can be solved in a straightforward manner. On the other hand, when it comes to the path approximation, the system error becomes unpredictable. Through our understanding of the problems to point, a distance error of 30.0m at 75th percentile can be reached easily. However, it is quite difficult to match the performance which is described in the literature. For stabilizing our results, we try to find some alternative ways to reduce the long-term drifting. These proposed methods are tested and have good results on the selected data, though, they are at the very beginning of the work. There are many rooms left to make our approaches improve on a larger scale.

References

[1] Khairi Abdulrahim, Chris Hide, Terry Moore, and Chris Hill. Using constraints for shoe mounted indoor pedestrian navigation. *Journal of Navigation*, 65(01):15–28, 2012.

[2] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016.

[4] Mark Euston, Paul Coote, Robert Mahony, Jonghyuk Kim, and Tarek Hamel. A complementary filter for attitude estimation of a fixed-wing uav. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 340–345. IEEE, 2008.

[5] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[6] Robert Harle. A survey of indoor inertial positioning systems for pedestrians. *IEEE Communications Surveys & Tutorials*, 15(3):1281–1293, 2013.

[7] M.B. Kjaergaard and C.V. Munk. Hyperbolic location fingerprinting: A calibration-free solution for handling differences in signal strength (concise contribution). In *Pervasive Computing and Communications, 2008. PerCom 2008. Sixth Annual IEEE International Conference on*, pages 110–116, March 2008.

[8] Nisarg Kothari, Balajee Kannan, Evan D Glasgwow, and M Bernardine Dias. Robust indoor localization on a commercial smart phone. *Procedia Computer Science*, 10:1114–1120, 2012.

[9] Hui Liu, Houshang Darabi, Pat Banerjee, and Jing Liu. Survey of wireless indoor positioning techniques and systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(6):1067–1080, 2007.

[10] Sebastian Madgwick. An efficient orientation filter for inertial and inertial/magnetic sensor arrays. *Report x-io and University of Bristol (UK)*, 2010.

[11] A. Moreira, M. J. Nicolau, F. Meneses, and A. Costa. Wi-fi fingerprinting in the real world - rtls@um at the evaal competition. In *Indoor Positioning and Indoor Navigation (IPIN), 2015 International Conference on*, pages 1–10, Oct 2015.

[12] Veljo Otsason, Alex Varshavsky, Anthony LaMarca, and Eyal De Lara. Accurate gsm indoor localization. In *International conference on ubiquitous computing*, pages 141–158. Springer, 2005.

[13] Francesco Potortì, Paolo Barsocchi, Michele Girolami, Joaquín Torres-Sospedra, and Raúl Montoliu. Evaluating indoor localization solutions in large environments through competitive benchmarking: The evaal-etri competition. In *Indoor Positioning and Indoor Navigation (IPIN), 2015 International Conference on*, pages 1–10. IEEE, 2015.

[14] O. Woodman. *Pedestrian Localisation for Indoor Environments*. PhD dissertation, University of Cambridge, 2010.

[15] Moustafa A Youssef, Ashok Agrawala, and A Udaya Shankar. Wlan location determination via clustering and probability distributions. In *Pervasive Computing and Communications, 2003.(PerCom 2003). Proceedings of the First IEEE International Conference on*, pages 143–150. IEEE, 2003.

[16] Pengfei Zhou, Mo Li, and Guobin Shen. Use it free: instantly knowing your phone attitude. In *Proceedings of the 20th annual international conference on Mobile computing and networking*, pages 605–616. ACM, 2014.

Starting Engagement Detection Toward a Companion Robot Using Multimodal Features*

Dominique Vaufreydaz[1,2], Wafa Johal[2], Claudine Combe[1]
[1]Prima/Inria-LIG, CNRS
[2]University of Grenoble - Alpes, LIG, CNRS

March 12, 2015

Abstract

Recognition of intentions is a subconscious cognitive process vital to human communication. This skill enables anticipation and increases the quality of interactions between humans. Within the context of engagement, non-verbal signals are used to communicate the intention of starting the interaction with a partner. In this paper, we investigated methods to detect these signals in order to allow a robot to know when it is about to be addressed. Originality of our approach resides in taking inspiration from social and cognitive sciences to perform our perception task. We investigate meaningful features, i.e. human readable features, and elicit which of these are important for recognizing someone's intention of starting an interaction. Classically, spatial information like the human position and speed, the human-robot distance are used to detect the engagement. Our approach integrates multimodal features gathered using a companion robot equipped with a Kinect. The evaluation on our corpus collected in spontaneous conditions highlights its robustness and validates the use of such a technique in a real environment. Experimental validation shows that multimodal features set gives better precision and recall than using only spatial and speed features. We also demonstrate that 7 selected features are sufficient to provide a good starting engagement detection score. In our last investigation, we show that among our full 99 features set, the space reduction is not a solved task. This result opens new researches perspectives on multimodal engagement detection.

Keywords: multimodal perception - affective computing - healthcare technologies - companion robots

1 Introduction

Companion robots are entities that are intended to be used as assistants in everyday life, those being personal coach, desktop manager, etc. They could help to come up with tools that can potentially improve quality of life in the long run. Among usual embedded functions, one can find entertainment, video conference, objects grasping, activity monitoring, serious games and frailty evaluation [37, 8, 10, 9]. Companion robots can assist therapy for autism [6]. This paper presents research on companion robots using the Kompai Robot (see Figure 1)

*Author version.

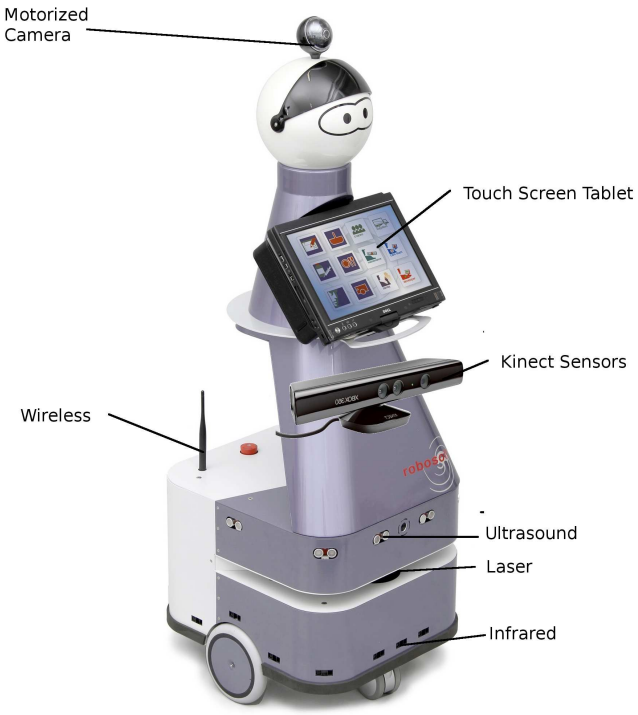


Figure 1: The Kompaï Robot from our partner Robosoft is equipped with a laser range finder, ultrasound and infrared telemeter, a tablet PC and a webcam on top. We added a Kinect for our experiments.

As argued in [20, 29], the primary challenge in building engaging companion robots is to provide social competency in perceiving, reasoning and expressing social and affective aspects of interactions with the human user. Companion robots are aimed to interact with humans in home environments. In order to stay credible, companion robots are expected to behave and react as per predefined manners corresponding to instructions and social signals used by the user.

As speech being multimodal in a face-to-face interaction, non-verbal communication also uses a variety of channels to convey messages. New areas explore techniques for the multimodal aspect of human communication in order to design robots able to read and express communicative signals in a social manner. Non-verbal cues of communications have been well studied for detection of emotions [30, 52, 46, 28]. In this paper, we propose to use these cues for intention recognition, and in particular an intention of interaction.

Recognition of intention is a basic skill acquired by infants early in their development. According to Vernon [45], among other skills, the perception of others' attention is crucial for the infant to master social interactions. The perception of intentions and emotions, present in newborn infants, helps to set their "preparedness" for social interaction [45]. Intention recognition allows the interacting agent to take quick decisions and to respond better to the user's need or state of mind. Some of the non-verbal communication signals are cues to subdued goals and intentions of the humans, and therefore a good way to improve adaptability of the robots' behaviors is by predicting their intentions. A part of human cognition is anticipation, allowing reading intentions and guessing goals in order to react quickly to stimuli. This skill is also very important for turn-taking in interaction.

In neurocognition, the Broca's area, responsible for language comprehension, action recognition & prediction and speech-associated gestures, would be the host of intention recognition in the human brain. According to Vernon, studies have shown that the activation of the Broca's area is significantly higher when a subject observes goal-directed actions with intentional cues rather than meaningless gestures.

As humans instinctively detect the intention of someone who wants to ask for way in the street, we are interested in the opening engagement phase of the process during which humans subconciously express their intentions to interact. Our goal is to investigate techniques to detect and recognize signals for non-verbal communication reflecting this intention and in our particular case, the intention of a user to engage an interaction with a robot.

Intention of engagement is a real question, especially when it comes to environments such as the work place or home, where people are not used to interact with robots [48]. Classically, the criterion for a user's intention of engagement is the spatial distance between the user and the communicant interface [11]. Some investigations have improved on this idea by also considering the speed of movement of the user [19]. These studies have chosen to use the relative spatial position of the concerned agents as criteria. The following assumption is made behind this choice: if the user is close to the robot, there is an intention to interact. Using distance and sometimes speed of the human provides with satisfactory results, but for a companion robot in real situations at home, close distance does not necessarily signal a desire for engagement. For instance, many times during the day, one can pass in front of the refrigerator without the wish to open it. Following the same logic, despite the physical distance of the user from the robot, a robot should be able to detect when it is about to be solicited, and anticipate the interaction in order to be more comfortable and socially acceptable.

In this study, we propose a multimodal approach for detecting a starting engagement using a RGB-D sensor mounted on a companion robot. Getting inspiration from social and cognitive sciences, our goal is to select features in order to improve the re-usability in other situations and/or with other sensors. In our approach, the idea here is to get rid of the usual way to do such experiment i.e. putting all available features together, combining them in a more optimized representation and let the training paradigm filter everything. Doing this, we might have good performances, but we may not learn anything about detecting intention of engagement. We will see that less than 10% of our features are crucial for starting engagement detection. In another context, one can make well-founded choices among sensors to reflect this knowledge. It will be more efficient to design a new device or robot knowing which particular features are of importance. This prospective research aim to build a set of meaningful features extracted from multimodal sensors useful for the description, recognition and discrimination of the intention of engagement.

This paper aims to contribute on the following statements :

- There exist subconscious social signals expressed by humans that characterize their will to interact with a robot and these signals are detectable.
- Some features from literature in the social and cognitive sciences are computable on a companion robot (notably Schegloff metrics [36]).
- Multi-modal features will perform better than spatial features to detect this starting of engagement in a home-like environment. A realistic dataset in a home-like environment can help us to validate this hypothesis.
- The set of relevant features for starting of interaction detection can be reduce without loss of performance using a feature space reduction process using the Minimum Redundancy Maximum Relevance (MRMR) method [13] never used in this context.

2 Multimodal Social Signal Processing For Non-Verbal Communication

2.1 Social Signal Processing

A communicative agent does not use only the verbal channel, but many channels to send and receive various messages while interacting [16]: human communication is intrinsically multimodal. To make human-robot communication fluent and acceptable, the robot has to decode these behavioral and non-verbal cues in order to act accordingly. For instance, computer systems and devices able to recognize

agreement or inattention, and capable of adapting and responding in real-time to these social signals in a polite, non-intrusive or persuasive manner, are likely to be perceived as more natural, efficient, and trustworthy[42, 45]. In the context of people assistance, social features seem to be crucial for the acceptance of a robotic companion in a domestic environment.

Argyle, in his book "Bodily Communication" [1] mentions different signals from different modalities used for non-verbal communication. The considered modalities are facial expressions, gaze, gestures & body movements, posture, contact, spatial behavior, clothing, and vocalizations. This work shows that recording of these modalities allows to recognize the mood of a person.

P. E. Bull [4] follows the idea that communication implies a socially shared signal system or code. Non verbal communication is argued to be intentional or non-intentional. Therefore, it is valuable information that allows to access intentions of the emitter that can be non-voluntarily transmitted. Bull claims the importance of posture and gesture in non verbal communication where these channels have been neglected compared to facial features and speech cues.

2.2 Intentionality in Human-Machine Interaction

The intention cues form a way of communication. As stated before, recognition of human's intentions, goals and actions is important in the improvement of non-verbal human-robot cooperation. [21] defines intention recognition as the process of estimating the force driving humans' actions based on noisy observations of human' interaction with their environment. Tahboub in [43] sees intention recognition as a substitution or a complement to reliable and extensive communication that is a prerequisite for coordination and cooperation. Indeed, in order to have a smooth interaction, intention recognition is essential. The DARPA/NSF final report on Human-Robot Interaction [5] recommends to improve the models of human-robot relationship and in particular to work on the intentionality issue.

In [19], it is proposed to recognize intentional actions using relative movements of a human towards a robot. An IR sensor embedded on the robot is monitored to track and estimate the velocity of a person. They then infer intentional actions such as approaches and departs using Hidden Markov Models (HMM) and position dependent models. Other related studies have defined human-robot proxemics in order to adjust inter-personal distance [47]. This work seems not enough to estimate engagement. On one hand, one can slow down near the robot without wishing to interact. On the other hand, someone passing swiftly by close to the robot might want to interact with it hastily.

The relative position and speed are not the only features that should be used to estimate intentionality. Multimodal fusion and usage of postural information have given good results in the measurement of quality of human-robot interaction and engagement of the user into the interaction [6, 34]. We aim to detect intention of interaction using also postural information and not only proxemics features.

In his study, Knight [18] points towards the importance for a robot to convey and hence to detect intentionality. It helps to clarify current activity and to anticipate goals. Learning from the engagement of humans, the robot should be able to anticipate the interaction and also to learn adequate moment when the robot itself can engage an interaction. In [39], engagement is defined as the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake. Engagement is in the frame of connection that can be a collaborative task, spoken language, gestures etc. Sidner et al. propose a model in three steps: (1) initiation of interaction, (2) sustainance of interaction, (3) disengagement. As presented in 3.3, we will see that our classification is based on this model.

3 Corpus for Engagement with a Companion Robot

A part of the work accomplished was to build a multimodal dataset including interactions with a companion robot equipped with a laser telemeter and a Kinect device. In this context, we focus on working with consumer devices and in a natural and non intrusive manner. Even though the tendency is the increasing usage of physiological sensors, such as R. Picard's pulse bracelet called Cardiocam, physiological signals still remain an invasive and relatively expensive option for users, for them to be released widely. The

physiological modality is not considered in this work, yet it might be enriching to include it in a future work that uses, for instance, contact-free heart rate measurements [31].

In order to validate our hypothesis in the context of interactions with a robot companion, the considered sensors are the ones commonly found on such robots: microphones, video sensors, depth sensors and telemeters. There exist available datasets in the field of social signals processing dealing with non-verbal communication which use multiple sensors. These datasets for emotion recognition are unfortunately more often built for face-to-face interaction where people sit and interact only with speech. The SSPNet association released the SEMAINE-DB dataset [41] where several persons were recorded in a face-to-face speech interaction. This database is suitable for a desktop environment that involves an interaction with a virtual commmunication agent. It is not well suited for human-robot interaction; especially as body cues in social signals are more diverse than facial expressions and speech characteristics. There exist other datasets using the Kinect sensor; such as Cam3D dataset centred on facial and hand movement associated with audio recording [22], or the LIRIS Human activities dataset [51] associated with human activity monitoring task. However, the proposition of a robot centred dataset for multi-modal social signal processing has not been made yet.

Looking at limitations of the existing multimodal datasets, the sensors equipping the Kompaï robot have been used to record a new robot view-point dataset where the users are interacting with the robot while standing. The scenarios included in this dataset will be presented in section 3.4.

3.1 Realistic Dataset

R. Picard in [30] states five variables that may affect data collection. (1) The first factor is the spontaneity of the behavior. The emotion can either be elicited by a stimulus or acted. (2) Another influence can come from the environment of the recording and the question here is: are the expressions of the participant similar in a lab setting and in a real-life situation? (3) Next question to be considered when recording affective data is: should the focus be on the expression or on the internal feelings of the participant? (4) The participant's awareness about the fact that he's being recorded. Indeed, what is the influence of open-recording in comparison with hidden recording on the recorded data? (5) Finally, should the participant be informed of the purpose of the experiment?

Regarding this research matter, (1) the engagement is relatively spontaneous, because the participants didn't act the interaction but were asked to interact whenever they wanted to with the robot. (2) The recording is made in a *living lab* environment, similar to a flat. The participants have no prior experience of this environment. This can create some fluctuations in their behavior. (3) We wanted to record intentionality of interaction, hence we focused on expression of social cues rather than to do a subjective evaluation. (4-5)We chose to not tell the participants that we were interested in the social cues of intention of interaction in order to collect more natural data.

3.2 Experimental implementation

The experimentation space is presented in figure 2. The apartment is divided into 3 areas: a living-room, a kitchen space and an empty space. To test our assumption that spatial information is not enough to detect an intention of interaction, furniture is placed so that participants will need to pass near the robot each time they want to go from one side to another of the experimentation room, even if they do not want to interact with the robot. This choice is an adverse condition as it leads us to distinguish someone passing close to the robot with or without intention of interaction.

In our recording, the robot is immobile. All features are robot centered but can also be computed with a mobile robot. The interaction in this dataset consists of playing a "tap the mole" game on the mounted tablet PC on the robot. Our hypothesis was that the interaction with a companion robot is also preceded by a pre-interaction phase (see section 3.3) where the participant shows some subconscious social signals of its intention of interaction. We also assume that these cues are detectable with the sensors that are equipped in our enhanced version of the companion robot (see figure 1).

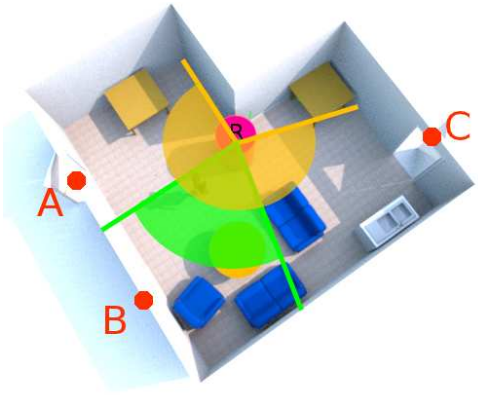


Figure 2: Home-like environment for our experimentation. The area has an L shape and 3 access doors (A, B and C). It is organized around 3 spaces: a living-room (near door B), a kitchen space (near C) an empty space (near A). The robot is place in the center of the area (purple cylinder). The view from the Kinect is depicted in green and the telemeter field of view in orange.

3.3 Steps of the interaction process

Sidner et al. in [39] proposed a model to describe the process of interaction in three steps: (1) initiation of interaction, (2) maintenance of interaction and (3) disengagement. Our work follows this approach by modeling these events as (illustrated in section A.1): (1) WILL_INTERACT, (2) INTERACT, (3) LEAVE_INTERACT. We added two more classes. The SOMEONE_AROUND event is when someone is detected in the room but with no wish of interacting with the robot. When nobody is in the room, it corresponds to the NO_ONE event.

3.4 Scenarios

The data is recorded in two different scenarios performed several times by one or several participants in a home-like environment where the Kompaï robot is present. The first one is dedicated to mono-user experiment. Only one user will be in the room at a time. The multi-users scenario addresses a more adverse condition. Three persons are already in the room and interact with each other. The idea behind this scenario is to check if we can detect starting engagement among social interaction between participant.

Each participant was given randomly one or several actions to perform in the room. As said, the room is similar to a small flat (Figure 2). It was asked to the participant to enter the room by different doors, perform some realistic actions and to go out. One action is to interact with the robot. The other actions were going across the room, walking, sitting, playing cards or pouring water from the sink.

3.4.1 Scenario 1: Passing By

In this scenario, each participant is asked to go through the room by different doors (A), (B) or (C). At this point, the given instructions did not mention the robot's presence in the living lab. Participants were not aware that they will interact later with the robot. After some crossings, the participant was invited to play the game on the robot's tablet. The Figure 3 shows the setting of this scenario.

3.4.2 Scenario 2: Playing cards together

In this second scenario, 3 or 4 persons were asked to enter the room and start playing cards in the living-room area. A telephone placed in the room was used to ask one of the participants to execute an action (interacting with the robot, or using the sink for instance). Once the participant was asked to perform a task, he could do it when he wanted to. The participants could sit wherever they wanted in

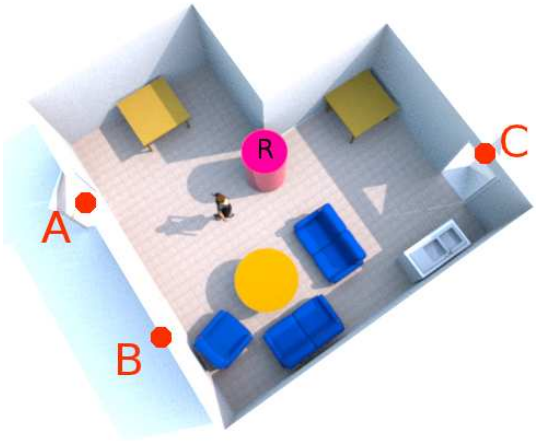


Figure 3: Scenario 1, Passing by

Figure 4: Scenario 2, Playing cards together

Figure 5: Perception from the robot point of view with a coming user toward a robot. Laser Telemeter (red lines), foot (blue spots) and pedestrian (green ellipse) information are depicted on left picture. In the middle, one can find RGB view from the Kinect with face detection (green circle). Right, the depth view with user (blue) and skeleton (yellow) tracking are drawn (note that with the first Kinect version, *Kinect for XBox*, there is a little shift between the RGB and depth views). Acoustic and other body features are computed on these data (see section 4).

the living-room area. The figure 4 shows this scenario when a new participant is entering in the room while two participants are already sitting.

4 Features Extraction

In order to characterize the engagement, features were extracted from the corpus previously introduced and then synchronized with a unique time scale. Our Kompaï robot, loaned by our partner Robosoft[1] (see figure 1), is composed of a mobile platform containing the wheel actuators, obstacle detection system, manual remote control utilities, etc. The mobile platform is topped by a tablet serving as interface with the user, a pair of microphones, a motorized web camera and a speaker device. We added a Kinect sensor to the robot. Novelty of this work is not the adjunction of the RGB-D device but the synchronous usage of information from all sensors to compute a multimodal feature set. The current version of the system is online and computes on the fly all features on the Kompaï.

[1] `http://www.robosoft.com/`

Feature extraction algorithms present a fair amount of noise in general and an interest of using multimodality is to be able to compensate one modality with another. The full feature set gathered and computed on our corpus is composed of 99 features. Then, a feature selection is made driven by social and cognitive science research on non-verbal communication cues depicted in section 2, the availability of sensors on our Kompaï robot and the performance of algorithms in our experimental conditions. We let aside important cues that might improve our results but are not usable in our context. For example, gaze direction and facial emotion recognition can not be computed; hand state and gestures are not reliable for instance. Raw (x, z, y, *confidence*) tuples for skeleton joints permit us to compute more interesting body features.

This section presents different feature extraction techniques used in our experiment. We chose to investigate a selection of 32 features including spatial information (spatial subset), body pose and video face detection (body subset), speech activity detection and sound localization (acoustic subset) in order to model the intention of engagement. We detail these subsets in the following subsections. These features are computed on several raw data channels: laser telemeter data coming from the robot, rgb video (section 4.3.3), depth view (section 4.3) and audio channels from the Kinect (see figure 5). The synchronization and labeling methods are then explained in section 4.4.

4.1 Spacial features

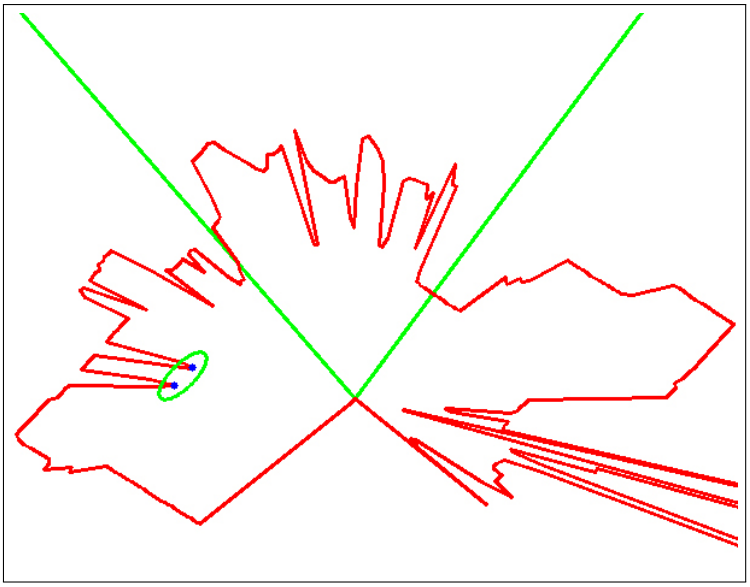


Figure 6: Spatial features: foot and pedestrians tracking using robot telemeter. Blues points are foot, ellipse represents tracked pedestrian. The green lines represent, in the robot frame, the Kinect field of view. On this figure, there is one tracked pedestrian. The pedestrian is out of the camera view.

Proxemic features are classically used to describe role, attention and interaction, and in particular to determine the intention of interaction. The tracking of the human trajectory can be done through visual based models or using laser telemeters. Telemeters provide planar information of the environment while covering a wider range angle than standard video camera with a good precision. The Kompaï robot is equipped with a single-row laser-range scanner at 20 cm above the ground. For pedestrian tracking, it is more likely that we can detect shins.

Classical proxemic features are the relative position of the individual to the robot and his speed. For a successful collaboration, the distance between the robot and the human should be optimum and the speed controlled [19]. It is important to know about distance so the person interacting with the robot does not feel uncomfortable [12]. [53] proposed a system for tracking pedestrians using multiple single row

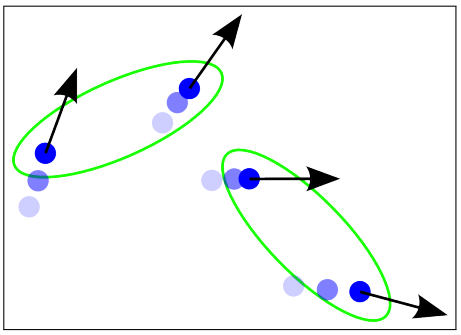


Figure 7: Feet grouping if the pair of foot satisfies a constant space between legs through time.

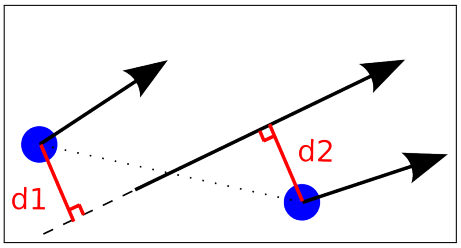


Figure 8: Space between legs is given by the sum of d1 and d2, the distances between feet and the main direction vector.

laser scanners. Their pedestrian's walking model is described and used to accurately track pedestrian feet according to their swinging phase. In our study, since a single scanner is used, foot occlusion is frequent as soon as a foot goes behind the other from the telemeter point of view. Therefore, it is difficult to predict swinging phases because of sparse data. Our human detection and tracking process is done through a feet-pairing process where a human is represented either by its both feet or one single foot when the other one is temporary hidden.

4.1.1 Feet detection

The laser sensors that equip the Kompaï robot give the distance values over 270 degrees every 80ms. An adaptive background subtraction on the telemeters values is used to detect moving objects in the room. Moving objects are candidates for being feet. The distance and the angle associated to the detected moving object are used to compute the positions x and y into the robot reference frame (see fig. 6).

4.1.2 Speed Estimation

A Kalman Filter is used on the set of moving objects detected by the laser to compute the speed and the acceleration at each frame. The Kalman filter is an iterative prediction estimation algorithm allowing to introduce measured data (in this case the position x and y of a moving object) and to estimate dynamics such as the position and speed in two dimensions $(cible\_dx, cible\_dy)$. The implementation of the Kalman Filter over the telemeters moving object data has been made using the OpenCV library. A direction vector is extracted from each foot tracker. It will be used for feet pairing in pedestrian tracking.

4.1.3 Pedestrian tracking

Grouping feet simply according to the distance between feet is not sufficient enough since pedestrians have different step length. Furthermore people standing side by side could be misidentified. Looking at the space between legs rather than between feet leads to a more robust parameter: even if the step length varies for the same pedestrian (standing, walking, running) the distance between legs along a direction vector remains constant during a natural walk because of the geometric properties of the human skeleton. Our process consists in pairing tracked feet that match a particular model, using a 2 stage filtering. First, for each frame, feet that are less than one meter apart are paired up together forming a potential pedestrian. This one meter threshold was empirically set to quickly exclude impossible pedestrians. Then, candidates are evaluated and an actual pedestrian is revealed if it satisfies that the space between legs through a short frame sequence is relatively constant (∼30 cm) as shown in figure 7.

The space between legs is computed as the sum of the distance between each foot and their projection on the main direction vector, which is the sum of two single foot direction vectors, figure 8. Pedestrian targets can be initialized as soon as they are in movement. One of the feet can then be hidden without causing the loss of padestrian's localisation, it will be paired up when the foot appears again. If both feet disappear, then the pedestrian tracker is lost and deleted.

Our tracking process is capable of tracking multiple walking pedestrians with frequent occluded feet from a single range laser telemeter.

Features Name	**Sensor**	**Frequency**
Positions ($cible\_x, cible\_y$)	laser range finder	12.5Hz
Speed ($cible\_dx, cible\_dy$)	laser range finder	12.5Hz
Distance ($cible\_dist$)	laser range finder	12.5Hz

Table 1: Features from the Space subset

Every 80 ms, we have the *Number of pedestrians* around the companion robot and for each pedestrian, we get an *id*, *cible_x* and *cible_y* position in the reference frame of the robot, his distance to the robot *cible_dist* and *cible_dx*, *cible_dy* the speed of the pedestrian in x and y axis (see Table 1 above).

4.2 Acoustic features

Pantic in [28] and [46] list some features from the audio signal that can be used to spot basic emotions such as happiness, anger, fear and sadness. It can be agreed on that some audio features such as pitch, intensity, speech rate, pitch contours, voice quality and silence are good parameters to classify the emotional state of an individual. Moreover, speech is an important information source for social glue with a companion robot [2]. Considering the recognition of the starting engagement in an interaction, only few papers in the literature use audio features in a multimodal frame. [27] proposes an engagement estimator using head pose associated to audio features in a face-to face conversational agent sitting interaction. Some articles invoke interest of sound localization in attention or focus estimation [23].

Features Name	**Sensor**	**Frequency**
Speech Activity (sad_event)	Kinect's Microphones	100Hz
Source localization ($beam, angle, confidence$)	Kinect's Microphones	8Hz

Table 2: Features from the Acoustic subset

The microphone array embedded in the Kinect sensor is a four-element linear microphone array processing acoustic echo cancellation and noise suppression. Using this audio stream, we can compute Speech Activity Detection (SAD) [44], which is indicative of the parts of the acoustic signal representing speech. The SAD labels the audio stream every 10 ms. The source localization outputs the stimulated beam (rough estimation) and the source position (more accurate angle) associated with a confidence. The frame rate of the acoustic localizer is 8Hz.

4.3 Body features

The Skeleton tracking of the Kinect sensor allows real time pose and gesture recognition. Our system outputs at depth camera frame rate the *number of skeleton*, and for each skeleton an *id* and 60 features giving x, z, y and *confidence* for each joint.

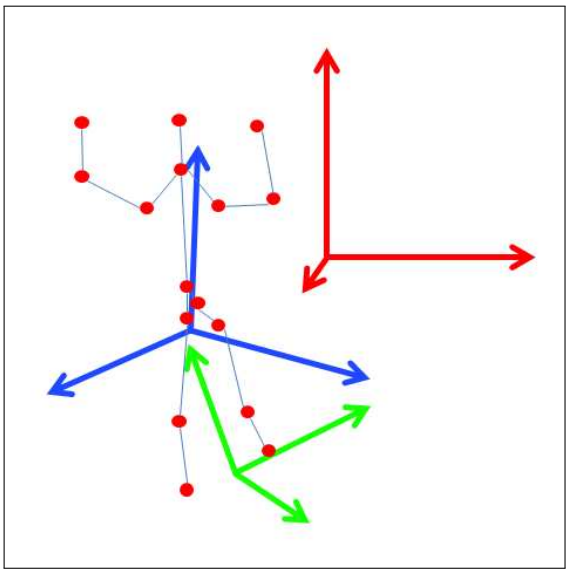


Figure 9: Stance (green) pose, hip (blue) pose and torque. Body pose features computed from the skeleton information of the Kinect sensor (positioned in red)

4.3.1 Body pose

As expressed previously, body pose features give clues on intention of interaction. These features can measure the level of engagement of a user into a task as in [35], proposing a measure of the Body Lean Angle. Psychologists have proposed many models to describe body pose metrics and their associated meaning. An overview of these metrics can be found in [24]. In [14], the authors propose a spacial model coupling attention estimation and distance metrics for a receptionist robot to infer the intentions of the human. Their results are promising, but the approach is limited to face-to-face interaction. [15] studied on spatial relationships in human robot interaction and concluded that human-human proxemic measures and social arrangement such as Hall's interpersonal distance system are not enough to achieve socially appropriate robot behavior. Psychologists such as Hall [12], Mehrabian [25] Schegloff [36] have proposed some metrics that have been used in computer assisted analysis of posture, but there is no consensus on one particular model. Posture is difficult to measure and evaluate using computer vision. Nevertheless, with new devices like the Kinect Sensor and other real-time 3D pose reconstruction systems, we are now able to evaluate the pose of a person.

The body features used in our experiments are based on Schegloff metrics presented in [24, 36] and computed from the Kinect skeletons. These features aim to depict the body pose of the individual. The accent is posed on the stance, the hips, the torso and the shoulders' position and orientation relatively to each other. We obtain 19 features containing Schegloff's metrics at depth frame rate.

4.3.2 Distance

What is interesting about body features is that they depict the orientation of the bodypart relatively to the Kinect sensor placed on the robot. A *skeleton distance* associated to the skeleton position is computed using the average z-value of several joints of the skeleton.

4.3.3 Face detection

In terms of affect & emotion detection and speech recognition, a lot of studies have published results with a combination of face and audio features [33, 40, 7, 17]. Within engagement, the orientation of the head and the gaze seem to be crucial. As shown in [32], a speaker can be detected more easily with the combination of different features as a mouth sensor. Face detection is already a first cue of interaction,

and orientation of the face toward the robot is a reliable sign of attention. Gaze tracking can give a better estimation of user' attention, but performance in uncontrolled real-life condition (small faces in video stream or untrained gaze angle for instance) are not good enough.

From the video extracted from the RGB stream, we propose to focus on face detection. We use a trained machine learning system using Haarcascasdes method. The training is provided by the OpenCV library [50]. The gathered features are the position of face(s) in the pixel reference frame, the {*0,0*} point is the center of the image. For each detected face, we have *x*, *y* and *face size*.

Features Name[2]	**Sensor**	**Frequency**
Stance ($*Pose\_x, *Pose\_y, *Pose\_z$, $*Pose\_rot$) for feet, hips, torso and shoulders	Kinect's Skeleton	30Hz
Relative torque angle ($*Torque$) for hips, torso, shoulders	Kinect's Skeleton	30Hz
Skeleton distance (skl_dist)	Kinect's Skeleton	30Hz
Face ($face\_x, face\_y, face\_size$)	RGB stream	30Hz

Table 3: Features from the Body pose subset.

4.4 Fusion, Synchronization and labeling of features

At this point, we have a selection of 32 features: pedestrian information (*x*, *y*, *speed_x*, *speed_y* and *distance* to robot), Schegloff metrics (computed from the skeleton see 4.3.1), face detection, speech activity detection and sound localization. This section details how we deal with sparse features, features synchronization and corpus labeling.

4.4.1 Sparse features

Multimodality has one major drawback. Space coverage is not the same for all sensors: the Kinect has a 60 degrees field of view, the laser telemeter 270 degrees, etc. Moreover, we do not have every feature all the time. Whereas video, depth and laser telemeter data, face detection, sound classification, skeleton and pedestrian tracking are not available all the time. One way to cope with these sparse data is to train several classifiers with all possible combinations for available features and to select the adequate one at the right moment. The problem with this approach lies in reducing the amount of data for training for each subtype of classifiers. Another way to solve this problem is to use specific neutral values for unavailable features. For example, when there is no pedestrian, we can set all pedestrian features (position, speed and acceleration) to 0. This set of values is considered neutral as it is impossible to find them in observed data. In these experiments, as we did not have enough data to train each subtype of classifiers, we chose the second method.

4.4.2 Features synchronization

We needed to synchronize the monitored data from the different modalities. Data collected through the Kinect sensor such as the skeletons' positions, the video and the depth are tagged with a time relative to the Kinect sensor's initialization. The laser data are labeled with an absolute time stamp thanks to the real-time micro-controller of the Kompaï robot. The telemeters' input is the steadiest one at a fixed 80 ms period, hence it is used as synchronization frame rate at 12.5HZ. The short time delay between frames prevents us to interpolate and allows to elicit the last value of each feature as the current value.

[2]Suffixes are presented with a "∗" character. For example, we compute *shoulderPose_x*, *shoulderPose_y*, *shoulderPose_z*, *shoulderPose_rot* and *shoulderTorque*.

4.4.3 Corpus labeling

The labeling of the dataset with the 5 classes (WILL_INTERACT, INTERACT, LEAVE_INTERACT, NO_ONE, SOMEONE_AROUND) is semi-automatic. The timestamped notes of the experimenter contain start time and end time of all events (participant p is entering the room, a specific action was asked, etc.). This annotation serves as first segmentation input. The first labels are then made automatically using both the tablet touching information and the available features. The INTERACT time-interval is labeled from first touch of the tablet until the last click. A WILL_INTERACT event starts when someone is coming to the Kompaï robot before an INTERACT event. If someone is moving or sitting, and after decides to come to the robot, only the direct path to the robot is labeled as WILL_INTERACT. The LEAVE_INTERACT labeling is made like the WILL_INTERACT even and corresponds to all direct paths after leaving the interaction. The NO_ONE and SOMEONE_AROUND events, correspond to the rest of the time, respectively, there is no one in the room and when a person is present, but there will be no interaction.

All labels have been reviewed by a human expert looking at the video recordings. No problems were reported as a result.

4.5 The dataset in numbers

In our dataset, each frame corresponds to 80 ms, provides a full feature set and has a unique label. The total number of recorded frames is 158200[3]. The number of synchronous frames for each event is not equal. The Figure 10 shows the data distribution of each event.

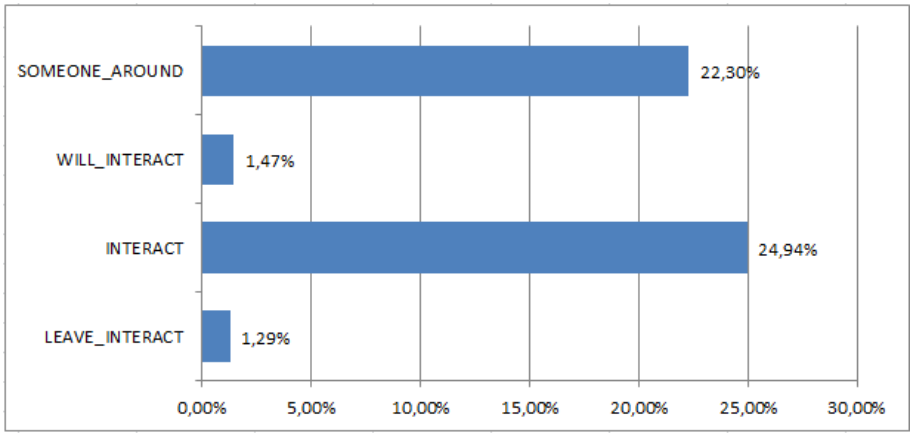


Figure 10: Percentage of each class in the dataset.

In real life, individuals do not express social signals the same way. A certain variability was introduced in the pool of 19 participants. They were from 20 to 35 years old, almost 50% male/50% female, students, administrative assistants and researchers. Voices, clothing styles (colors, trousers or skirts, etc.) and statures vary to challenge perception algorithms. 15 participants did 1 or 2 interactions from 2 to 10 minutes according to their will, 9 were recorded both in mono-user and in multi-users scenarios.

In total, the corpus includes 29 interactions with the robot, made by 15 different participants. The total size of the uncompressed data set is around 300 GB. One can find samples of the corpus in A.1.

5 Multimodal detection of engagement

In this work, we first choose to test all the modalities that can help us to detect intention of interaction. Then, a selection can be made among the most relevant multi-modal features (section 5.2.3). The evaluation focuses on comparing the detection of the intention of interaction by using multimodality

[3]The total recording time is 3:30:56.

Telemeters condition

Class	Precision	Recall
NO_ONE	0,95	1,00
WILL_INTERACT	0,91	0,77
INTERACT	0,77	0,96
LEAVE_INTERACT	0,00	0,00
SOMEONE_AROUND	0,75	0,35

Multimodal condition

Class	Precision	Recall
NO_ONE	0,95	1,00
WILL_INTERACT	0,90	0,87
INTERACT	0,84	0,95
LEAVE_INTERACT	0,21	0,01
SOMEONE_AROUND	0,76	0,41

Table 4: Results for Neural-Network 5-classes classification using Weka. Left table presents results for the telemeter condition. Right table for the multimodal condition.

versus simple spatial information. We want to confirm that these state-of-the-art approaches based on spatial features only are not enough in home-environment with furniture. The Scipy library through Sklearn [38] and the Weka toolbox [49] were used for the classification. The techniques used for the classification are the Multi-class Support Vector Machine (SVM) from Sklearn and the Artificial Neural Network (ANN) technique from Weka.

5.1 Prepare the dataset for the Classification (K-Cross Folding)

In order to train a model and to test it afterwards, the dataset needs to be split in a training set and a test set. A way to randomize this splitting is the k-cross folding process. In this method, the dataset is partitioned in k subset. One subset is kept for testing and the $k-1$ others are used for training the model. This splitting process is repeated k times so that each subset is used once for testing against others subsets. K-cross validation allows to ensure that the splitting is quite random. Since the events (interaction phases) are not equally probable and temporally related, we used a *stratified* k-fold-cross validation that keeps the same proportion of the different classes in the splitting process.

For our experiment, using $k = 10$, the train and test sets are composed respectively of 140292 and 15587 frames[4].

5.2 First classification experiment

We chose to use two kinds of classification in this experiment. Even if many other techniques could have been applied, we decided to focus on Neural Network and Support Vector Machine (SVM) (sections 5.2.1 and 5.2.2). For this two techniques, we built and tested two classifiers one for the multimodal condition (including 32 features) and one for laser telemeter only condition (a subset of the multimodal dataset including spatial information only), see [3]. In 5.2.3, we try to determinate if some features are more relevant for our task.

5.2.1 Neural Network

The Artificial Neural Network is a graphical layered model commonly used to infer model from observation. In our case, we suppose that our features set can characterize the starting of engagement. ANN is a good classifier to build prospective detection especially with large features vector. The test results of the ANN classification are presented in left table in 4 for the telemeter, and the right table for the multimodal dataset. Notably, one can see that that LEAVE_INTERACT was not classified in the telemeter condition. The INTERACT precision increased in multimodal condition combined with a small loss in recall. Concerning the WILL_INTERACT class, the system returns more relevant events using multimodality (higher recall score) even if its precision decreased. In multimodal condition, the precision is improved for most of classes. The Neural Network classifier gives always better recall rate in this condition.

[4]Note that the *stratified* splitting process let aside 2321 frames.

Telemeters condition

Class	Precision	Recall
NO_ONE	0,68	1,00
WILL_INTERACT	0,80	0,68
INTERACT	0,00	0,00
LEAVE_INTERACT	0,00	0,00
SOMEONE_AROUND	0,76	0,01

Multimodal condition

Class	Precision	Recall
NO_ONE	0,92	0,88
WILL_INTERACT	0,92	0,71
INTERACT	0,54	0,77
LEAVE_INTERACT	0,04	0,10
SOMEONE_AROUNDd	0,52	0,29

Table 5: Results for SVM 5-classes classification using Sklearn. Left table presents results for the telemeter condition. Right table for the multimodal condition.

For the intention of engagement detection, in a practical point of view, the accent has to be put on the good performance in term of recall associated to a low false-positive rate. Using Neural Network in multimodal condition seems to fulfill this requirement.

5.2.2 Multi-Class Support Vector Machine

The results of the 5-classes classification using support vector machine for the multimodal condition are presented on right table in 5, on left table for the telemeter condition. Analyzing these tables, one can see that the False-Positive rate is higher in the telemeter condition. The INTERACT class is not classified at all. The precision and recall scores for WILL_INTERACT class are improved by the multimodality. The aim of our work was especially to decrease the rate of misclassifying an event as WILL_INTERACT, hence the system has fewer chances to predict an interaction when there is one user with no intention of interaction. In the case of an SVM classifier, multimodality is thus interesting for this purpose.

Scores for INTERACT, LEAVE_INTERACT and SOMEONE_AROUND classes are of interest. In multimodal condition, the SOMEONE_AROUND scores drop while INTERACT and LEAVE_INTERACT are better classified. This fact is a first clue of the closeness of our classes in the feature space. The section 5.2.4 will discuss this topic.

5.2.3 Minimum Redundancy Maximum Relevance experiment

A dimensional reduction of the features space was made using the Principal Component Analysis (PCA) and the Linear Discriminant Analysis (LDA) using the Sklearn tool-kit. The results were not conclusive, the dimensionality reduction gave strictly the same performance during the classification where we were expecting an improvement.

The Minimum Redundancy Maximum Relevance [13] (MRMR) technique was performed in order to highlight the best features for our detection system. This dimensionality reduction technique has the advantage of giving the more relevant features instead of building new features from the observed ones. Using mutual information, correlation and t-test/F-test metrics, the MRMR algorithm selects a feature subset maximizing dissimilarity of features and statistical characterization of the classification. It could allow eventually to discard less relevant features in order to optimize the detection of engagement process.

In order to evaluate the relevant features for the multimodal detection of intention of interaction, we used a MRMR dimensionality reduction from a vector of 32 features before performing an SVM learning. The Figure 11 shows the feature reduction's impact on the precision. The precision drops at the 6-features reduction. From the 32-features till 7-features along the feature space reduction, the precision remains pretty stable. These results confirm that there are many redundancies in the 32-feature space. Some of these features seems to be fundamental for a better detection with a higher precision than the telemeters' one. Equivalent conclusions can be made on the Figure 12 regarding the recall performances.

The first remark on these results is that the seven highest rated features are coming from heterogeneous modalities. The *face size* and *face x* are respectively the relative size and position of the face in the Kinect view. The *beam* and the *angle* are the sound localization features from the microphone array. The telemeter information are considered as relevant, with the high selection rate of the speed *speed_x*

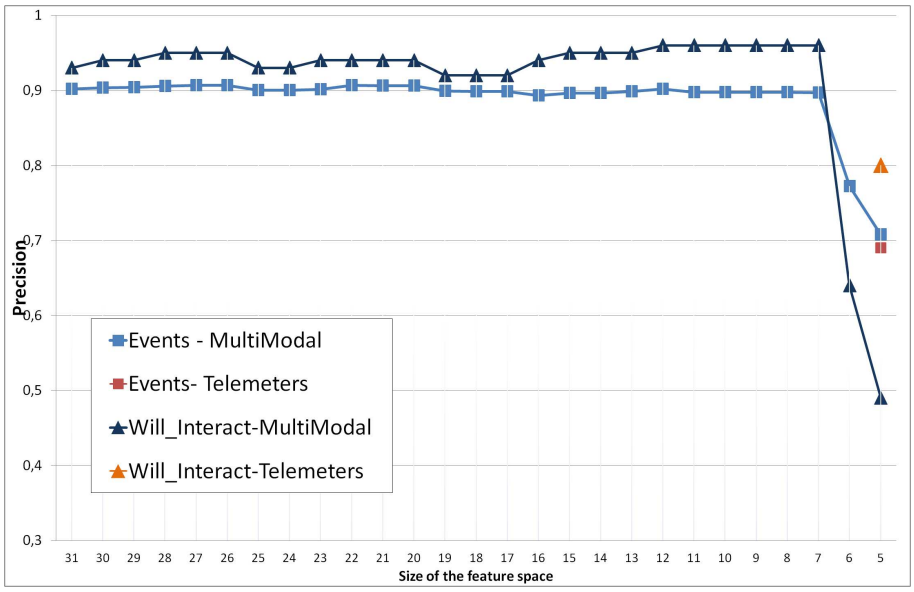


Figure 11: Precision evolution with the decreasing number of multimodal features in comparison with the telemeter condition for all events and for the WILL_INTERACT event

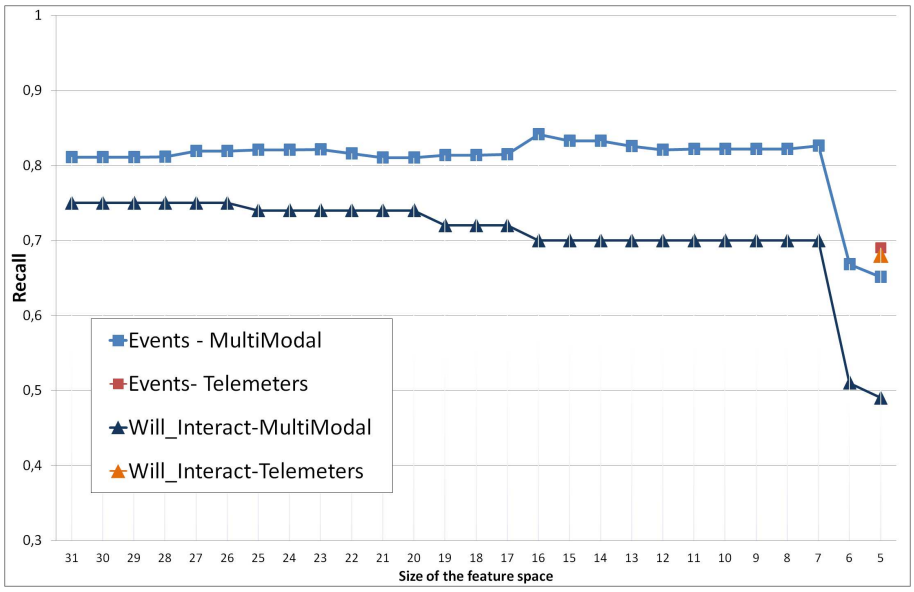


Figure 12: Recall evolution with the decreasing number of multimodal features in comparison with the telemeter condition for all events and for the WILL_INTERACT event

and y position. The fact that these spatial features are selected among the most relevant ones in our multimodal set is not surprising. They were part of previous state-of-the-art researches. Last, the *shoulder pose rotation* corresponds to the relative orientation of the shoulder in the body, and is extracted from the skeleton information.

5.2.4 Discussion about first experiment

The previously presented results support our assumptions about multimodal recognition of intention of interaction. The body pose, especially the shoulders orientation were shown to be relevant for intentionality detection. Spatial information coming from the audio and telemeter streams are also important. The position and size of detected faces in the video confirms that facing the robot is a sign of intentionality of engagement.

These evaluations were conducted using an a priori selection of 32 features. This selection was inspired by our literature searches in human-robot interaction, social sciences and cognitive science fields. Results are improved, but can we conclude that our results are generic enough? For instance, we replaced all skeleton information by Schegloff metrics. Even if results validate our hypothesis, we need to check if we do not have interesting information in the left aside 67 other features.

From the results[5], we see that LEAVE_INTERACT is never well classified in the telemeter condition. In the multimodal condition, the precision and recall scores get slightly improved. Anyway, LEAVE_INTERACT is most of time confused with SOMEONE_AROUND. Several explanations may enlighten this result. When someone is interacting with the Kompaï, he is close to the robot. We have spatial features computed from laser telemeter but no information about his body from the Kinect (see table 7). A more intuitive point can be that less social signals are expressed when leaving interaction. For closeness reasons, the INTERACT class presents also low classification results. We do not actually need to detect it: intention of engagement is a prior state to interaction.

In preliminary conclusion, we can say that our primary hypothesis is validated: multimodality can improve engagement detection on a companion robot with embedded sensors. Nevertheless, other experiments must be conducted with a three classes approach (NO_ONE, SOMEONE_AROUND and WILL_INTERACT) and all the available features.

5.3 Second experiment

In this second experiment, we will tackle our classification task in regards to the lessons learned in 5.2.4.

5.3.1 New 3 labels corpus

Validation of our 3 classes We need to validate our hypothesis about the confusion of the LEAVE_INTERACT and SOMEONE_AROUND classes. We conducted clustering experiments using the k-means algorithm. We wanted to check if it is difficult to separate these 2 classes in the features space. K-means is an algorithm that produces the best clustering knowing the number of wanted clusters. We ran K-means from 2 to 1500 clusters with our 7-features set and checked the distribution of each feature vectors in these clusters. In all clusterings done, there is no significant separation between our 2 classes, i.e. the feature vectors of each class are equitably distributed among clusters. We did clustering with every feature set up to our 32 features and distribution remained diffused. This result also corroborates that it is likely to say that either people do not express strong social signals when they leave interaction with the robot or that in our hardware setup, we cannot compute them.

New labeling We modified our labeling using this time only our 3 classes (see discussion 5.2.4). All frames from the INTERACT class were removed and all instances of LEAVE_INTERACT were replaced by SOMEONE_AROUND. From the remaining 124282 frames, using k-cross folding (see 5.1) with $k = 10$, we

[5]As we did a *stratified* k-fold-cross validation, we have many confusion matrices. Presenting one will not correspond to the k-fold-cross validation result (table 4 and 5), showing all is not possible.

Class	Precision	Recall
WILL_INTERACT	0,95	0,93
NO_ONE	0,93	1,00
SOMEONE_AROUND	0.99	0,84

Table 6: Results of multimodal SVM 5-class classification using Sklearn with our 3 classes.

computed train and test sets respectively of 111854 and 12428 frames. Repartition of each class is given in the figure 13.

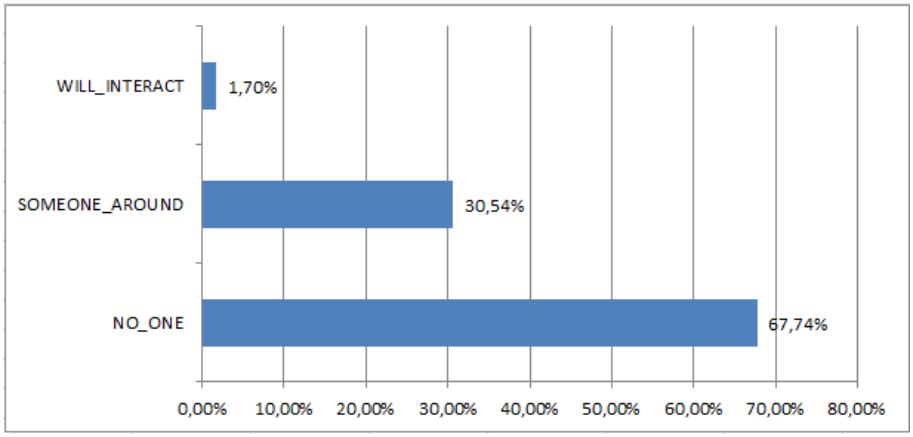


Figure 13: Percentage of frames per event with 3-classes labeling

5.3.2 Classification

Using the new dataset, we re-performed the classification process using SVM. Results are shown in the table 6. We can see that we have an overall improvement in our results on the 3 remaining classes, mainly on the recall score. We increased both scores on the SOMEONE_AROUND class.

Training on this data was more successful, but new experimentation in other conditions, in other places, with different lighting conditions, with more participants has to be done in order to conclude definitely on these results.

5.3.3 Feature selection among all available features

We re-ran our experiment using the MRMR technique to select in our total 99 features set the most relevant ones. Results in this case differ from the first experiment. Some intuitive features, like the *facing coefficient* and the *skeleton distance*, were selected. Surprising features appear among the more relevant ones. Indeed, the most important feature is the *right ankle x* position. This fact is peculiar when one knows that more than 70% of the frames have no information about a skeleton. Moreover, many skeletons have noisy feet information due to the position of the robot in the living lab. During our recordings, participants passed very closed to the robot, as we wanted them to do (see SOMEONE_AROUND example in A.1). In this case, the mounted Kinect did not manage to compute confident 3D feet position.

The MRMR technique is not efficient on this task. For a 99 features space, we may not have enough data to determine reliable metrics (mutual information, correlation, t-test/F-test...) used by the algorithm. Nevertheless, this experiment confirms that selecting human readable features inspired from social and cognitive sciences could be an alternative method for feature space reduction.

6 Conclusion

Psychologists working on the acceptance of a robot by the elderly and people with disabilities have pointed out the need for more natural and acceptable interactions with the companion robot in the home environment. A starting engagement is the first step of interaction. It corresponds to the phase preceding the actual interaction, when the user implicitly signifies his wish to interact. For a companion robot, the skill to detect engagement , and thus, to anticipate human will is a key feature to make it socially acceptable.

Our goal was to evaluate and measure humans' cues for engagement in an interaction with a robot. They were classically detected using the position and the speed of the user. We have shown the limits, in term of recall, in performance of this technique, confronting it with realistic scenarios of engagement towards a robot in a home environment. Indeed, the proximity of the user with the companion robot is not a sufficient criteria when predicting the engagement. Spatial based detection of intention of interaction used in previous approaches gave good results in lab environment. However, the congestion of home-environment leads to situations where humans pass close to the robot without the will of interaction and where spatial based detection gives false positive responses.

Having built realistic scenarios involving the interaction of participants with the Kompaï robot, we have collected sensory data of various engagement sequences. Several features were computed over multiple modalities. From the video, we detected the size and position of the face in the image. The skeleton data gave us clues to compute body poses. The audio was used for the sound localization and for the speech activity recognition. Telemeters gave us an estimation of the position and the speed of the pedestrians.

A cross-fold validation allowed us to segment our dataset into training and testing sets. These subsets where used by two different classifiers, a Neural Networks and a Support Vector Machine. These classifiers, trained on multimodal and telemeters features set, gave better performances for the multimodal condition. This fact showed that spatial and speed features used in related works are not enough in a home environment. Multimodality improved the recall of the engagement detection significantly, which was the hypothesis of this research.

6.1 Key points

Transposing social and cognitive sciences results and using human readable features can be an alternative approach for feature selection. Using this methodology, we enhanced spatial information with the selected body related and acoustic features and get better detection scores, notably in terms of recall. As far as we know, we validated experimentally for the first time, that shoulder pose rotation, a metric from Schegloff's research in Sociology, is of importance to the detection of intention of engagement.

The high correlation between the features also made the classification more difficult. On one hand, the Minimum Redundancy Maximum Relevance (MRMR) feature selection algorithm helped us get to a set of measurable multimodal features sufficient to detect intention of interaction towards a robot based on human selected features set. On the other hand, trying to deals with all 99 available features to elicit the more relevant ones fails. New experiments need to be conducted with more variability in order to improve the scores of the intention detector using the reduced feature space.

Current work about high-level features fusion and analysis is ongoing. We want to remove some artifacts that penalize feature selection and classification algorithms. For instance, we are combining face detection, skeleton tracking and depth data to improve feature association. These features now belong to the same user, i.e. when we compute Schegloff and face features they refer to the tracked pedestrian. As far as we can say from our preliminary experimentation, doing so, results are improved but not significantly for the multi-user scenario.

6.2 Impact of this research

With this research we provide deeper knowledge about meaningful features that can facilitate robot's social abilities. We computed new features inspired from the literature in social sciences, notably the Schegloff's features. A ranked list of 32 most relevant features for our starting engagement can be found

in A.2. This presented work gives design guidelines and praise multi-modal sensor embedding on robot to facilitate human-robot interaction.

The feature selection process depicted in this article was inspired from research on genome [26]. Principal Component Analysis (PCA) and the Linear Discriminant Analysis (LDA) did not provide significant classification improvement. The MRMR algorithm however showed that selecting features can be more relevant than combining them. The direct ranking provides information about meaningful features for a classification task, for example. The feature selection method can be applied in many contexts where space reduction is of interest.

This work is one more step into the use of multimodality for social signal processing applied to human-robot interaction. Multimodality can be very useful in decoding and recognising affect signals and hence in improving the human-robot relationship. With more and more powerful embedded systems deployed on robots, we can expect such multimodal detection to be generalized in real-time and to allow robots to predict intentions of the users. The prediction of the engagement is a first step towards a smoother and socially acceptable human-robot interaction.

7 Acknowledgments

The author would like to thank the Robosoft company for loaning us the Kompaï robot, Inria and French Ministry of Education and Research for their support.

8 References

References

[1] Michael Argyle. *Bodily Communication*. Methuen & Co Ltd, 1975.

[2] V. Aubergé, Y. Sasa, T. Robert, N. Bonnefond, and B. Meillon. Emoz: a wizard of oz for emerging the socio-affective glue with a non humanoid companion robot. *Workshop on Affective Social Speech Signal (WASSS2013), sattelite event of Interspeech 2013*, aug 2013.

[3] Wafa Benkaouar and Dominique Vaufreydaz. Multi-Sensors Engagement Detection with a Robot Companion in a Home Environment. pages 45–52, Vilamoura, Algarve, Portugal, October 2012.

[4] Peter E. Bull. *Posture and Gesture*. International Series in Expreimental Social Psychology. Pergamon Press, 1987.

[5] Jennifer L Burke, Robin R Murphy, Erika Rogers, Vladimir J Lumelsky, and Jean Scholtz. Final report for the darpa/nsf interdisciplinary study on human-robot interaction. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 34(2):103–112, 2004.

[6] Ginevra Castellano, André Pereira, Iolanda Leite, Ana Paiva, and Peter W McOwan. Detecting user engagement with a robot companion using task and social interaction-based features. pages 119–126. ACM, 2009.

[7] Tsuhan Chen and Ram R Rao. Audio-visual integration in multimodal communication. *Proceedings of the IEEE*, 86(5):837–852, 1998.

[8] Juan Fasola and Maja J. Matari. Socially assistive robot exercise coach: Motivating older adults to engage in physical exercise. In Jaydev P. Desai, Gregory Dudek, Oussama Khatib, and Vijay Kumar, editors, *Experimental Robotics*, volume 88 of *Springer Tracts in Advanced Robotics*, pages 463–479. Springer International Publishing, 2013.

[9] David Feil-Seifer and Maja J Mataric. Defining socially assistive robotics. In *Rehabilitation Robotics, 2005. ICORR 2005. 9th International Conference on*, pages 465–468. IEEE, 2005.

[10] David Fischinger, Peter Einramhof, Walter Wohlkinger, K Papoutsakis, Peter Mayer, Paul Panek, T Koertner, S Hofmann, Antonis Argyros, Markus Vincze, et al. Hobbit-the mutual care robot. In *Workshop-Proc. of ASROB*, 2013.

[11] James Glasnapp and Oliver Brdiczka. A Human-Centered Model for Detecting Technology Engagement. *Human-Computer Interaction*, LNCS 5612:621–630, 2009.

[12] Edward Twitchell Hall and Edward T Hall. *The hidden dimension*, volume 1990. Anchor Books New York, 1969.

[13] Peng Hanchuan, Long Fuhui, and Ding Chris. Feature Selection Based on Mutual Information : Criteria of Max-Dependency, Max-Relevance and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[14] Patrick Holthaus, Karola Pitsch, and Sven Wachsmuth. How Can I Help? *International Journal of Social Robotics*, 3(4):383–393, September 2011.

[15] H. Huettenrauch, K. Severinson Eklundh, A. Green, and E.A. Topp. Investigating spatial relationships in human-robot interaction. pages 5052–5059, 2006.

[16] Steven H. Kaminski. *Communication Models*. 2002.

[17] Kostas Karpouzis, Amaryllis Raouzaiou, Athanasios Drosopoulos, Spiros Ioannou, Themis Balomenos, Nicolas Tsapatsoulis, and Stefanos Kollias. Facial expression and gesture analysis for emotionally-rich man-machine interaction. *3D modeling and animation: synthesis and analysis techniques*, pages 175–200, 2004.

[18] Heather Knight. Eight lessons learned about non-verbal interactions through robot theater. *Social Robotics*, pages 42–51, 2011.

[19] Seongyong Koo, Dong-soo Kwon, and Extended Kalman Filtering. Recognizing Human Intentional Actions from the Relative Movements between Human and Robot. *Nonlinear Dynamics*, pages 939–944, 2009.

[20] Nicole C Krämer, Sabrina Eimler, Astrid von der Pütten, and Sabine Payr. Theory of companions: what can theoretical models contribute to applications and understanding of human-robot interaction? *Applied Artificial Intelligence*, 25(6):474–502, 2011.

[21] Peter Krauthausen and Uwe D Hanebeck. Situation-specific intention recognition for human-robot cooperation. pages 418–425. Springer, 2010.

[22] Marwa Mahmoud, Tadas Baltrušaitis, Peter Robinson, and Laurel D Riek. 3d corpus of spontaneous complex mental states. *Affective computing and intelligent interaction*, pages 205–214, 2011.

[23] Jerome Maisonnasse, Nicolas Gourier, Oliver Brdiczka, Patrick Reignier, and James Crowley. Detecting privacy in attention aware system. IET, IET, jul 2006.

[24] Ross Mead, Amin Atrash, and Maja J Matarić. Proxemic Feature Recognition for Interactive Robots : Automating Metrics from the Social Sciences. pages 52–61, 2011.

[25] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.

[26] Piyushkumar A Mundra and Jagath C Rajapakse. Svm-rfe with mrmr filter for gene selection. *IEEE Transactions on NanoBioscience*, 9(1):31–37, 2010.

[27] Ryota Ooko, Ryo Ishii, and Yukiko I Nakano. Estimating a user?s conversational engagement based on head pose information. *Intelligent Virtual Agents*, pages 262–268, 2011.

[28] Maja Pantic and Leon JM Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[29] Sylvie Pesty and Dominique Duhaut. Artificial companion: building a impacting relation. In *Robotics and Biomimetics (ROBIO), 2011 IEEE International Conference on*, pages 2902–2907. IEEE, 2011.

[30] Rosalind W. Picard. *Affective Computing*. International Series in Experimental Social Psychology. The MIT Press, 2005.

[31] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 18(10):10762–10774, 2010.

[32] James M Rehg, Kevin P Murphy, and Paul W Fieguth. Vision-Based Speaker Detection Using Bayesian Networks. *Pattern Recognition*, (Cvpr 99):110–116, 1999.

[33] Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace L. Sidner. Recognizing engagement in human-robot interaction. *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 375–382, March 2010.

[34] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W McOwan, and Ana Paiva. Automatic analysis of affective postures and body motion to detect engagement with a game companion. pages 305–311. IEEE, 2011.

[35] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W McOwan, and Ana Paiva. Automatic analysis of affective postures and body motion to detect engagement with a game companion. *6th ACM/IEEE International Conference on Human-Robot Interaction (HRI2011)*, pages 305–311, 2011.

[36] Emanuel A Schegloff. Body Torque. *Social Research*, 65(3):535–596, 1998.

[37] C. Schroeter, S. Mueller, M. Volkhardt, E. Einhorn, C. Huijnen, H. van den Heuvel, A van Berlo, A Bley, and H.-M. Gross. Realization and user evaluation of a companion robot for people with mild cognitive impairments. pages 1153–1159, May 2013.

[38] Scikit-learn. `http://scikit-learn.org/stable/`.

[39] Candace L Sidner, Christopher Lee, and Neal Lesh. Engagement rules for human-robot collaborative interactions. *IEEE International Conference On Systems Man And Cybernetics*, 4:3957–3962, 2003.

[40] De Silva and Palmerston North. Audiovisual Recognition. pages 649–654, 2004.

[41] SSPNet. Social signal porcessing network `http://sspnet.eu/2010/04/semaine-corpus/`, 2010.

[42] SSPNet. Social signal porcessing network `http://sspnet.eu/`, 2012.

[43] Karim A. Tahboub. Intelligent Human-Machine Interaction Based on Dynamic Bayesian Networks Probabilistic Intention Recognition. *Journal of Intelligent and Robotic Systems*, 45(1):31–52, March 2006.

[44] Dominique Vaufreydaz, Rémi Emonet, Patrick Reignier, and Reignier Patrick Vaufreydaz Dominique, Emonet Rémi. A Lightweight Speech Detection System for Perceptive Environments. *3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, Washington : United States, 2006*, 2006.

[45] David Vernon, Claes Hofsten, and Luciano Fadiga. *A Roadmap for Cognitive Development in Humanoid Robots*, volume 11 of *Cognitive Systems Monographs*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[46] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[47] Michael L Walters, Kerstin Dautenhahn, René Te Boekhorst, Kheng Lee Koay, Dag Sverre Syrdal, and Chrystopher L Nehaniv. An empirical framework for human-robot proxemics. *Procs of New Frontiers in Human-Robot Interaction (2009)*, 2009.

[48] Lin Wang, Pei-Luen Patrick Rau, Vanessa Evers, Benjamin Krisper Robinson, and Pamela Hinds. When in Rome: the role of culture & context in adherence to robot recommendations. *ACM*, pages 359–366, March 2010.

[49] Weka3. Data mining software and toolkit in java `http://www.cs.waikato.ac.nz/ml/weka/`.

[50] WillowGarage. Open cv library `http://opencv.org/`.

[51] C. Wolf, J. Mille, E. Lombardi, O. Celiktutan, M. Jiu, M. Baccouche, E Dellandrea, C.-E. Bichot, C. Garcia, and B. Sankur. The liris human activities dataset and the icpr 2012 human activities recognition and localization competition. March 2012.

[52] Zhihong Zeng, Maja Pantic, and Thomas S Huang. Emotion recognition based on multimodal information. *Affective Information Processing*, pages 241–265, 2009.

[53] Huijing Zhao and R. Shibasaki. A novel system for tracking pedestrians using multiple single-row laser-range scanners. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 35(2):283–291, 2005.

APPENDIX

A.1 Samples of the corpus

The table 7 shows samples of the dataset that we propose for 4 of the events WILL_INTERACT, INTERACT, LEAVE_INTERACT and SOMEONE_AROUND. As one can see the event LEAVE_INTERACT can be quite confusing with the INTERACT state.

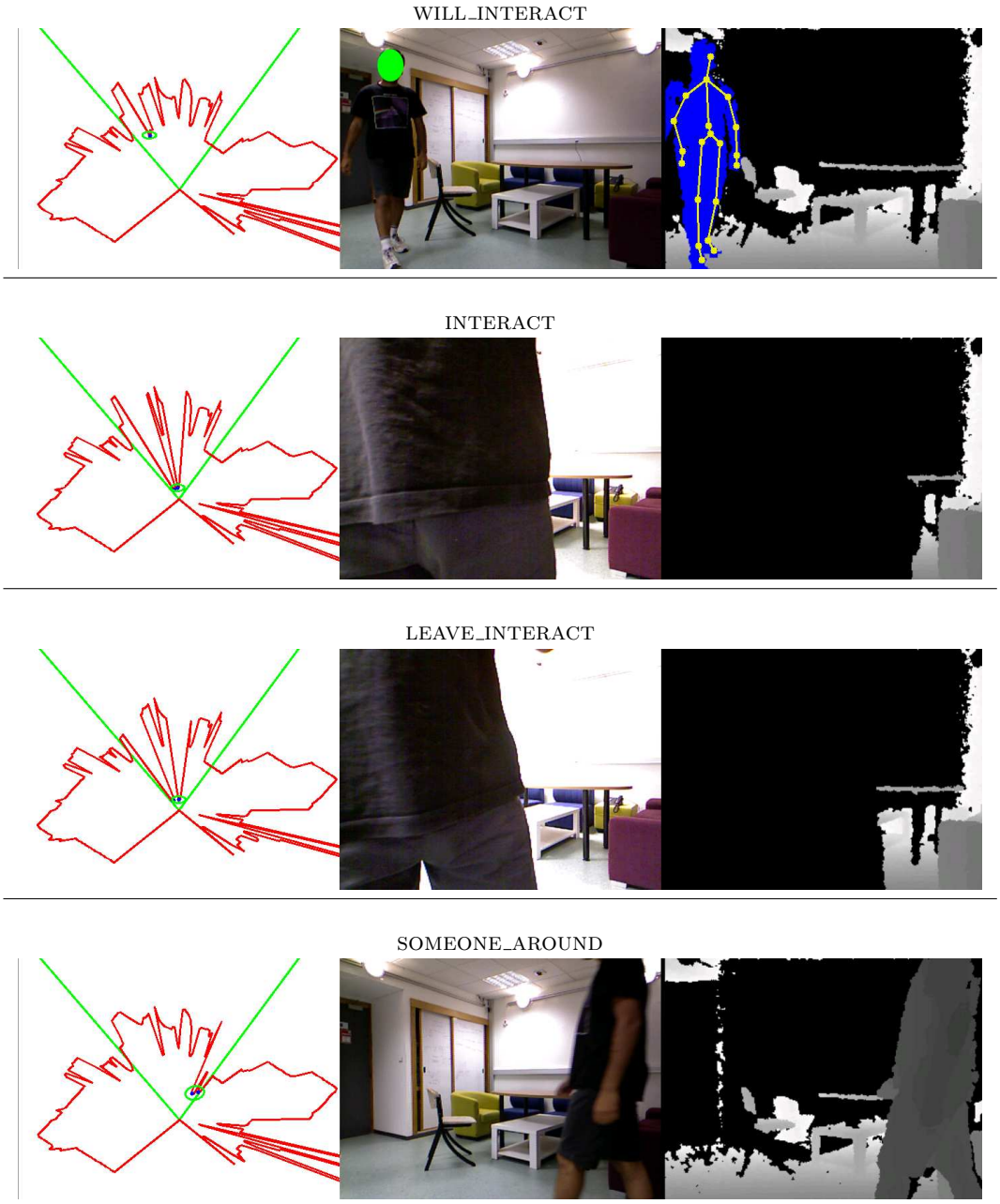


Table 7: Samples of data recorded with the Kompaï equipped with a Kinect sensor and a laser telemeter. For each view, one can find spatial information at left, rgb camera view with face detection in the middle and depth camera with people and skeleton detection.

A.2 Ordered list of 32 most relevant features

Using the Minimum Redundancy Maximum Relevance (MRMR) algorithm, we ranked 32 features (see section 5.2.3). In the following table, the reader must note that *cible_* prefix identifies pedestrian related features. Other body features are related to Schegloff work (see section 4.3.1). Features not listed in table 8 are depicted in section 4.

Order	Short name	Unit	Description
1	shoulderPose_rot	*radian*	Rotation of the shoulder
2	cible_dx	$meter.seconde^{-1}$	Speed in x of pedestrian
3	cible_y	*meter*	position on Y axis of pedestrian
4	face_size	*pixel*	Size of face in the RGB frame
5	face_x	*pixel*	Lateral position of the face
6	beam	*radian*	Activated audio beam
7	angle	*radian*	Audio localization (azimut)
8	hipPose_x	*meter*	Hip X attribute
9	hipPose_y	*meter*	Hip Y attribute
10	hipPose_rot	*radian*	Hip rotation angle
11	face_y	*pixel*	Height of the face
12	sad_event	Speech/Not speech	Speech activity detection tags
13	stancePose_rot	*radian*	Stance rotation
14	torsoPose_rot	*radian*	Torso rotation
15	shoulderTorque	*radian*	Shoulder torque
16	shoulderPose_y	*meter*	Shoulder Y attribute
17	source_confidence	$[0; 1]$	Audio localization confidence
18	torsoTorque	*radian*	Torso torque
19	stancePose_z	*meter*	Stance Z attribute
20	skl_dist	*meter*	Distance of the tracked skeleton
21	cible_x	*meter*	Position on X-axis of pedestrian
22	hipTorque	*radian*	Hip torque
23	torsoPose_y	*meter*	Torso Y attribute
24	torsoPose_x	*meter*	Torso X attribute
25	shoulderPose_x	*meter*	Shoulder X attribute
26	stancePose_x	*meter*	Stance X attribute
27	cible_dy	$meter.seconde^{-1}$	Speed on Y-axis of pedestrian
28	cible_dist	*meter*	Distance of the pedestrian
29	torsoPose_z	*meter*	Torso Z attribute
30	stancePose_y	*meter*	Stance Y attribute
31	hipPose_z	*meter*	Hip Z attribute
32	shoulderPose_z	*meter*	Shoulder Z attribute

Table 8: MRMR algorithm output on the 32 features set.

Impact of Head Motion on the Assistive Robot Expressiveness - Evaluation with Elderly Persons

Fabien Badeig[1,2], Pierre Wargnier[3,4], Maribel Pino[3,5], Philippe De Oliveira Lopes[3,5], Emeric Grange[1,2], James L. Crowley[1,2], Anne-Sophie Rigaud[3,5], Dominique Vaufreydaz[1,2,6]

[1] Inria
[2] Laboratoire d'Informatique de Grenoble (LIG)
[3] EA 4468, Living Lab Lusage, Université Paris Descartes
[4] MINES ParisTech, Centre de Recherche en Informatique
[5] Pôle de gériatrie, hôpital Broca, GH Paris Centre, Assistance Publique - Hôpitaux de Paris
[6] University of Grenoble Alpes

Author Version

Abstract

In the near future, robots will support human to perform tasks in many domains (industrial, domestic, educational and health tasks). Such robot behaviors need to take into account the social interaction between robot and human. In this context, we focus on the expressiveness of a moving head for an assistive robot for the elderly. We designed a new moving head for the *Kompaï* companion robot. On one hand, this new head improves its perception capabilities. On the other hand, we expect to jointly increase its social skills and thus its acceptability. This new head is composed of a tablet to animate a virtual face according to 4 facial expressions and a mechanical neck with 4 degrees of freedom to enhance the robot's expression. Before improving face expressions and adding more complex head movements, it is essential to evaluate the combination of simple head movements with virtual face expressions. A study was held jointly with physicians (psychologists, ergonomists) at the Broca Hospital in Paris to assess the impact to combine head movements with virtual face expressions, and the global acceptability of the *Kompaï* head by the elderly.

1 Introduction

After industrial application, recent improvements in robotics will lead in the near future to an increased role of robots in many news domains: domestic, educational or, as seen more recently, health tasks at home. Robots are increasingly seen as potential companions for everyday life. As stated by Dautenhahn in [1], *"A robot companion in a home environment needs to 'do the right things', i.e. it has to be useful and perform tasks around the house, but it also has to 'do the things right', i.e. in* ***a manner that is believable and acceptable to humans"***. To meet these requirements, a robot must have abilities to interact socially with humans. Two kinds of inter-correlated interactions are important for Human-Robot Interaction (HRI): physical interaction and social interaction. Physical interactions have been extensively studied for manufacturing use, medical use, etc. Social interaction is particularly challenging and needs to consider two points of view: 1) Interpretation by the robot of the human intentions, and 2) Interpretation by the human of the robot intentions. The first point requires databases of human interaction in different contexts where each data is annotated. These databases are the input of machine learning techniques (computer vision and multimodal perception domains). The second point is based on human interpretation. The only way to evaluate the contribution of proposals is to perform experimentation with human using methods from ergonomics.

In earlier work, we addressed embedded perception of humans [2] on companion robots. For this perception task, a new head was designed for *Kompaï*, a companion robot. *Kompaï* (Figure 1) is manufactured by Robosoft[1]. Its dimensions are 140 centimeters in height and 55 centimeters in width. *Kompaï* is a mobile robot which can either act autonomously or be driven by a user using a remote control or voice commands. It is equipped with various sensors for its environment: Lidar, infrared and ultrasonic telemeters, cameras, RGB-D sensor (Kinect 1). Using these sensors, *Kompaï* can perform elementary actions such as locate itself in the environment, stop before obstacles, and detect people. Its new head is based on a tablet with front/rear cameras and servomotors to mimic neck moves. Moving the head and looking at a specific position can improve

[1] `http://www.robosoft.com/`

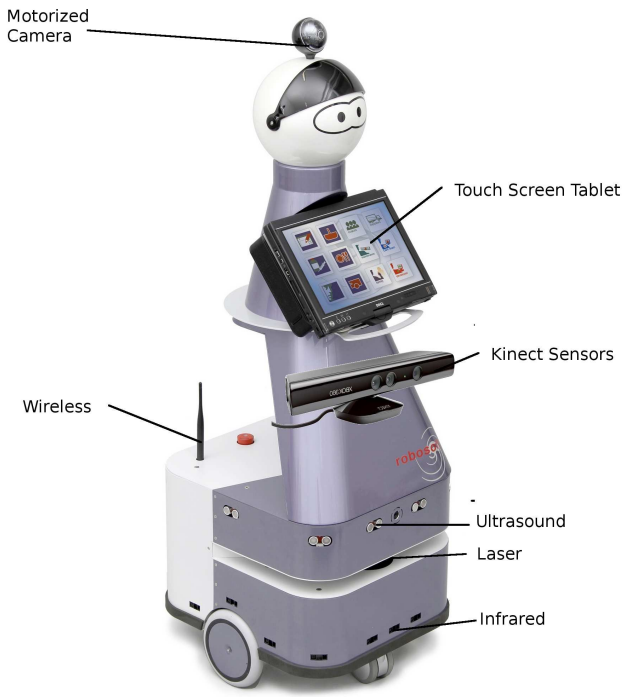


Figure 1: *Kompaï* is equipped with a laser range finder, ultrasound and infrared telemeters, a tablet PC, a camera in its base, a webcam on top and a Kinect on the torso.

detection of objects, human faces, bodies or hands. This allows the robot to anticipate interaction by rotating the head and align the robot's gaze with its human partner.

In this article, we address the impact of this moving robot head in the interaction loop with human. More precisely, in the context of a companion robot interacting with elderly people, we want to evaluate if head movements:

- do not disrupt perception of expressed emotions;
- can improve expressiveness of the robot, thereby improving its overall acceptability.

We describe the results of a study conducted jointly with physicians (psychologists, ergonomists) at the Broca Hospital in Paris in early 2015. In this study, we have evaluated several criteria: qualitative evaluation of the emotions expressed by the robot, impact of movements on the emotion expression and global acceptability of the robot head.

In the following, Section 2 presents related works. Section 3 describes the design and technical implementation of the expressive moving head. This is followed by a section on evaluation protocol and the results of the study with elderly people. The paper ends with conclusion and perspective.

2 Related Works

As robotic technologies progress, an increasing number of studies have addressed the impact of design on human robot interaction. Such studies often use a Wizard-of-Oz technique (WoZ), as in [3, 4], combined with user interviews on specific key aspects of the experimentation [5]. [6] and [7] have shown that the shape of the robot and its autonomous movements have important influence on the anthropomorphic perception of the robot. Several experiments have demonstrated that even simple devices, such as cleaning robots, can elicit empathy in humans [8]. This tendency to "*anthropomorphize*" is even more important with humanoid robots and increases expectations for overall performance [1].

Numerous companion robots can be found in the literature [5, 7, 9–13]. Such robotics can be grouped into 3 categories. The first group includes pet robots. Historically, the dog-like Aibo, was the first pet robot sold at large-scale, with early models reaching the market in 1999. More recent examples of this category include iCat, Paro (a baby seal) and Pleo (a dinosaur). The second category contains utility robots, usually with some anthropomorphic attributes such as gaze, head movement, facial expression, and arm gestures. In this category, one can find Sparkly, IROMEC, Pearl, Robotcare, Care-o-bot, Max from the Serroga project, the *Kompaï* used

Figure 2: New design for the robot head. Left, the original head, right the new design with a tablet.

in this project as well as the Aldebaran Pepper. The last category includes humanoid robots such as the NAO, KASPAR, a small child robot, and the Geminoid series, a series of humanoid robots used for telepresence.

For the design of the head of a companion robot for elderly people, we drew inspiration from robots in the second category. The first inspiration came from Sparkly [7], a small mobile robot with an expressive head. By changing the shape of metal wires on its face, it can express several emotions: *happiness*, *sadness*, *surprise* and *anger*. The IROMEC robot is similar but a screen replaces the metal wires. The facial designs of both robots are easily understandable with their smiley design. Other robots could be more massive, almost at human size, with an increase social presence. Care-o-bot is a companion robot comparable to the *Kompaï* robot. It is not anthropomorphic but has a mobile torso. In a user study [14], joint attention was elicited by rotating the robot's torso to simulate gaze direction. Results demonstrated that participants preferred this condition and that movements increase the social impact of the robot. Within the Serroga project, an evaluation of *Max*, a companion robot for elderly, has been conducted [5]. Its design integrates a topped head with 2 expressive eyes but no neck. *Max* was tested from 1 up to 3 days at home with elderly people. Beyond the functional evaluation done in this experiment, participants demonstrated acceptance of the robot as a health assistant and as a social companion. In their conclusion, the authors state that "*robots provide psychosocial and instrumental advantages due to their embodiment, mobility, and social presence.*".

Several studies have been conducted with Nao, KASPAR and some other robots [9–11,15]. The results indicate that head, gaze, embodiment and movements have an impact on the social presence and on the acceptance of the robot.

3 Experimental device

The original head of the *Kompaï* is a plastic spherical head with a camera on top. Based on state-of-the-art results and our own past experience [16], we made several design choices to replace it. The new head is composed of an Android tablet mounted on a mechanical neck. Figure 2 depicts the plastic spherical head and the new head.

3.1 Facial expression design

The Android tablet displays several animated facial expressions according to the situation and the selected visual design. For the design of the expressive face, we compared the smiley approach used with Spackly and IMOREC (see section 2) with a virtual agent approach called *Louise*. The *Kompaï* smiley design[2] was inspired by the original plastic head design. The virtual human face used in this study is from the virtual character *Louise*, meant for a future version of an assistive embodied conversational agent project [17]. The character model was created using Autodesk Character Creator and the expressive face images were created by animating and rendering this model using Autodesk Maya [3].

In our context to assess the impact of an animated robot head for elderly people, we choose to study the following basic emotions: *happiness*, *sadness*, *surprise*, *anger*. Figure 3 shows the facial expressions for *Kompai* and for *Louise* faces. An unpublished internal evaluation about *Kompaï* facial expression recognition has been conducted on 156 persons from 6 to 89 year old, half female/male and 90.77% right-handed. Recognition rates were 97.92% for *happiness*, 77.93% for *sadness*, 81.41% for *surprise* and 93.39% for *anger*.

[2] The smiley faces were provided by the Robosoft company in collaboration with Inria researchers.
[3] `http://www.autodesk.com/products/maya/overview`

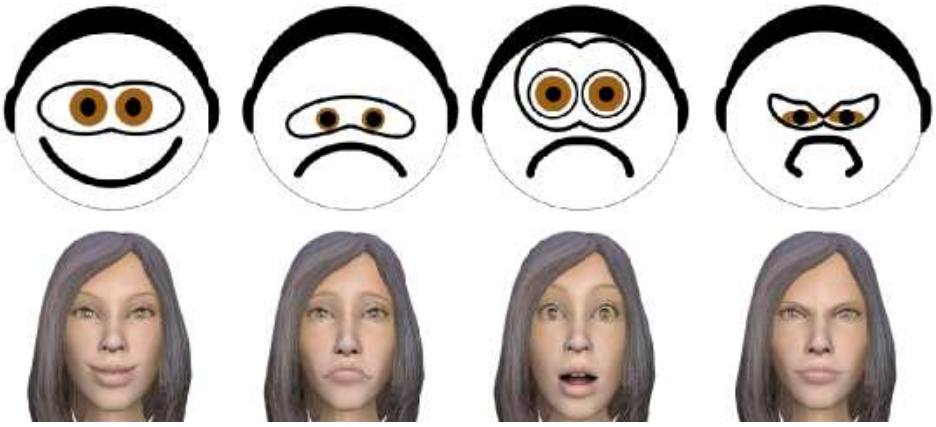


Figure 3: *Kompaï* and *Louise* expressive faces designed for Android tablet application. From left to right, one can find *happiness*, *sadness*, *surprise* and *anger* expressions.

3.2 Mechanical neck design

The objective is to animate the robot head to combine facial robot expressions with a neck movement like human behavior. The mechanical neck is composed of 4 servomotors: 1 in the base for rotation, 3 in the upper part for movements. The mechanism is closed to the robot used in human honesty experiment by Hoffman et al. [18]. The neck is in constructed from wooden and was made using a laser cutter. The "L" shape allows the head to lower onto its neck. It also acts as vibration absorber without power consumption while the robot is moving. Figure 4 depicts the mechanical design of the *Kompaï* head.

To control the mechanical neck, we developed SmartServoFramework[4], an open source project freely available. This framework lets us control movements for all expressions using widely spread servomotors, in this case Dynamixel© servomotors. Gesture associated to facial expressions can be described as follow:

- the *happiness* movement moves up proudly;
- for *sadness*, the head moves to look down;
- *surprise* consists of a recoiled movement;
- to express *anger*, the head goes forward and nods laterally.

In order to be comparable, these movements are the same for *Kompaï* face and *Louise* face.

3.3 Wizard of Oz and head remote controller

Based on a REST architecture (*REpresentational State Transfer*), in this study, the expressive robot head can be remotely controlled by a Wizard-of-Oz (WoZ) experimenter using a smartphone application (depicted on figure Figure 5). It is possible to switch between *Kompaï* face and *Louise* faces, or to activate the head animation associated to each facial expression. The experimenter can choose which facial expression to play. We added 2 joysticks (4 degrees of freedom) to control the neck. Using these joysticks the experimenter can dynamically change the behavior of the robot. For example, one can give the illusion that the robot is looking around to find a person.

4 Evaluation and discussion

4.1 Assessment protocol

We conducted collective and individual assessments at the Broca Hospital in Paris supervised by computer scientists and physicians (psychologists and ergonomists). To assess the acceptability of our mobile expressive robot head by elderly people, we retain several criteria: qualitative evaluation of the emotions expressed by the robot, impact of movements on the emotion expression and global acceptance of the robot head. Movements associated to the facial expressions are quite short. To reduce impact of the presentation order, each stimulus was played randomly several times.

[4] `https://github.com/emericg/SmartServoFramework`

Figure 4: Head mechanical design. The head is built with 4 Dynamixel© servomotors, laser cutted wood and Plexiglas for the structure and the tablet support. During experimentation, the neck is covered with a textile muff.

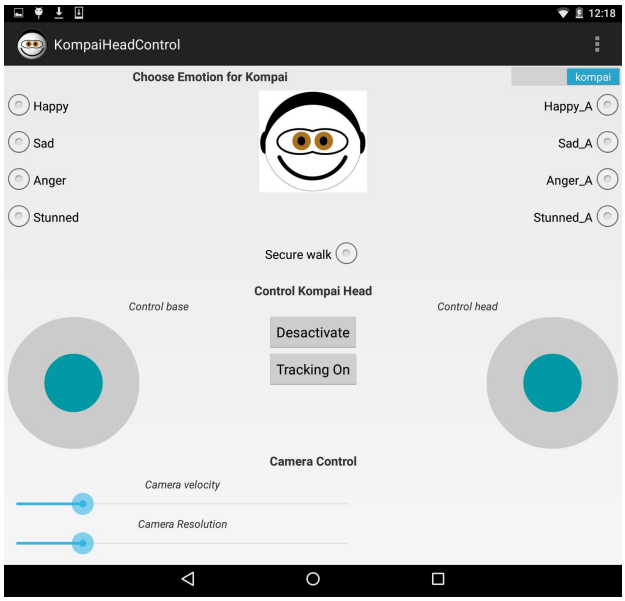


Figure 5: Remote control of the expressive robot head. Using this remote control, the experimenter can switch from the *Kompaï* to *Louise* rendering, launch predefined animations and control the 4 degrees of freedom of the head.

4.1.1 Gathered information

To compile sociodemographic data, all participants were asked to complete a form relative to their situation: age, gender, education, marital status, monthly income and work situation (see table 1). Participants had to meet several criteria to be included in the study: be over 54 year old, be in good cognitive and physical health, without severe visual or hearing pathology, and must understand French.

Each participant assessed the robot head's expressiveness with and without head movement. To remove issues related to technical vocabulary, we asked participants to associate an emotional state to the expressive head using a word of their choice for all 4 conditions: *Kompaï* or *Louise*, with or without mechanical animations. These words were tagged by a psychologist according to the semantic proximity between the word and the robot's expression. We retained 3 semantic distances: *irrelevant concept* when the word is out of the semantic field of the facial expression, *close concept* when the word is in its semantic scope but does not match with it, and *correct concept* when the word is relevant.

At the end, we asked participants to complete a survey asking information about their perception and acceptance of the robot head. All the sessions with elderly participants were recorded and annotated by psychologists. These annotations are useful to get free comments from the participants about the robot in context.

4.1.2 Collective assessment

The collective assessment was organized as a discussion on the topic of "Robots and expressiveness". The first step was to familiarize the elderly subjects with the robot *Kompaï*. Afterward, an open discussion was held about the requirements to improve the quality of robot-human interaction using an expressive head. Underlying topics are acceptability and usefulness of such head. The discussion was oriented towards the expressiveness of a robot head with a mechanical animation associated to a virtual animation of the face (discussion about ergonomics and behavior of robots in human-robot interaction).

For this experiment, 7 participants (6 men and 1 woman) were supervised by 3 physicians (2 psychologists and 1 ergonomist) of the Living Lab and 3 computer science researchers. The session in the Living Lab lasted two hours. The participants assessed the way they perceived each emotion for each condition: *Kompaï* and *Louise*, with and without mechanical animation. They also compared these conditions with the original plastic

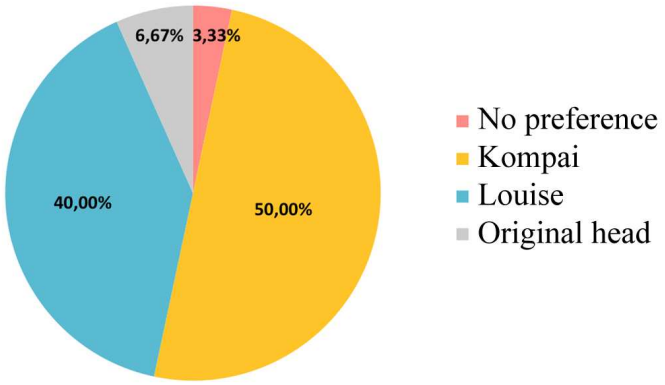


Figure 6: Preference among head version for the elderly.

Variables	Classes	Frequency (n=30)	Percentage
Age	≤ 70	15	50.00%
	> 70	15	50.00%
Gender	M	6	20.00%
	F	24	80.00%
Education	Primairy school	2	6.67%
	Secondary and high school	5	16.67%
	Higher learning	23	76.67%
Martial status	Single	17	56.67%
	Couple	13	43.33%
Monthly income	$< 2,500$ €	13	43.33%
	$> 2,500$ €	17	57.00%
Work situation	Unemployed	27	90.00%
	Working	3	10.00%

Table 1: Sociodemographic representation of participants

head. The results of these assessments are described in the section 4.2. Only 2 participants from this collective assessment appear in table 1 as we have not complete sociodemographic data for the others.

4.1.3 Individual assessment

We recruited 28 volunteers to take part in an individual assessment of the expressive robot head. The session took place at the Broca Hospital. Each individual interview was supervised by 2 researchers. The participants assessed the way they perceived each emotion for each face (*Kompaï* and *Louise*) with and without mechanical animation. And they compared with the previous plastic head. The results of these assessments are described in Section 4.2.

4.2 Results & discussion

As explained in the previous section, final results include 30 participants. Sociodemographic information of the sample is presented in table 1. Data shows balanced distribution regarding age (in range [54,90], half ≤ 70, $avg = 72.66$, $\sigma = 9.03$), marital status and monthly income. Males and higher education are respectively under and over represented. As expected for elderly people, only 10% are still working.

Table 2 presents the results of the expressive robot head assessment. For the *Kompaï* head with or without mechanical animations, happiness is the most relevant expression: 83% and 63% of participants correctly recognized this expression. For the *Louise* head with or without mechanical animation, surprise is the most easily recognized expression with correct recognition by 90% and 86% of participants. The *Kompaï* condition obtains its worst score (16% and 10%) for this expression with a high rating for the irrelevant expression (76% without mechanical animations and 86% with mechanical animations). In our internal evaluation of the smiley design (see section 3.2), the score for the static version of surprise shows a higher score (81.41%). From this,

Happiness	**Kompai (static)**		**Kompai (moving)**		**Louise (static)**		**Louise (moving)**	
	Freq.	*%*	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*
Irrelevant concept	3	10	6	20	10	33.33	9	30
Close concept	2	6.67	5	16.67	4	13.33	1	3.33
Correct concept	25	83.33	19	6.33	16	53.33	20	66.67

Sadness	**Kompai (static)**		**Kompai (moving)**		**Louise (static)**		**Louise (moving)**	
	Freq.	*%*	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*
Irrelevant concept	7	23.33	14	46.67	21	70.00	22	73.33
Close concept	12	40	2	6.67	0	0	2	3.33
Correct concept	11	36.67	14	46.67	9	30.00	6	20.00

Surprise	**Kompai (static)**		**Kompai (moving)**		**Louise (static)**		**Louise (moving)**	
	Freq.	*%*	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*
Irrelevant concept	23	76.67	26	86.67	2	6.67	4	13.33
Close concept	2	6.67	1	3.33	1	3.33	0	0
Correct concept	5	16.67	3	10	27	90.00	26	86.67

Anger	**Kompai (static)**		**Kompai (moving)**		**Louise (static)**		**Louise (moving)**	
	Freq.	*%*	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*
Irrelevant concept	5	16.67	11	36.67	22	73.33	22	73.33
Close concept	2	6.67	5	16.67	2	6.67	2	6.67
Correct concept	23	76.67	14	46.67	6	20.00	6	20

Table 2: Rating of the head expressions for each configuration

we have concluded that the surprise expression may have been improperly designed and should undergo a redesign. The sadness expression is quite balanced between the 2 versions. The anger expression is not well interpreted with the *Louise* head but it is unclear whether the expression design is weak for *Louise* or whether this expression is not well accepted by the participants.

The mechanical animation maintains the relevance for head expressions. But with the *Kompaï* head, mechanical animation can reduce the interpretation of the head expression. With the *Louise* version, this does not occur. Results with and without mechanical animations are very similar. While the mechanical animation does not modify the acceptability of the robot head, it also does not significantly improve its expressiveness.

From the data, one can observe that, with a relevant design of the facial expressions, the elderly interpret more easily a facial expression with a "smiley" face design rather than a virtual agent. This result matches with the participants' preferences depicted in Figure 6. Indeed, the elderly people prefer the tablet system at 90% with a small inclination for the *Kompaï* version (50% versus 40%). The original spherical head is less supported (6.67%).

To interpret more deeply these results, we run a bivariate analysis between the *Static* and *Moving* conditions for all facial expressions (*Happiness*, *Sadness*, *Surprise*, *Anger*). There are few significant differences to highlight. The motion is significant for *Louise* for the Happiness and Sadness facial expressions with *p-value* respectively equals to 0.003 and 0.008. For the *Surprise* and *Anger* expressions, the motion is significant for *Kompai* with *p-value* equals to 0.007 and 0.0009. All other comparisons are not significant. From these results, we can affirm that, in our experiment, most of the time there is no significant difference between static and moving conditions. In the few cases of a significant difference, there is a decrease of the recognition using motion except for *Hapiness* with *Louise* (table 2). As we write this article, we cannot definitely conclude about reasons of this deterioration. Several hypotheses could be envisioned: design of the motions associated to the expressions, flatness of the head or interdependence between tablet rendering and movements for instance. We are currently working on a new version of the robot to question these hypotheses.

Last, free comments from people provide complementary information about these results. For the original head, people like its 3D volume that "*let see the head from a 3/4 view*". The tablet is appreciated as it is more expressive. Within comments, there is no clear consensus about the anthropomorphism choice to make for the head. Some people prefer the *Louise* version precisely because it is anthropomorphic. Others said that the smiley version is better because it "*stays a robot*" and it is more playful.

5 Conclusion

In this article, we focus on social Human-Robot interaction from the human point of view. More precisely, we evaluate the impact of an expressive moving head for an assistive robot interacting with elderly people. The expressive moving head was designed with 4 degrees of freedom for our robotic platform. This new head is composed of a tablet that provides an animated virtual face according to 4 facial expressions and a mechanical neck to enhance the expressiveness of the robot.

To assess the impact of head movements combined with virtual face expressions, and the global acceptability of this head by the elderly, a study was held jointly with physicians (psychologists, ergonomists) at the Broca Hospital in Paris. This study compares a virtual 3D human character (*Louise*) to a virtual "smiley" face (*Kompaï*), with or without mechanical animations to gauge impact of movements. From the experiments we draw two conclusions:

- The recognition of the facial expressions by Elderly people is easier with a virtual "smiley" face design rather than with a virtual 3D human character;
- The mechanical animation maintains the relevance of expression. While it does not significantly affect the acceptability of the robot's head, it also does not improve its expressiveness.

These conclusions have raised several questions. Which neck movement associated to a virtual facial expression to improve the global robot expression? Are there expectations on the animation of the neck according to the virtual face representation (difference between *Kompaï* and *Louise* condition)? New experiments should be conducted to generalize the results. In addition, it would be interesting to create and to evaluate new head animations to complete these results.

6 Acknowledgement

The authors would like to thank the Robosoft Company for providing the *Kompaï* robot and necessary materials. These researches have been conducted within the PRAMAD2 FUI project, founded by the French Ministry of Education and Research and the French National Research Agency (ANR). Prototyping was done using the Amiqual4Home facilities (ANR-11-EQPX-0002).

References

[1] K. Dautenhahn, "Human-robot interaction," in *The Encyclopedia of Human-Computer Interaction, 2nd Ed.*, M. Soegaard and R. F. Dam, Eds. Denmark: The Interaction Design Foundation, 2014, ch. 38, [Online https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/human-robot-interaction; accessed 01-February-2016].

[2] D. Vaufreydaz, W. Johal, and C. Combe, "Starting engagement detection towards a companion robot using multimodal features," *Robotics and Autonomous Systems*, vol. 75, pp. 4–16, 2015.

[3] E. S. Kim, L. D. Berkovits, E. P. Bernier, D. Leyzberg, F. Shic, R. Paul, and B. Scassellati, "Social robots as embedded reinforcers of social behavior in children with autism," *Journal of autism and developmental disorders*, vol. 43, no. 5, pp. 1038–1049, 2013.

[4] P. Worthy, M. Boden, A. Karimi, J. Weigel, B. Matthews, K. Hensby, S. Heath, P. Pounds, J. Taufatofua, M. Smith, S. Viller, and J. Wiles, "Children's expectations and strategies in interacting with a wizard of oz robot," in *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*, ser. OzCHI '15. New York, NY, USA: ACM, 2015, pp. 608–612. [Online]. Available: http://doi.acm.org/10.1145/2838739.2838793

[5] H.-M. Gross, S. Mueller, C. Schroeter, M. Volkhardt, A. Scheidig, K. Debes, K. Richter, and N. Doering, "Robot companion for domestic health assistance: Implementation, test and case study under everyday conditions in private apartments," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on.* IEEE, 2015, pp. 5992–5999.

[6] B. J. Scholl and P. D. Tremoulet, "Perceptual causality and animacy," *Trends in cognitive sciences*, vol. 4, no. 8, pp. 299–309, 2000.

[7] B. Friedman, P. H. Kahn Jr, and J. Hagman, "Hardware companions?: What online aibo discussion forums reveal about the human-robotic relationship," in *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM, 2003, pp. 273–280.

[8] L. D. Riek, T. C. Rabinowitch, B. Chakrabarti, and P. Robinson, "Empathizing with robots: Fellow feeling along the anthropomorphic spectrum," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, Sept 2009, pp. 1–6.

[9] J. Broekens, M. Heerink, and H. Rosendal, "Assistive social robots in elderly care: a review," *Gerontechnology*, vol. 8, no. 2, pp. 94–103, 2009.

[10] L. Damiano, P. Dumouchel, and H. Lehmann, "Towards human–robot affective co-evolution overcoming oppositions in constructing emotions and empathy," *International Journal of Social Robotics*, vol. 7, no. 1, pp. 7–18, 2014. [Online]. Available: http://dx.doi.org/10.1007/s12369-014-0258-7

[11] J. Wrobel, M. Pino, P. Wargnier, and A.-S. Rigaud, "Robots and virtual agents to assist older adults: A review of present day trends in gerontechnology," {*NPG*} *Neurologie - Psychiatrie - Griatrie*, vol. 14, no. 82, pp. 184 – 193, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1627483014000403

[12] S. S. Kwak, Y. Kim, E. Kim, C. Shin, and K. Cho, "What makes people empathize with an emotional robot?: The impact of agency and physical embodiment on human empathy for a robot," in *RO-MAN, 2013 IEEE*. IEEE, 2013, pp. 180–185.

[13] M. Watanabe, K. Ogawa, and H. Ishiguro, "Field study: can androids be a social entity in the real world?" in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 2014, pp. 316–317.

[14] J. Saez-Pons, H. Lehmann, D. S. Syrdal, and K. Dautenhahn, "Development of the sociability of non-anthropomorphic robot home companions," in *Development and Learning and Epigenetic Robotics (ICDL-Epirob), 2014 Joint IEEE International Conferences on*. IEEE, 2014, pp. 111–116.

[15] W. Johal, S. Pesty, and G. Calvary, "Towards companion robots behaving with style," in *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*. IEEE, 2014, pp. 1063–1068.

[16] R. Barraquand, "Designing sociable technologies," Ph.D. dissertation, Grenoble INP, 2012.

[17] P. Wargnier, G. Carletti, Y. Laurent-Corniquet, S. Benveniste, P. Jouvelot, and A.-S. Rigaud, "Field evaluation with cognitively-impaired older adults of attention management in the embodied conversational agent louise," in *4th International Conference on Serious Games and Applications for Health (IEEE SeGAH 2016)*, 2016.

[18] G. Hoffman, J. Forlizzi, S. Ayal, A. Steinfeld, J. Antanitis, G. Hochman, E. Hochendoner, and J. Finkenaur, "Robot presence and human honesty: Experimental evidence," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 181–188.

Natural Vision Based Method for Predicting Pedestrian Behaviour in Urban Environments

Pavan Vasishta[1], Dominique Vaufreydaz[2] and Anne Spalanzani[3]§

Abstract

This paper proposes to model pedestrian behaviour in urban scenes by combining the principles of urban planning and the sociological concept of *Natural Vision*. This model assumes that the environment perceived by pedestrians is composed of multiple potential fields that influence their behaviour. These fields are derived from static scene elements like side-walks, cross-walks, buildings, shops entrances and dynamic obstacles like cars and buses for instance. Using this model, autonomous cars increase their level of situational awareness in the local urban space, with the ability to infer probable pedestrian paths in the scene to predict, for example, legal and illegal crossings.

1 INTRODUCTION

As the race to attain and deploy fully autonomous vehicles on urban roads heats up, the concept of Situational Awareness (SA) takes centre stage. Situational awareness is the natural human ability to understand, react and predict the environment based on previously learnt parameters, the utilisation of which is most frequently seen while driving. Human drivers need to balance different variables – speed, route selection, positions of pedestrians, cyclists, other cars, etc. – while trying to predict their future states. In fact, errors in maintaining situation awareness are the most frequent cause of errors in real-time tasks such as driving [1] and can be attributed to many accidents.

Situational Awareness can be described in three incremental abstract levels [2] as *Perception*, *Comprehension* and *Projection*. A human driver in an urban street cycles through these levels continuously. Objects and elements– pedestrians, obstacles, other cars, cross-walks, interesting areas– in the environment are identified. These elements are contextually understood with regard to the environment that they are in– the answer to the question "Why is that element there?" For example, the answer to "Why is there a cross-walk on the street" is to facilitate a crossing from one side of the road to another. Finally, the objects and elements are understood together and their future interactions are projected with a certain probability: a cross-walk *may* be used by a pedestrian if he/she is close to it. A human driver's specific course of action is decided by these continuously evolving projections.

The main motivation of this work is to increase the situational awareness of an autonomous car in the context of driving in urban streets. A major driving force is the adoption of sociological ideas for understanding pedestrian behaviour in inner city areas. Pedestrian behaviour has been postulated to be a function of the built environment; i.e. their movement is a consequence of the presence of certain positive and negative *attractors* [3]. This behaviour, called Natural Motion, is an extension of Gibson's *Natural Vision* which envisages human behaviour as wanting to move in a direction that interests them the most in their field of view [4]. These positive attractors, here called "Points of Interest (POI)", may be present as an element in the scene. They can be monuments, places of public interest, public transportation... Other, more common, POIs are areas of commercial interest - stores, restaurants, etc., that are seen very frequently in an urban centre [5]. The presence of these POIs in any scene influences the behaviour of pedestrians within it. Understanding these influences allows the autonomous car to perform actions that are instinctive in a human driver - project pedestrian future states and intuit areas of legal and illegal crossings.

*[1] Pavan Vasishta is a PhD student in the CHROMA team (Univ. Grenoble Alpes, Inria, 38000 Grenoble, France, email: Pavan.Vasishta@inria.fr)

†[2] Dominique Vaufreydaz is an associate professor in the Pervasive Interaction team (Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP*, LIG, 38000 Grenoble France, email: Dominique.Vaufreydaz@inria.fr)

‡[3] Anne Spalanzani is an associate professor in the CHROMA team (Univ. Grenoble Alpes, Inria, 38000 Grenoble France, email: Anne.Spalanzani@inria.fr)

§* Institute of Engineering Univ. Grenoble Alpes

The major contribution of this work is the creation of a new framework for quick comprehension of urban streets. It also forms a base to project future states of pedestrians without the need of their presence in the scene, analogous to a human driver's intuition. We model the scene as attractive and repulsive potential fields. The novelty of our approach is the introduction of POIs whose attractiveness influences pedestrian behaviour.

The paper is divided into five sections. Section II deals with related work in the field of pedestrian motion prediction, followed by section III, the theoretical basis of the framework and the methods used to project future states. Section 4 discusses the implementation of this framework, its results and validation of a conducted experiment. Section 5 concludes this paper with a discussion on the current work and envisaged future work.

2 RELATED WORK

As far as the authors know, there has been little to no work done in the field in accounting for Point of Interest influences in urban pedestrian prediction. Much work, however, has been done in modelling and prediction of pedestrian routes and route-choice behaviour [6]. Most pedestrian behaviour prediction algorithms depend on learning frameworks based off of observed data. Modelling the inherent pedestrian variables as one technique. A data driven approach on this, minimising an energy function that accounts for many personal factors like speed, grouping etc., can be found in [7]. A similar approach, based on the *Social Force Model* can be seen in [8].

Others use Markov Decision Processes (MDPs) and its variants to predict the beliefs of pedestrian crossings in a scene [9]. Destinations are assumed to be known. In close spirit to our work are [10], [11] and [12]. These works build a cost function based on the environment. [10] learns the cost function of the environment via observed trajectories. An MDP is solved with rewards based on observed trajectories as well with known destinations. Working on static scenes, it requires previously observed trajectories to model the cost function and thus is infeasible for rapidly changing scenes on autonomous vehicles. [11] continues this work by semantically segmenting the observed environment to construct a cost function fed to an hMDP to predict pedestrian positions, even without pedestrian observations. While this knowledge is transferable, it is computationally expensive. The computation problems are solved in [12], yet being applied only for static scenes with learned trajectories. Last, [13] looks at the distance of the pedestrian from the kerb, position and velocity of the car from the cross-walk to learn an "*Inner city model*" to predict pedestrian crossings.

These works require learning from observed data to predict pedestrian positions. Our framework envisages to solve the issues of computational complexity and dynamicity while considering environmental and social factors. It also does *not* use learning for building environmental models which makes it very useful to deploy in unknown environments, with very few dependencies.

3 THEORETICAL FRAMEWORK

Pedestrian crossing behaviour in urban areas can be classified into two broad categories - legal crossings and illegal crossings. Legal crossings are such movements of a pedestrian that account for the safest path from one side of the street to the other. These generally happen on a cross-walk. An illegal crossing is an abnormal behaviour wherein the pedestrian decides to not take the cross-walk to cross the street.

In a structured urban environment, for legal crossings to occur, certain assumptions are made:

- The edges of the road repel pedestrians such that their paths are restricted to the side-walk.
- A cross-walk acts as a conduit between the two sides of the street and offers no resistance to crossing
- The road acts as a barrier for crossing, repelling pedestrians towards the side-walks.
- Static and Dynamic obstacles on the road are repulsive in nature, increasing the resistance of the road and pushing back pedestrians towards side-walks.
- Side-walks offer no resistance to pedestrian movement.
- Points of Interest are a reason for pedestrians to cross from one side of the street to another.

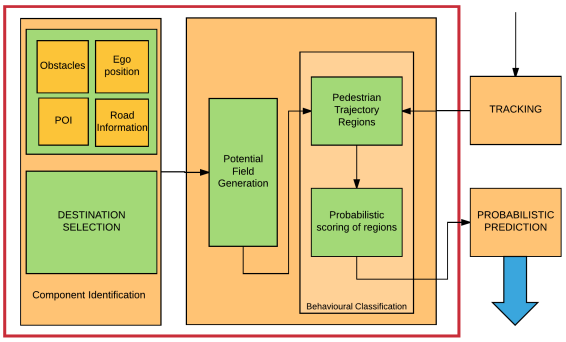


Figure 1: Architecture of the Framework. Only the red block has been implemented in this work

An illegal crossing occurs when at least one of these assumptions is violated. Predicting these areas of illegal crossings leads to a higher level of SA. Looking at these assumptions, it can be seen that a system of potential fields [14] can be a good fit as a model for explaining urban behaviour. Each of the assumptions made earlier can be represented as a function of a potential field.

The architecture of the proposed framework can be found in Fig. 1. From the observed scene, certain informations need to be extracted. These informations are road width, number of lanes, the closest Points of Interest (POI) from the observer, their orientation, the position of the closest cross-walk, etc. Static and Dynamic obstacles on the road also need to be identified.

Destination points in the scene are chosen and fed forward along with the scene information to the Potential field generator. This potential field generation step generates a grid with a potential "Map" based on the extracted data. The generated "Map" is used by the Behaviour classification module, first to generate pedestrian trajectory regions and to score these regions with a probability.

Firstly, to demonstrate the model, it is assumed that the scene under consideration is well-structured. This implies that there is an observable demarcation between the road and the side-walk, the road and the cross-walk and that the lanes on the road are easily observable. It is also assumed that the width of the road (L_{road}), the width of the lane (L_{Lane}), and the POI positions are known, can be computed using sensors embedded into the autonomous car or can be retrieved from a global map. Given the position of the ego vehicle, the distance to the cross-walk and the side-walk can be extrapolated from known information. Considering the pedestrian as a self-driven particle under the influence of attractive and repulsive forces, a potential field can be constructed which produces certain motion behaviours [5].

A destination, by definition, draws a self driven particle towards it. Thus, a POI can be a destination. Conversely, all destinations in a scene are points of interests. Making this assumption, the viable ends of the observed scene are designated as POI and the potential field is recalculated.

A grid is defined for the observed area with its origin at the top left corner and extending to (X_m, Y_m), the maximum grid values with a specified grid resolution. Each cell on the grid can take the attributes *road*, *cross-walk*, *empty*, *POI*, *obstacle* and *edge* and a range of values between 0 and 1. Taking inspiration from work done in [15], the resultant model is a linear function of all component potential values at each cell in the grid. Thus,

$$\begin{aligned} \mathbf{U_{total}} = \mathbf{U_{Edge}} + \mathbf{U_{Road}} + \mathbf{U_{Obs}} + \\ \mathbf{U_{CW}} + \mathbf{U_{POI}} \end{aligned} \tag{1}$$

where $\mathbf{U_{Edge}}, \mathbf{U_{Road}}, \mathbf{U_{Obs}}, \mathbf{U_{CW}}$ and $\mathbf{U_{POI}}$ are potentials associated with the road edges, the road, obstacles on the road, the cross-walk and the POIs.

3.1 Computing Potentials

For each cell in the grid, the different potentials can be calculated as follows.

3.1.1 Edge Potential

The edge potential must repel pedestrians towards the center of the side-walk. An illegal crossing occurs when the self driven particle can exert enough force to overcome this potential. For each cell with a center (x, y), the value of the potential is defined by

$$\mathbf{U_{Edge}^{ij}} = \frac{1}{2}\eta\left(\frac{1}{\rho(x^{ij}, y^{ij})}\right) \tag{2}$$

where $\rho(x, y)$ is the distance of the i^{th} and j^{th} edge cell from all other cells in the grid. η is a scaling factor dependent linearly on L_{Road}. The calculated values are ceiled to an appropriate value. The total edge potential is a summation of potential values of all edge-containing cells.

3.1.2 Road potential

Based on a sociological study conducted in France [16], it can be inferred that the propensity of illegally crossing a road is linearly dependent on its width. Thus, a narrow road entices a pedestrian to cross illegally while a wider one does not.

For cell (C^{ij}) with an attribute *road*, the calculated potential value is:

$$\mathbf{U}_{Road}^{ij} = \beta_{Road} \exp\left(-\left[\left(\frac{x_{ij}-x_{road}}{\sigma_x}\right)^2 + \left(\frac{y_{ij}-y_{road}}{\sigma_y}\right)^2\right]\right) \tag{3}$$

β_{Road}, σ_x and σ_y are dependent on the width of the road as explained earlier.

The total road potential is the summation of potential values of all road-containing cells.

3.1.3 Obstacle Potential

Obstacles in the scene can be distinguished as static and dynamic obstacles. For either classification, the response of the self-driven particle under the effect of the obstacle remains the same. The self driven particle cannot cross through the obstacle and the approach to the obstacle is slow. A Yukawa potential [17] is considered a fit for the expected behaviour.

A static obstacle takes the shape that it is perceived to be. A dynamic obstacle(for example, other cars in the scene), is described as a rectangular shape with a triangular shape extending forward in the direction of motion. Thus, the potential is described by:

$$\mathbf{U}_O^n = \Lambda \frac{\exp\left(-\alpha \mathbf{K}\right)}{\mathbf{K}} \tag{4}$$

Where Λ and α decide the behaviour of U_O^n. Larger the values, sharper the drop off of the potential near the obstacle.

$\mathbf{K}$ is the distance of the obstacle from every point on the workspace, i.e.,

$$\mathbf{K} = \|C^{ij} - C^{Obs}\| \tag{5}$$

The total effect of all the obstacles in the workspace is given as

$$\mathbf{U_{Obs}} = \sum_{n=0}^{N} \mathbf{U}_O^n \tag{6}$$

Where N is the total number of obstacles observed. The extremely large values that are generated are truncated to a maximum viable value.

3.1.4 POI Potential

A Point of Interest (an inexhaustive list of what may be considered as a POI can be found in [5]) generates an attractive pull in the scene. With sufficient motivation, the self-driven particle can escape the influence of a POI. A POI is also a terminal point in the scene - the implication being that all exits in the scene are POIs. The potential of a POI situated at a cell defined by (x_{poi}, y_{poi}) is a Gaussian function centered at (x_{poi}, y_{poi}). $\beta_{poi}, \sigma_x, \sigma_y$ depend on the global importance of the specific Point of Interest.

3.1.5 Cross-walk Potential

The cross-walk connects the two side-walks of the street and acts as a resistance-less conduit for the self driven particle. Thus, the potential of the cross-walk is the smallest value in the area.

3.2 Behavioural Classification

3.2.1 Trajectory Regions

Pedestrian route choice behaviour, in inner city limits, can be described in terms of optimisations. A pedestrian either tries to perform a distance optimisation to a destination at one extreme or optimises for safety at the other. Thus, all possible pedestrian paths can captured between these two behaviours. From each destination in the scene to all the others, an A* search is performed with the heuristic:

$$h(s) = \alpha C(s, s+1) + (1-\alpha)\rho(g, s+1) \tag{7}$$

Where $C(s, s+1)$ is the cost of moving from the current state s to the next state $(s+1)$. $\rho(g, s+1)$ is the normalised distance between the goal (destinations) and the next state. α is the parameter contributing to the integration of safe trajectory optimisation and shortest distance optimisation. Varying this parameter allows for simulating the different trajectories pedestrians might take, like partially optimising for distance and partly for safety.

3.2.2 Probability map of pedestrian trajectories

Trajectories generated for each entrance and exit are collated to create regions of probable pedestrian trajectories. These are then analysed to find regions of overlap. The larger the value of the region of overlap, higher the probability that a pedestrian might be present there.

4 IMPLEMENTATION AND RESULTS

In the current work, the red block in Fig. 1 has been implemented on the dataset provided in [18]. This is a dataset generated in Martigny, Switzerland from a static camera overlooking a city square. The camera captures a scene containing 2 POIs and a crosswalk. It also captures pedestrian and traffic circulation. POI positions have been extracted from OpenStreetMaps. The scene at a specific instant is shown in Fig. 2

This scene is segmented into a grid containing features of the component potential fields - road, POI, crosswalk and sidewalks. Road parameters are taken from Swiss national standards and a potential field resulting from the constituent components are created as explained in 3. This potential field "Map" is then used to determine areas of probable trajectories and pedestrian probability.

4.1 Validation

To show that this approach follows *Natural Vision* [4] is sufficient validation to use it to efficiently model the environment. Naive and sufficient validations are described for the approach, based on chosen destinations.

Fig. 4 forms the crux of our work. This scene, taken from the acquired dataset, is the representation of the observation at a specific instant (leftmost image in Fig. 4). The width of the road and lanes were extracted manually based on Swiss national standards. With this data, the scene is reconstructed by placing its constituent elements at relative positions on a grid. The obstacle has been identified and tracked using the YOLO2 framework. Attributes of each cell on the grid is manually defined as crosswalk, side-walk, road, edge or obstacle, as well as for the POI. With these data populated on the grid, a resultant potential field "Map" has been generated as defined in section 3.

4.1.1 Primary Validation

A first naive validation of this method is to show that observed pedestrian presence matches those areas predicted by the framework. In this scene (Fig. 2), there are 4 potential destinations, marked (1), (2), (3) and (4). Destination (2) is the entrance of the visible POI while the others are the ends of the scene. Choosing one of these as a starting point, we compute trajectories of the safest and the shortest paths and everything in between to all the other exits. By repeating this at each exit, a map of areas of most probable pedestrian positions emerges. For instance on Fig. 3, two destinations – (1) and (4) – were chosen and the different trajectories determined. Superimposing areas leads to the pedestrian trajectory probability map at right on the same figure.

Pedestrians were detected using the YOLOv2 framework [19]. When the YOLOv2 framework failed, there was a manual annotation of pedestrians for detections. Trajectories were tracked based on these

Figure 2: Scene at a specific instant from the dataset. No dynamic obstacles on the road are observed at this instant.

detections. By accumulating these trajectories, a probability map based on observations was created (Section 3.2.2). Fig. 3 compares the predicted probability areas from our model and the observation.

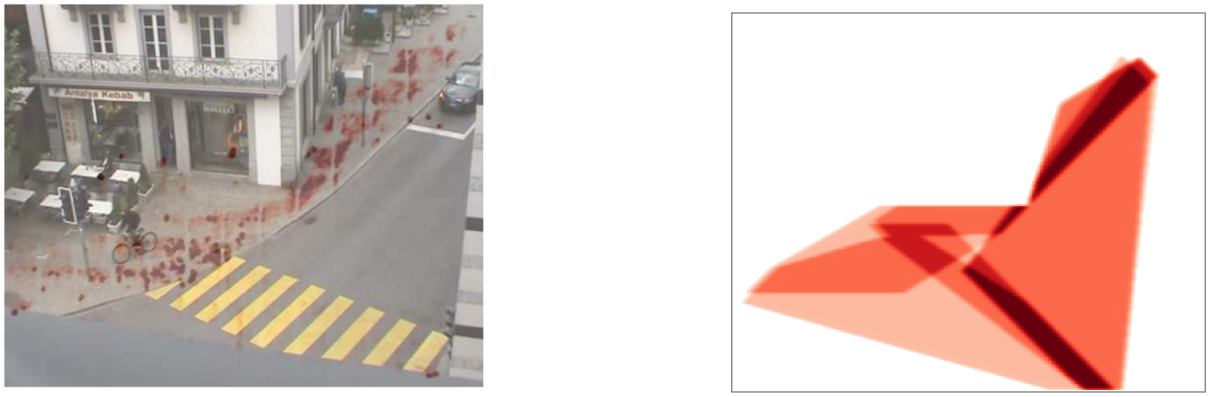


Figure 3: Comparison between observed probability map and predicted pedestrian trajectory probability map. The left image shows the ground truth. Right image is the predicted probability map of the scene. Darker the colour, higher the probability of pedestrians being present in that area.

In the observed probability map, it can be seen that there is a high incidence of pedestrians on the side-walk. The probability of a legal crossing is much higher compared to an illegal one, as observed. The predicted pedestrian trajectory probability map shows that there is a high probability that this scene has more chances of pedestrians crossing legally compared to an illegal crossing. It also predicts a high number of pedestrian trajectories into destination (3). Comparing to the ground truth, even though there are a few stray illegal crossings, they are overwhelmingly outnumbered by the number of legal ones. It is also observed that pedestrians continue to walk towards destination (3) compared to people crossing across into destination (2) in the scene.

4.1.2 Secondary Validation

A secondary validation of the framework is to prove that it predicts behaviour based on Natural Vision. Pedestrians stay within certain bounds while going towards a destination. Knowing this, areas of illegal crossings can be predicted. From the dataset, 38 trajectories were chosen and classified into eight different behaviours. Each case is classified based on the entry and exit destinations of the scene and the legality

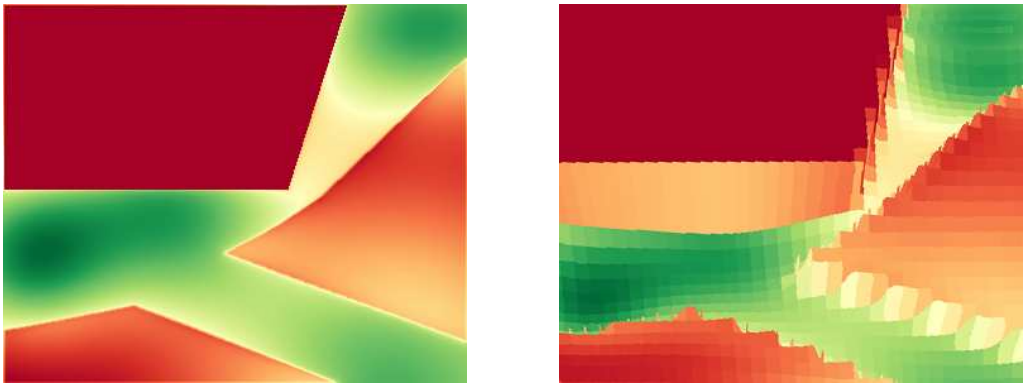


Figure 4: Resultant potential field of the observed scene. The scene contains a visible POI, a total of four destinations and a cross-walk. The right image is the 3D representation of the same scene. At this instant, there are no dynamic obstacles observed, the presence of which, changes the resultant field.

Table 1: Different behavioural cases observed in the dataset. Destinations can be seen on Fig 2.

Case	Destinations		Crossing
	From	To	Legality
I	1	3	Legal
II	1	2	Legal
III	4	2	Legal
IV	4	1	Legal
V	1	4	Illegal
VI	1	4	Legal
VII	4	1	Illegal
VIII	4	3	Legal/illegal

of crossing between them. The behavioural classes for pedestrians walking in the scene (Fig. 2) can be found in Table 1.

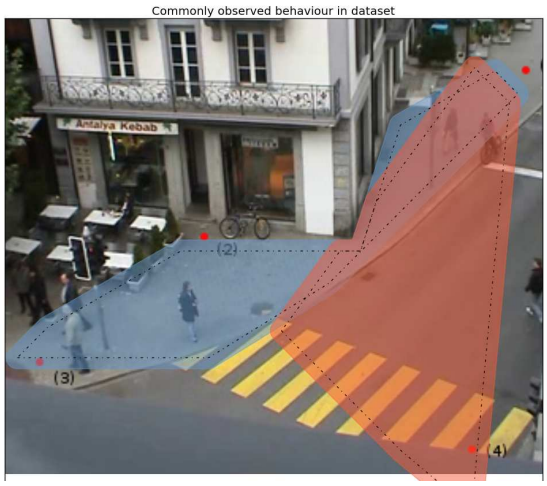


Figure 5: Validation of the two most common behaviours in the dataset: case I and case VI (see table 1). The dashed line corresponds to the A* predicted regions for each case. The coloured boundaries represent the extended zone.

As seen in section 3.2.1, the A* algorithms returns a sharp trajectory, going straight for the goal based on the heuristic provided. It does not meander when there are no explicit potential modifications. The zones predicted by just using the A* algorithm can be seen as the dashed line in Fig. 5. Human motion very rarely follows a perfect straight line. Thus, to account for these random motions, the zones were dilated by 40 cms on all sides. This leads to a much better prediction score as can be seen in Table 2. This is sufficient to show that the principle of *Natural Vision* is valid and can be used to determine pedestrian behaviour.

Table 2: Quantitative analysis of trajectories within predicted regions

Case	Nb Trajectories	A* Predicted zone		Extended zone (40cms)	
		Inside zone (%)	Outside zone (%)	Inside zone (%)	Outside zone (%)
I	10	84.11	15.88	96.88	3.11
II	2	30.07	69.92	88.17	11.82
III	1	29.10	70.89	51.49	48.50
IV	9	68.48	31.51	77.43	22.56
V	7	72.38	27.61	83.26	16.73
VI	9	39.63	60.36	50.26	49.73
VII	12	76.79	23.20	83.28	16.71
VIII	10	64.67	35.32	69.98	30.01

4.2 Discussion

A very high percentage of the most common behaviours observed in the dataset, I and VII, are seen to be within the predicted zone, regardless of extending the zone. For some behavioural cases, like cases II and III, not many pedestrian trajectories could be found in the chosen dataset. These zones are bounded by the shortest route to the goal and the safest. Our results are a validation of this. These scores can be ameliorated by substituting better parameters to build the potential field model. For example, the *attractiveness* parameter of the POIs in the scene were assumed to be equal for all POIs in the scene. These parameters could change based on the POIs global importance, the time of day, etc., all of which could be encoded on a map for a given city. By accurately estimating the values, pedestrian behaviour in urban centers could be much better predicted. Thus, a conclusion can be reached that pedestrian behaviour is not random in nature. Their movement can be accurately predicted by utilising well established sociological ideas of *attractors* and *Natural Vision* as our work demonstrates. This helps in providing prior knowledge of the observed scene which could then be used for predicting pedestrian intentions. An effect of the primary validation is that the resultant pedestrian trajectory probability map could provide prior knowledge of hazardous zones in an urban street that can be used for safer navigation.

5 CONCLUSION

In this work, we have established a new framework for increasing the Situational Awareness of an autonomous car on urban roads. This is done by adapting the sociological principle of Natural Vision as a function of a potential field composed of different elements of the urban environment. This allows the car to understand pedestrian behaviour in previously unobserved areas. We have also demonstrated that pedestrian behaviour in urban areas is not random but is a function of the built environment they are in. The main contributions of this paper are – a) the usage of sociological principles and the integration of POIs into understanding pedestrian behaviour, b) quick computation of probable pedestrian movement zones even when there are no pedestrian observations in the scene.

The current work utilises a dataset that captures pedestrian behaviour with a mounted, stationary camera overlooking a single view. Future work will deal with the application of this framework from the ego-perspective of an autonomous car. Algorithms like the CMC-DOT [20] can be used for estimating the occupancy grid and positions of dynamic obstacles in real time. This can then be used as an input for our framework. Another work that needs to be done is to have online extraction of POIs and track individual pedestrians based on the priors determined by our framework.

ACKNOWLEDGMENT

These researches have been conducted within the VALET project, funded by the French Ministry of Education and Research and the French National Research Agency (ANR-15-CE22-0013-02).

References

[1] L. Gugerty, "Evidence from a partial report task for forgetting in dynamic spatial memory," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 40, no. 3, pp. 498–508, 1998.

[2] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 37, no. 1, pp. 32–64, 1995.

[3] B. Hillier, A. Penn, J. Hanson, T. Grajewski, and J. Xu, "Natural movement: or, configuration and attraction in urban pedestrian movement," *Environment and Planning B: planning and design*, vol. 20, no. 1, pp. 29–66, 1993.

[4] J. J. Gibson, "The ecological approach to visual perception." 1979.

[5] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[6] E. Papadimitriou, G. Yannis, and J. Golias, "A critical assessment of pedestrian behaviour models," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 12, no. 3, pp. 242–255, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.trf.2008.12.004

[7] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?" in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1345–1352.

[8] S. Pellegrini, K. Schindler, and L. van Gool, "You'll Never Walk Alone: Modeling Social Behavior for Mluti-target Tracking," *Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV)*, no. Iccv, pp. 261–268, 2009.

[9] T. Bandyopadhyay, C. Z. Jie, D. Hsu, H. Marcelo, A. Jr, D. Rus, and E. Frazzoli, "Intention-Aware Pedestrian Avoidance," *The 13th International Symposium on Experimental Robotics*, pp. 963–977, 2013.

[10] B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. Srinivasa, "Planning-based prediction for pedestrians," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. IEEE, 2009, pp. 3931–3936.

[11] K. Kitani, B. Ziebart, J. Bagnell, and M. Hebert, "Activity forecasting," *Computer Vision–ECCV 2012*, pp. 201–214, 2012.

[12] D. Vasquez, "Novel planning-based algorithms for human motion prediction," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 3317–3322.

[13] S. Bonnin, T. H. Weisswange, F. Kummert, and J. Schmuedderich, "Pedestrian crossing prediction using multiple context-based models," *2014 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014*, pp. 378–385, 2014.

[14] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," *The International Journal of Robotics Research*, vol. 5, no. 1, pp. 90–98, 1986.

[15] M. T. Wolf and J. W. Burdick, "Artificial potential functions for highway driving with collision avoidance," in *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*. IEEE, 2008, pp. 3731–3736.

[16] M. C. Montel, T. Brenac, M.-A. Granie, M. Millot, and C. Coquelet, "Urban environments, pedestrian-friendliness and crossing decisions," in *Transportation Research Board 92nd Annual Meeting*, 2013, p. 13p.

[17] R. Volpe and P. Khosla, "A theoretical and experimental investigation of impact control for manipulators," *The International Journal of Robotics Research*, vol. 12, no. 4, pp. 351–365, 1993.

[18] J. Varadarajan and J.-M. Odobez, "Topic models for scene analysis and abnormality detection," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1338–1345.

[19] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.

[20] L. Rummelhard, A. Nègre, and C. Laugier, "Conditional monte carlo dense occupancy tracker," in *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*. IEEE, 2015, pp. 2485–2490.

MobileRGBD, An Open Benchmark Corpus for mobile RGB-D Related Algorithms

Dominique Vaufreydaz, Amaury Nègre
Prima - Inria, LIG, Univ. de Grenoble, CNRS
Zirst Montbonnot, 655 avenue de l'Europe
38334 Saint Ismier cedex - France
Email: dominique.vaufreydaz@inria.fr, amaury.negre@imag.fr

***Abstract*—Since the commercialization of low cost RGB-D sensors, like the Kinect, more and more indoor robots have been equipped with this kind of sensors to perform tasks as people tracking or gesture recognition. Nevertheless, as far as we know from the literature, studies do not consider the limits of the sensors in term of motion speed, position of the sensor on the robot, etc. In this work, we propose to provide a corpus dedicated to low level RGB-D algorithms benchmarking. Originality of our approach is the use of dummies in order to play static users in the environment. This idea let us vary other variables that can impact algorithm performance: linear/angular speed of the robot, trajectory of the robot, RGB-D sensor height and vertical angle of view, number and relative position of dummies and furniture position. This paper first describes the experimental platform used to perform the acquisitions and the environment setup required to reproduce the dataset. Then, a precise description of all available data is given. We will see that, as this corpus contains a lot of configurations, it will allow researchers to investigate how these variables impact the results of their algorithms.**

I. Introduction

Body shape, gesture, body language detection and evaluation are not new areas of research. Several years ago, one can use computer vision algorithms in order to detect pedestrians, bodies or objects from a video stream. Later on, researchers have used Time-Of-Flight (TOF) cameras in order to achieve this task but these devices are expensive. Nowadays, many researches on this topic are done using Kinect of other RGB-D sensors [1].

Many robotic researches, especially but not only in researches about companion robot, are using RGB-D sensors like the Kinect for many tasks: 3D SLAM, people tracking and skeleton tracking, gesture recognition, body pose estimation, engagement toward a robot or for elderly at home, detecting falls or searching for a fallen person ([2], [3], [4], [5], [6], [7], [8]). We totally agree with [9], benchmarking is one pillar of research in these domains: how can one states that an algorithm performs better without a common reference? For instance, in [10], a fall detection system is depicted with an interesting approach and good performances but description of the train/test corpus is not detailed enough. Thus, this corpus is not reproducible and results not comparable to other approaches.

RGB-D related algorithms are difficult to benchmark as there is no reference corpora with ground truth labels for every task or/and condition. Focusing on all body related tasks, it is even more difficult to record data. Labeling depth data with skeletons is not an easy task and recording many times people enhanced variability but increase difficulty of the labeling task.

Several corpora using RGB-D are available for many tasks. One can find corpora for language signing [11] or mental state computation [12] for instance. [13] proposed a Human activities dataset for ICPR 2012 human activities recognition and localization competition. This dataset exposes humans in everyday-life tasks recording using the Kinect sensors. Labeling is done on human activities, i.e. at a high level of annotation. One cannot state performances of skeleton or face detection using this corpus even these features can be used for the activity recognition. Few corpora are available on with a mobile robot equipped with an RGB-D sensor. In [2], the RGB-D data are dedicated to SLAM. Mobile paths were done using a man handled and a robot mounted Kinect device. For the robot recordings, linear speed is from 0.1 to $0.23\ m.s^{-1}$ and angular speed is around $12\ deg.s^{-1}$. Ground truth labels are provided by a calibrated motion capture system, but no body features are provided. One can find a more correlated approach in [3]. The authors use a Kinect mounted on a cleaning robot in order to evaluate upper body skeleton tracking. People were asked to play predefined gestures in front of the robot while it is rotating. In this study, we have no information about the robot or its angular speed. Moreover, variability in this case is not only due to the speed of the robot but also to human gestures. Between two records, humans' moves are different, thus can introduce variability on detection algorithms. These artifacts can lead to different results for several speeds, not depending on the speed itself but on external variables.

In this paper, we first present our design path and objectives in section II. Section III depicts our experimental environment, i.e. our experimentation room and robotic platform. We give clues about control of the robot in order to make reproducible recording trajectories and RGB-D sensor vertical angle calibration. Gathered data using the Kinect 2 device mounted and the robot sensors are detailed in section IV.

II. Design path and objectives

From our experience of using RGB-D sensors on mobile robot, we know it could be difficult to isolate which changes/variables impact performance of a specific algorithm: is it due to the robot speed, the number of persons at the same time, their relative position, etc. Sometimes, we even improve performance by changing position or field of view of

the mounted RGB-D sensor on the robot. Looking at design of some mobile platforms like the Hobbit [14], the turtlebot[1] [15], a tobotic Wheelchair [16] or our version of the Kompaï [7], positions of using mounted RGB-D sensor differ. Most of the time, it was parallel to the floor plane. These choices are often driven by human, technical or design considerations without evaluating impact on the algorithms. From all the reasons above, we can extract a set of variables that we want to address in our corpus:

- linear and angular robot speed;
- number and position of persons in the environment;
- furniture placement and body occlusions;
- height and vertical field of view of the RGB-D sensor, i.e. looking down or up at several positions;
- body orientation toward the sensor;

We thought to reverse the corpus recording paradigm. Our goal is to facilitate ground truth annotation and reproducibility of records among speed, trajectory and environmental variations. As we want to get rid of unpredictable human moves, in our benchmark corpus, we use dummies (see 2). Interest of dummies resides in the fact that they do not move between two recordings. It is possible to record the same robot move in order to evaluate performance of detection algorithms varying speed.

Our goal is to provide a benchmark corpus for "low level" RGB-D algorithm family like 3D-SLAM, body/skeleton tracking or face tracking using a mobile robot. Using this open corpus, researchers can find a way to answer several questions:

- what is the algorithm performance in multiples conditions?
- on a mobile robot, what is the maximum linear/angular speed supported by the algorithm?
- which variables impact the algorithm?
- evaluate suitable height/angle of the mounted RGB-D sensor to reach goals: monitoring everyday live is different from searching fallen persons on the floor;
- finally, what is the performance on an algorithm with regards to others?

III. Experimental setup

A. Room description

The experimental room is a flexible space designed to be representative of a home-like environment. It has an 'L' shape with a kitchen place with a sink, a diner and a lounge space. The size of the room is 6x8 m. A view of the kitchen space is shown on the fig. 2. Except the sink, furniture can be moved *ad libitum* to reflect experimental needs.

In our setup, dummies are in the kitchen. The robot moves from the dining room toward the dummies. Some furniture is available to create different setups and placed to make some occlusions: sofa, table and chairs. Augmented reality tags are on the walls and let us have more 3D information that information gathered from the robot laser range finder (see III-B).

[1] http://www.turtlebot.com/

Fig. 1. Robotic platform used for recording: a Robulab10 mobile robot from the Robosoft compagny. It is equipped with a laser ranger finder, a Kinect 2.0 and a laptop computer.

Fig. 2. View of the experimental room. The room is designed to simulate a real living apartment. Some augmented reality tags have been placed around the two dummies.

B. Platform description

Our recording platform is depicted on fig 1. For the robotic part, we used a Robulab10[2] mobile platform with 2 lateral propulsive and 2 castor wheels. Each propulsive wheel can been control independently. The Robulab10 is equipped with a laser range finder, 8 ultrasound and 16 infrared telemeters. On its top, we mounted a flexible structure with a Kinect 2. This structure lets us change manually the height of the sensor

[2] The Robulab10 platform is sold by the Robosoft compagny http://www.robosoft.com/

and, using servomotor, one can programmatically modify the vertical angle of view. Last, a laptop is used for control and record facilities.

C. Localization and trajectory following

The robot trajectory repeatability is one of the key aspects to compare different configurations. In this sense, an absolute accurate robot localization is required. To provide such localization a map of the room is exploited in combination of the laser sensor data. The chosen solution consists on a variant of Iterative Closest Point algorithm [17] based on a point-to-line measure to compute the transformation that best matches the laser point cloud to the walls and static furniture (see Fig. 3). This localization is performed in real time and we experimentally measured a centimeter accuracy.

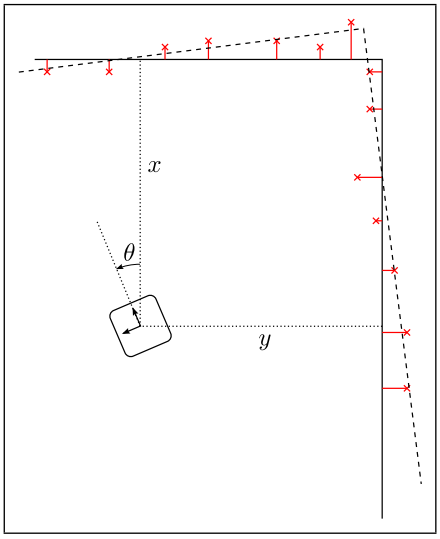


Fig. 3. Robot localization, an ICP point-to-line method is used to compute the transformation that minimize the distance between laser points and map.

Once the robot localization is known, we implement a rectilinear trajectory following algorithm. The robot command is decomposed on a linear speed and an angular speed. To control the angular speed, we first compute a control point in front of the robot, located in the linear trajectory (see Fig. 4). Then we calculate the error (ψ) as the angle between the robot direction and the control point direction. This error is injected to a PID controller to obtain a filtered angular speed command. For the linear speed control, we simply use a trapezoidal speed profile, decomposed in three steps to avoid hard acceleration. In the first step, the speed v increased linearly with the time t:

$$v(t) = a \cdot t^2$$

In the second step, the speed is constant at requested value. Last, the robot decelerates progressively, the speed decrease is computed using the square root of the distance to the goal d:

$$v = \sqrt{2 \cdot a \cdot d}$$

D. Sensor height and tilt mechanism

Height of the RGB-D device, i.e. its basement position relatively to the floor, can be modified manually. This position varies from 48 to 128 cm. The sensors are 4 cm over

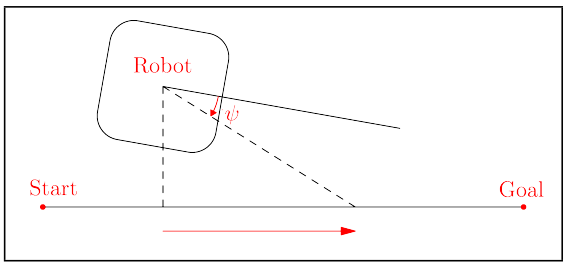


Fig. 4. Trajectory servoing.

the basement, thus, the recording positions of the mounted Kinect 2 are from 52 to 132 cm with a 10 cm step. In our hardware structure, we added a servomotor to rotate vertically the Kinect. Behind this idea, we want to investigate if it is more suitable to have a lower Kinect looking up, a central or a higher one looking straight or down on a future robot. This setup also deserves to improve search algorithm for a fallen person, i.e. we must have a floor view using the Kinect.

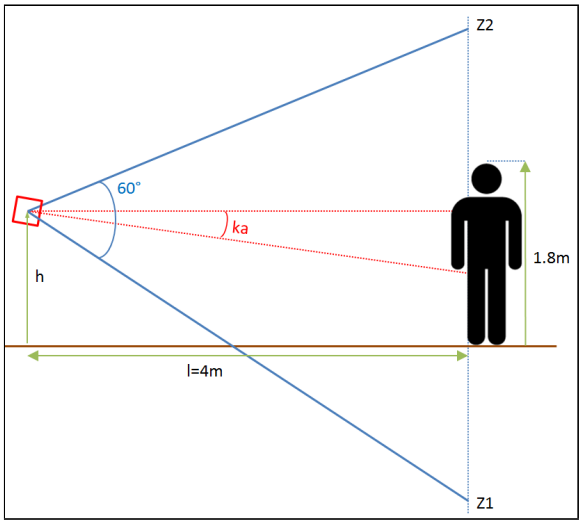


Fig. 5. From the height of the RGB-D sensor (*h*), computation of the percentage of a 1.8m person in the field of view at *l=4m* for a specific angle *ka*. *Z1* and *Z2* are respectively lowest and highest visible points at a *4m* distance. One can find an example in table I.

Step of the servomotor is $(360/4096)° < 0.1°$. To calibrate it for our specific angle set, we first extracted from the depth data the normalized coefficients of the floor plane *x*, *y*, *z* and *w* the actual height of the sensors. For each servomotor position, we computed the Kinect angle (*ka*) in regards to the floor:

$$ka = atan2(FloorPlane.z, FloorPlane.y)$$

Obviously, recording angles depend on the height of the RGB-D sensor. On lower position, we did not record floor view and on higher position we did not record ceil view. As the max depth value is *4.5m*, We decided to record every angle that permits at *4m* to see at least 50% of a *1.80m* body (see Fig 5). Using *ka*, we computed *Z1* and *Z2*, respectively the lowest and highest visible points at a *4m* distance. The percentage of visible body is then computed directly:

$$visible\ \% = ||\overline{Z1, Z2} \cap \overline{0, 1.80m}||/1.80$$

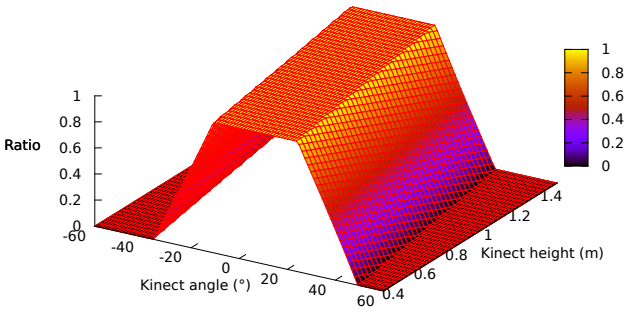


Fig. 6. Varying the height of the sensor (*h* in fig. 5), compute percentage of a 1.80m body in the field of view at *l=4m* regards to a set of angles. For each height, retained recording angle are over 50%.

Fig 6, shows all body percentage values varying *h*. In the table I, one can find example using the lowest sensor height (*h=0.52m*). In this case, we recorded using angles from $-20°$ to $35°$ using a $5°$ step. For the highest position (*h=1.32m*), retained angles are from from $-35°$ to $20°$ with the same step.

IV. Corpus acquisition

A. Scenarios

Among variables, our scenarios first must handle dummies and furniture placements within the experimental space. Dummies are set by one or two in several positions: in front of the rear wall, in front of the sink, sited on the sofa or behind a table to create partial body occlusion for instance (see fig. 2). Other variables for our recording scenarios are recording speed and path. On fig. 7, one can see some example of recording trajectories (more trajectories are tackled in the corpus). Each blue arrow represents a robot trajectory recorded at several speeds forward and backward. We do strait trajectories ending far or closer to the rear wall. Curved paths and pure rotations are also used. Last, as seen on fig. 7, trajectory at $45°$ are followed by our robot.

For each trajectory, several records are made setting the linear speed from 0.1 to 1.1 $m.s^{-1}$ with a 0.1 step. The maximum linear speed of the propulsive wheels 1.3 $m.s^{-1}$, thus we must limit the maximal speed in order to be able to make the robot turn. For the pure rotations, angular speed is set from $0.1\,\text{rad}.s^{-1}$ to $2\,\text{rad}.s^{-1}$.

We envision to record more complex paths among a full home-like setup and in some corridors.

B. Data acquisition

This project started with the release of the new Kinect 2 sensor[3]. At the recording time, OpenNI under Linux [4] does not handle correctly this new device. Thus, we choose to use the sensor under the Windows Kinect SDK. We recorded

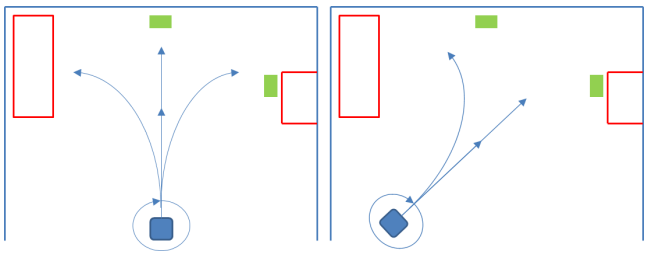


Fig. 7. Exemples of robot moves while recording. The dummies (green rectangles) and furniture (red lines) setup is the one presented on figure 2. Blue arrows represent robot trajectories.

synchronously all available streams from our mobile platform (see section III-B). Features are all robot centered.

The Kinect sensor exposes several data streams. Some of them are shown in fig. 1. Some difficulties raised when real-time acquisition of all the sensors equipping the Kinect is done. Asking for multiple streams at the same time can lead to variations on streams frame rate. For instance, when a skeleton is detected, the RGB frame rate slightly decreases. Optimization on the acquisition code was made in order to record all streams without information loss. An equal care was made to gather data from the Robulab10 platform.

Finally, we record at the highest possible frame rate several robot centered information. RGB, depth and infrared streams, detected skeletons are gathered from the Kinect 2. From the robot, laser range finder, acoustic and infrared telemeters, battery level, wheel odometry, current linear and angular speeds are monitored. All information is stored in uncompressed format in order to remove coding/decoding artifacts on video streams for instance. Each frame/event is tagged using timestamps in ms from epoch time[5] and can be synchronized for viewing or processing. The data recorded to build the corpus are summarized in the table II.

1) Videos streams: The resolution of the RGB image is of 1920×1080 pixels at 30 frames per second. The RGB sensor has a $70°$ horizontal and $60°$ vertical field of view wide-angle lens. Data are stored in 2 files. The first one is a binary raw concatenation of all YUV received images. The second one is a text file containing all timestamps associated with video frames. A compressed video file is provided for human needs. Portable C++ source code is also provided for reading frames using OpenCV [6].

2) Depth and Infrared streams: In the Kinect 2, the depth stream is a real Time-Of-Light (TOF) approach whereas previous Kinect versions that use structured light. It is also an active IR device, thus there is an IR video stream associated to the depth information. The depth range of the Kinect is limited from 0.4 meter to 4.5 meters. The depth is measured in meters from the camera along the Z axis, X and Y axis are in pixel coordinates. The resolution of the depth and IR images is of 512×424 pixels. The frame rate is at maximum 30 frames per second. As for the RGB streams, IR and depth streams are both stored using a binary raw file and associated timestamps in a separate text file.

[3]We are members of the Kinect 2 for Windows beta test program.
[4]http://www.openni.org/?

[5]The timestamps are expressed in time since 1^{st} of January, 1970.
[6]http://http://opencv.org/

Angle in degree	-25	**-20**	**-15**	**-10**	**-5**	**0**	**5**	**10**	**15**	**20**	**25**	**30**	**35**	40
Z1	-5.19	-4.25	-3.48	-2.84	-2.28	-1.79	-1.35	-0.94	-0.55	-0.19	0.17	0.52	0.87	1.23
Z2	0.87	1.23	1.59	1.98	2.39	2.83	3.32	3.88	4.52	5.29	6.23	7.45	9.10	11.51
visible %	48.33%	**68.07%**	**88.43%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**90.55%**	**71.11%**	**51.67%**	31.93%

TABLE I. KNOWING THE HEIGHT OF THE SENSOR (HERE *h=0.52m*), COMPUTATION OF *Z1*, *Z2* (SEE FIG. 5), AND PERCENTAGE OF BODY IN THE FIELD OF VIEW FOR A SET OF ANGLES. IN BOLD, RECORDING ANGLES WITH PERCENTAGE OVER 50%.

Fig. 8. Robot sensor view. At left, the laser range finder view. In the middle, the RGB stream. Last, the infrared view. The depth view (not on this figure) is aligned with the infrared one.

3) Body information: The Kinect on Windows SDK supports up to 6 skeletons tracked at the same time (it used to be 2 with the previous sensor version). Each skeleton is associated to a bag of pixels from the depth image. They now contains 25 joints (20 previously) more morphologically placed within the body shape. New joints concern neck, thumbs and hand tips. 3D information about joints is enhanced by quaternions giving rotations of body parts. We also have information about face detected in the RGB stream and/or from the depth stream. For portability, timestamped body data are saved in JSON format[7] in a text file. A portable C++ source code is provided for reading data.

4) Robulab10 sensors: As described, Robulab10 is equipped with many sensors. Even if we use the laser data for controlling the robot (see III-C), the nine ultrasound and the sixteen infrared telemeters are monitored. Wheel odometry, internal Robulab10 localization, commands sent to the robot and battery level are also gathered. All these data are saved in JSON with timestamps.

C. Labels

To make reproducible experimentation, one needs ground truth information and labels. For each experimental setup, several information are provided:

- a static map and a SLAM Based map of room and furniture;
- a 2D position and a 3D bounded box for each dummy;
- 10 seconds of full recording of each dummy are available (see section IV-B). Among this information, one can find an automatic skeleton from the Kinect and its projected information in the room space. This recording is done with a localized and still robot, using a Kinect at middle height and parallel to the ground. The robot is manually positioned at a strategic place, i.e. a place where we have an skeleton for the dummy and at least 2 visual tags (see Fig. 2).
- Another skeleton is manually annotated using the Depth data.

D. Corpus availability

The corpus is freely available for research teams through a web site[8]. As we said, it will come with C++ source code provided to read synchronously data. These source codes work under Window and Linux. We envision providing also a ROS "*bag*"[9] in order to facilitate integration of our corpus. After the final release of the Kinect 2, we plan to release our recording source code.

We have several hundreds of Gigabytes available for the community. We are preparing a web solution to conveniently distribute it. Indeed, we need to find a solution to distribute it the right way. We plan to propose a website where people can download a selection of the available data. Doing this, people may choose first subsets to achieve first experiments. Then, they can better choose other subsets in order to validate their first results.

We encourage researchers from other research teams to contribute to our effort. Using their own robot, they can use the same techniques to record other scenarios in other conditions.

V. CONCLUSION

In this paper, we have presented MobileRGBD a new freely available dataset for benchmarking RGB-D related algorithms on a mobile platform. At the writing time of this paper, we are still recording data. This corpus contains color images, depth maps and IR images, body information from the Kinect 2 and localization information of our Robulab10 platform. Originality of our corpus is the use of dummies in order to play static

[7] http://www.json.org/

[8] http://www-prima.inrialpes.fr/Vaufreydaz/OpenRGBDBenchmarkCorpus/

[9] http://www.ros.org/

Data	Sensor	Information	Frame rate
Telemeters distances	Laser range finder	20 meters maximum	12.5Hz
Ultrasound distances	Ultrasound telemeters	3 meters maximum	12.5Hz
IR distances	IR telemeters	1.5 meters maximum	12.5Hz
Commands	Control program	linear and angular speeds, stop command	12.5Hz
Odometry	Robulab10	internal localization information	12.5Hz
Battery level	Robulab10	in percentage	1Hz
Body	Kinect 2	maximum 6 skeletons with 25 joints and body parts rotations and faces Skeletons computed using OpenNi and face detection using OpenCV will be provided.	30Hz max
RGB Video	Kinect 2	1900x1080p	30Hz
Depth Video	Kinect 2	512x424p from 0.4 to 4.5m	30Hz
IR Video	Kinect 2	512x424p	30Hz

TABLE II. SUMMARY OF COLLECTED DATA IN THE CORPUS WITH THEIR CORRESPONDING SENSORS, INFORMATION AND MAXIMAL ACQUISITION FRAME RATE.

users in the environment. This idea let us vary other variables that can impact algorithm performance: linear/angular speed of the robot, trajectory of the robot, RGB-D sensor height and vertical angle of view, number and position of dummies and furniture position.

We propose a dataset that allows researchers to evaluate what is the performance of their algorithms and which variables impact their results. Knowing that, they can make more enlighten choices for the design of a future robotic platform or to solve specific problems. A multi-platform specific C++ source code is providing to facilitate the use of our data. The corpus will be available though a Web Site http://www-prima.inrialpes.fr/Vaufreydaz/OpenRGBDBenchmarkCorpus/.

ACKNOWLEDGMENT

The authors would like to thank Inria and French Ministry of Education and Researches for their support. This work was done using the Amiqual4Home facilities (ANR-11-EQPX-0002).

REFERENCES

[1] L. Cruz, D. Lucio, and L. Velho, "Kinect and rgbd images: Challenges and applications," in *Graphics, Patterns and Images Tutorials (SIBGRAPI-T), 2012 25th SIBGRAPI Conference on*, Aug 2012, pp. 36–49.

[2] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, Oct 2012, pp. 573–580.

[3] H. Kim, S.-H. Hong, S. H. Kim, P. Youn, S. Ha, and H. Myung, "Gesture recognition for moving rgb-d sensor," in *Robotics (ISR), 2013 44th International Symposium on*, Oct 2013, pp. 1–3.

[4] G. Mastorakis and D. Makris, "Fall detection system using kinect's infrared sensor," *Journal of Real-Time Image Processing*, pp. 1–12, 2012.

[5] F. Negin, F. Özdemir, C. B. Akgül, K. A. Yüksel, and A. Erçil, "A decision forest based feature selection framework for action recognition from rgb-depth cameras," in *Image Analysis and Recognition*. Springer, 2013, pp. 648–657.

[6] K. Buys, C. Cagniart, A. Baksheev, T. D. Laet, J. D. Schutter, and C. Pantofaru, "An adaptable system for rgb-d based human body detection and pose estimation," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 39 – 52, 2014, visual Understanding and Applications with RGB-D Cameras. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047320313000515

[7] W. Benkaouar and D. Vaufreydaz, "Multi-Sensors Engagement Detection with a Robot Companion in a Home Environment," in *Workshop on Assistance and Service robotics in a human environment at IEEE International Conference on Intelligent Robots and Systems (IROS2012)*, Vilamoura, Algarve, Portugal, Oct. 2012, pp. 45–52. [Online]. Available: http://hal.inria.fr/hal-00735150

[8] S. M. Anzalone, S. Ivaldi, O. Sigaud, and M. Chetouani, "Multimodal people engagement with icub," in *Biologically Inspired Cognitive Architectures 2012*. Springer, 2013, pp. 59–64.

[9] K. Berger, "The role of rgb-d benchmark datasets: an overview," *arXiv preprint arXiv:1310.2053*, 2013.

[10] A. Davari, T. Aydin, and T. Erdem, "Automatic fall detection for elderly by using features extracted from skeletal data," in *Electronics, Computer and Computation (ICECCO), 2013 International Conference on*, Nov 2013, pp. 127–130.

[11] K. Stefanov and J. Beskow, "A kinect corpus of swedish sign language signs," in *Proceedings of the 2013 Workshop on Multimodal Corpora: Beyond Audio and Video*, 2013.

[12] M. Mahmoud, T. Baltrušaitis, P. Robinson, and L. D. Riek, "3d corpus of spontaneous complex mental states," in *Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 205–214.

[13] C. Wolf, J. Mille, E. Lombardi, O. Celiktutan, M. Jiu, M. Baccouche, E. Dellandrea, C.-E. Bichot, C. Garcia, and B. Sankur, "The liris human activities dataset and the icpr 2012 human activities recognition and localization competition," March 2012. [Online]. Available: http://liris.cnrs.fr/voir/activities-dataset/

[14] D. Fischinger, P. Einramhof, W. Wohlkinger, K. Papoutsakis, P. Mayer, P. Panek, T. Koertner, S. Hofmann, A. Argyros, M. Vincze *et al.*, "Hobbit-the mutual care robot," in *Workshop on Assistance and Service Robotics in a Human Environment Workshop in conjunction with IEEE/RSJ International Conference on Intelligent Robots and Systems, 2013*, 2013.

[15] C. Xiong and X. Zhang, "An exclusive human-robot interaction method on the turtlebot platform," in *Robotics and Biomimetics (ROBIO), 2013 IEEE International Conference on*, Dec 2013, pp. 1402–1407.

[16] D. Vasquez, P. Stein, J. Rios-Martinez, A. Escobedo, A. Spalanzani, and C. Laugier, "Human aware navigation for assistive robotics," in *Experimental Robotics*, ser. Springer Tracts in Advanced Robotics, J. P. Desai, G. Dudek, O. Khatib, and V. Kumar, Eds. Springer International Publishing, 2013, vol. 88, pp. 449–462. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-00065-7_31

[17] A. Censi, "An icp variant using a point-to-line metric," in *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*. IEEE, 2008, pp. 19–25.

Figurines, a multimodal framework for tangible storytelling

Maxime Portaz[1], Maxime Garcia[2], Adela Barbulescu[2], Antoine Begault[2], Laurence.Boissieux[1], Marie-Paule Cani[2,3], Rémi Ronfard[2], Dominique Vaufreydaz[1]

[1] Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP*, LIG, 38000 Grenoble, France

[2] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LJK, 38000 Grenoble, France

[3] Ecole Polytechnique, CNRS, Université Paris-Saclay, LIX, 91128 Palaiseau, France

Dominique.Vaufreydaz@inria.fr

Author version

Abstract

This paper presents *Figurines*, an offline framework for narrative creation with tangible objects, designed to record storytelling sessions with children, teenagers or adults. This framework uses tangible diegetic objects to record a free narrative from up to two storytellers and construct a fully annotated representation of the story. This representation is composed of the 3D position and orientation of the figurines, the position of decor elements and interpretation of the storytellers' actions (facial expression, gestures and voice). While maintaining the playful dimension of the storytelling session, the system must tackle the challenge of recovering the free-form motion of the figurines and the storytellers in uncontrolled environments. To do so, we record the storytelling session using a hybrid setup with two RGB-D sensors and figurines augmented with IMU sensors. The first RGB-D sensor completes IMU information in order to identify figurines and tracks them as well as decor elements. It also tracks the storytellers jointly with the second RGB-D sensor. The framework has been used to record preliminary experiments to validate interest of our approach. These experiments evaluate figurine following and combination of motion and storyteller's voice, gesture and facial expressions. In a make-believe game, this story representation was re-targeted on virtual characters to produce an animated version of the story. The final goal of the *Figurines* framework is to enhance our understanding of the creative processes at work during immersive storytelling.

Index Terms:Puppetry, Storytelling, Multimodal data fusion, RGB-D sensor, IMU sensor

1 Introduction

Storytelling is the art of creating or sharing a narrative. As famously emphasized by French structuralist Roland Barthes, forms of narrative are numerous and diverse

* Institute of Engineering Univ. Grenoble Alpes

Figure 1: View from a storytelling session with a 6-year old child using the *Figurines* framework. At left, the frontal view with body and face detections (eyes, ears, nose, neck, shoulders, arms, hands) and facial landmarks of the storyteller using OpenPose [1]. At right, top view with body and hand detection, and decor tracking (blue points). Tracking of figurines is done using a hybrid IMU/RGB-D approach.

around the world, and they always played an important role in human society [2]. Narrative is widely considered to be a fundamental part of human cognition and understanding [3]. Today, storytelling is increasingly performed digitally and it is becoming easier to create stories using available digital technologies [4]. Storytelling is especially important for children education by supporting and enhancing creative expression and learning, as well as encouraging team work and sharing of personal experience [5, 6, 7].

Several authoring tools exist to help a narrator develop and construct his story [8]. In the film and game industries, motion capture is often used to create the 3D representation of movements or object manipulations. Existing motion capture systems generally require expensive hardware setup and/or attached marker [9]. They are not usable by the general public, especially by young children. The emergence of RGB-D sensors (like the Kinect) and of Fab-Labs with their new widespread technologies (3D printing, laser cutting ...) lets people envision developing lighter and cheaper systems. In parallel, the increasing performance of human perception algorithms thanks to Deep Learning [1, 10, 11, 12] tends to improve human analysis capabilities of interaction systems. Even if interaction systems take advantages of these recent progresses, there are still challenging problems to address. A first challenge concerns the underlying algorithms for object tracking and occlusion. These problems, still under investigation in many laboratories [13], are addressed in our system by a hybrid IMU/RGB-D approach. This challenge is not in the main scope of this paper and will not be detailed. A second challenge is to develop these tools while allowing rich narrative creation, with free character expressions and motion. The intrusiveness of the acquisition system must be reduced to a minimum, so that it does not disrupt the narrative flow of the storyteller. This challenge is the core of our system design.

This paper presents *Figurines* , a narrative capture/playback system for adults and children. Its first goal is to record one or two narrators while they are telling and playing a story. A story is composed of animated characters, evolving with rigid decor elements, in a given scene. In order to provide a natural interaction and to help the storytellers to be immersed in the story world, we propose to use tangible interaction with *diegetic* objects like figurines and decor elements. Diegetic objects are part of the story and they are present (i.e. exist) in the story, and not just tools or symbols artificially

added to represent the story components. The second goal of the framework is to produce a complete synchronous representation of the story: 3D position and orientation of figurines, position of the decor elements and acting from the storytellers. One future purpose of this information is to analyze playing sessions to understand storytelling mechanisms. In our first experiments, this information was used to produce a 3D animated movie from the storytelling session. The final goal of the *Figurines* framework is, analyzing these session data, to enhance our understanding of the creative processes at work during immersive storytelling.

In Section 2, we present existing systems and related work. In Section 3, we explain our design and implementation. Finally, in Section 5, we present preliminary experimental results and evaluations made to test our system. Several storytelling sessions with one or two child narrators are analyzed. Limitations and lessons learned are discussed and conclusions are drawn.

2 Related Work

During the past years, many systems have been developed. Harley et al. [8] propose a survey and a classification of these systems. Different tangible interfaces exist to help children to experiment storytelling. Sylla et al. [14] and Chu et al. [15] present storytelling tangible interfaces for children. These systems use tangible objects as story components. They are suitable for children but they limit possible interaction and story structure. Character animation and orientation have been investigated with fully augmented puppets [16, 17] or even using the body of the puppeteer [18]. These interfaces allow precise and accurate control of the character. Tangible interfaces also exist with a simple webcam to track moves of a toy robot [19]. A large amount of work is needed to create the puppet and make it available to children. *ShadowStory* and *iTheater* [20, 21] let children animate puppets using accelerometer enabled devices.

Recently, several storytelling systems have taken advantage of RGB-D sensors. *PuppetX* [22] uses skeleton or hand/finger motions acquired from an RGB-D sensor. It retargets animation on a specific articulated puppet with servo-motors using manually defined rules to match between moves and puppet degrees of freedom. *3D puppetry* [23] takes advantage of the depth and associated color data. Using rigid colorized 3D models, it tracks poses of so-called puppets (cars, boats, etc.) to compute their 3D positions in the scene. *MotionMontage* [24] uses a similar approach to animate a virtual object along a path using a physical rigid object. One problem with purely vision-based systems is that their performance degrades rapidly in cases of occlusion. The storyteller must be very careful to keep the puppet visible from the camera at all times while acting. One way to do so is to provide feedback to the storyteller but this may interfere with the narration fluidity, moreover for children. Narrators may look to the feedback screen regardless of their narrative goals. To tackle this problem in the *Figurines* framework, we decided to combine IMU (Inertial Motion Unit) sensors with an RGB-D sensor in order to reduce occlusion problems as much as possible. This choice leads us to address the drifting problem of IMU path reconstruction with improvement over existing algorithms (see section 4.1). Similarly to *PuppetX* [22] and to *i-marionette* [18], we also want to benefit from storyteller body language to increase relevance of the resulting animation. Therefore, we include a second RGB-D camera to track the storytellers and record their body poses, facial expressions and voices.

3 Design and Implementation

In this section, we first define the recording scenario, prior of the design of our system. Then, the acquisition setup is presented. Last, design of the narrative elements and gathered data about figurines and storytellers are described.

3.1 Recording scenario

To increase playfulness, the system must be non intrusive and must not impose narrative schemes. The storytellers can use any object of any shape to play their story. The first type of objects is called *figurine* in the framework. A figurine is an object of importance in the story (prince, animal, car...). The second type of objects is *decor element*. As far as they fit into the playground (see 3.2), any object can be a decor element.

The recording scenario of the narrative session is quite simple. The narrators place themselves in front of the recording table. Several decor elements and different figurines are available:

1. storytellers choose among decor items and figurines to play with;
2. they start to organize the stage at their convenience to tell the story;
3. they can freely play their story as long as they want.

To increase the recreational dimension of the storytelling session, all calibrations are done off-line without the narrators. No specific action is mandatory from the narrator to help the system. This property is obviously even more important for children.

3.2 Acquisition setup

The acquisition system records everything from the augmented figurines, the storytellers and the decors. In its current configuration, due to space constraint and camera view angle, the system handles at maximum two simultaneous narrators and a 70cm x 70cm playing area. The full setup is shown on figure 2. It tracks figurines on the stage and the storytellers with several devices. An overhead Kinect and a set of Inertial Motion Units (IMU) track the figurines. The top RGB-D device is accurate enough for tracking moves and distance of figurines, and decor elements on the stage. The frontal Kinect complete the top one to record the storytellers and their behaviors (see fig. 2 and 1).

As stated before, the main difficulty to address is occlusion problems. Occlusions can be caused by the storyteller himself, by another figurine, or by a decor element. Figurines can disappear from the overhead camera. That's why we included IMUs in our setup (see 3.3). This brings additional advantages over previous works: it is not mandatory to scan the figurines before tracking, and most importantly, we can use non-rigid objects (articulated, dressed and/or soft puppets for instance).

The acquisition setup records lots of raw data. All the streams from the RGB-D devices are recorded synchronously in uncompressed format at full frame rate: RGB, depth, infrared, skeletons, faces and audio streams. All IMUs information is synchronously stored in the system. Due to technical reason, data processing has to be done off-line after the recording. Data from the IMUs cannot be synchronized on-line, due to the low power Bluetooth 4.0 that could not transfer data when the acquisition

Figure 2: Acquisition setup. There are two Kinect devices: one looking down to follow figurines, one following narrators. The playground area is the table in the middle. Its size in our experiment is 70cm x 70cm.

frame rate is 200Hz. Moreover, real time processing would cause a limited choice for computer vision algorithms used in the framework. As on-line processing is not mandatory for the storyteller, and would not improve the session, it is not a constraint in the framework.

3.3 Figurine and decor design

In the last years, new widespread and affordable digital fabrication technologies are available for researchers but also, within Fab-Labs, for the general public. We decided to take benefit from these technologies to build personalized figurines and decor elements. As can be seen, we created a specific box to serve as basement for 3D printed figurine (figures 3 and 4) or as IMU container for enhanced usual puppet (figure 5). One benefit of this process is that we can create upon request almost any figurine. Another benefit arises when we produce a 3D movie from the storytelling session (see figure 6). The printing models can be reused for the 3D rendering.

As said, figurines are of importance for the story. This justifies the need for a fine tracking. Figurines are thus augmented with an IMU to improve its monitoring (figures 3 and 4). Any puppet or toy that can be enhanced with an IMU can be integrated in a story. Even decor elements, if they need to be active part of the story can be equipped and become a figurine. For instance, Figure 5 in Appendix shows a handmade car build by a child with bricks. Preliminary experiments with different consumer

Figure 3: At left, 3D models used to print an enhanced princess figurine. At right, the result dressed figurine.

Figure 4: Figurine examples. Left, the IMU in the basement of the soldier figurine. Center and right, the soldier and prince dressed figurines respectively.

IMUs demonstrated that even a 50Hz frame rate is not reliable enough to reconstruct 3D path because of acceleration information sparsity and drift. Among professional available IMUs, we selected the 10-degrees-of-freedom Hikob Fox IMU[1]. This choice was driven by several technical aspects. These sensors are very light (~20 gr with their embedded battery, memory card and printed basement). They are able to record gyroscope, accelerometer and magnetometer data at high frame rate (up to 200Hz) on their memory card. Their storage and their rechargeable battery allow an autonomy of several hours. Last crucial point, using wireless synchronization, all IMUs share a common time reference. For decor elements, everything that fit the playing area can be used and tracked by the framework. In our prototype, we designed decor objects with a laser cutter (see figure 2). These decor elements fit perfectly with the figurines, as their scale has been chosen accordingly.

[1] `http://www.hikob.com/wp-content/uploads/2015/06/HIKOB_FOX_ProductSheet_EN.pdf` (last seen 07/2017)

4 Storytelling data

This section describes data computed on the figurines and the decor elements on stage, and about the storytellers. All the processing described here are offline after the recording of the narrative session. In this article, we do not detail underlying mathematics and algorithms used in the framework but present the general principle.

4.1 Stage tracking

Using the overhead RGB-D device, it is possible to track decor elements. The first step is to use the depth data to do their automatic detection over the playground. This automatic detection can be corrected at any time while processing for mis-detected objects. In a second step, the system tracks moves from decors elements using a dense optical flow. In the current implementation, decor tracking follows center of objects in 2D position (x,y), that is, the system tracks them but does not provide neither their orientation nor their height ($z = 0$). Once again, if this information is mandatory for the storyteller (a magic tree for instance), the decor element can be transformed into a figurine.

One can say that the figurine tracking is a hybrid IMU/RGB-D algorithm. The main tasks to address are figurine identification, tracking in 3D space and computing orientation over time. For the identification task, when a new mobile object is detected in the depth data, it is compare to synchronous IMUs data of actual moving figurines. If one figurine matches the current motion, its label is tagged over the mobile object. For the tracking aspect, as far as a figurine can be seen in the Kinect view, standard tracking paradigm using a Kalman filter is applied. When a figurine disappeared (under a decor element, hidden by the narrator hand, out of the camera view, ...), the only available information are gathered from IMUs. The IMUs give us the acceleration and the orientation, so we can compute the position from the last known position integrating twice the acceleration. A drifting problem appears fast with this method. Our implementation corrects the drift using an improved version of Neto's algorithm [25] and let us reconstruct the 3D path of the figurine until it is identifiable again. Finally, using magnetometer and gyroscope information, the figurine orientation is computed using Madgwick's method [26].

4.2 Storytellers' information

The tracking of the storytellers is done using both RGB-D devices. As seen on figure 1, the recordings include body tracking (frontal and from top), face tracking and sound. Body tracking is limited to upper-body tracking (head, shoulders, torso, arms, wrist and hands). The body tracking algorithm uses a modified version of the *Realtime Multi-Person Pose Estimation* algorithm [10]. Faces and hands are tracked using OpenPose [1, 10, 11, 12]. Using this software, we are able to compute facial landmark, head pose, eyes and facial Action Units. The recorded sound is tagged into voice segments. Optionally, speaker diarization can be applied to partition the audio stream according to the storyteller identity [27]. Using such a system could improve information gathered by the system and the speech slot association with figurines within the storytelling session.

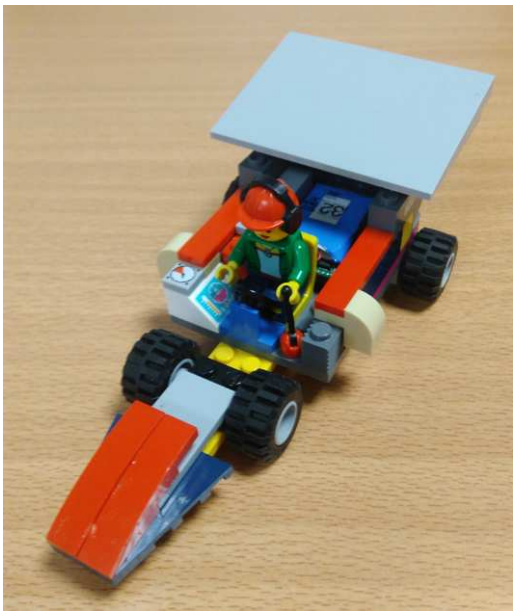


Figure 5: Child handmade Lego© car with an IMU (behind the driver's seat).

5 Evaluation of narrative session recordings

Preliminary evaluations were conducted in a dedicated room. According to the scenario (see section 3.1), we filmed narrative sessions with several users to check the usability of the system. We conducted experiments with children to verify simplicity and usability of the system: two children play the storyteller role, individually then together. We also wanted to check that narrative process is not disturbed by the recording system. Several sessions with adults have been made during the development phase. As the room does not have one-way mirror, the examiner remained in the room with the narrators for the entire session. Observations about these sessions are discussed in the following paragraphs.

In this first experiment, the storytellers did not model their own characters. We tried to provide them with enough different figurines recurring in folktales: the prince, the princess, the witch, the soldier/knight, the dragon, the horse and the wolf. We provided standard decor elements like several different trees, bridge, house, tower and etc.

5.1 Children sessions

Three narrative sessions with two children were recorded. Each child records one session alone (6 minutes for the 6-year old boy, 20 minutes for the 7-year old girl). They perform also an 11 minutes narrative session together.

The system was able to reconstruct the figurines movements and to track the decor elements. The main problems come from the storyteller tracking. The children faces were often lost on the video, due to three reasons:

1. To take advantage of the whole scene, children have to stretch their arm and to come closer to the table. Doing so, they are hidden by their arm (see fig 1) or too close to the RGB-D device to get depth data.

2. The children are sometimes hidden by the decor (the tower in our experiment).

3. In one session, the child tells his story looking at the experimenter, thus his face in profile was not detected.

For these children sessions, we recovered the upper body information 94.03% of time. The percentage of face detection is also 74.36% over time (only 11.48% using the standard Kinect2 face detection). There is no significant difference with sessions with one or two children.

Regarding the storytelling aspects, children seem not to be disturbed by the acquisition system. With one child, the presence of the examiner was a discomfort. As said, the child looked at the examiner and tended to explain the story to him, instead of playing each character role. In further experiments, we will equip the room with a one-way mirror to solve that problem.

5.2 Adult sessions

Four adults played with all the figurines and the scenery we built for a total of 13 minutes. The system was able to track each figurine, the stage and the storyteller. The face was detected and tracked with more accuracy than with children (98.02% of the time). Adults have longer arm and do not need to get closer to the table. The upper body is tracked 100% of the time. Contrary to children, in our experiments, adults have much more difficulties to imagine stories with imposed characters. As we did not let them print or construct their own figurines for these preliminary experiments, some of the participants expressed discomfort using the provided figurines. Partly for this reason, all adult storytelling sessions are shorter.

5.3 Make-believe games

We also used our framework to record imaginary dialogues for a make-believe game [28]. The storyteller's voice, gestures and facial expressions combined with movements of instrumented figurines were transferred to virtual characters in order to obtain an animated version of the dialogue. A rendering example is presented in figure 6 and in this video https://hal.inria.fr/hal-01518981v2/file/wicedcrc.mp4.

6 Lessons learned

Figurines our storytelling framework, benefits from RGB-D sensors and IMU technologies. However, it has intrinsic limitations. Due to the setup (figure 2) and the Kinect angle of view, a maximum of 2 storytellers can be recorded at a time. The number of figurines has been limited to 4 in all our experiments. This is not a strict limitation but one can figure out that increasing the number of figurines may lead to less accurate figurine identification. For the RGB-D acquisition system, we learned some lessons. The frontal Kinect must be higher and further away from the stage. It will overhang the decor elements and prevent from occlusions. This will improve perception of child storytellers.

The 3D reconstruction is efficient in our context but it is not perfect. Even small collisions of the figurines with solids have a huge effect on the measured IMU accelerations. These perturbations cannot trivially be filtered. 3D path reconstruction is actually under investigation. Manual corrections may be needed to improve quality of

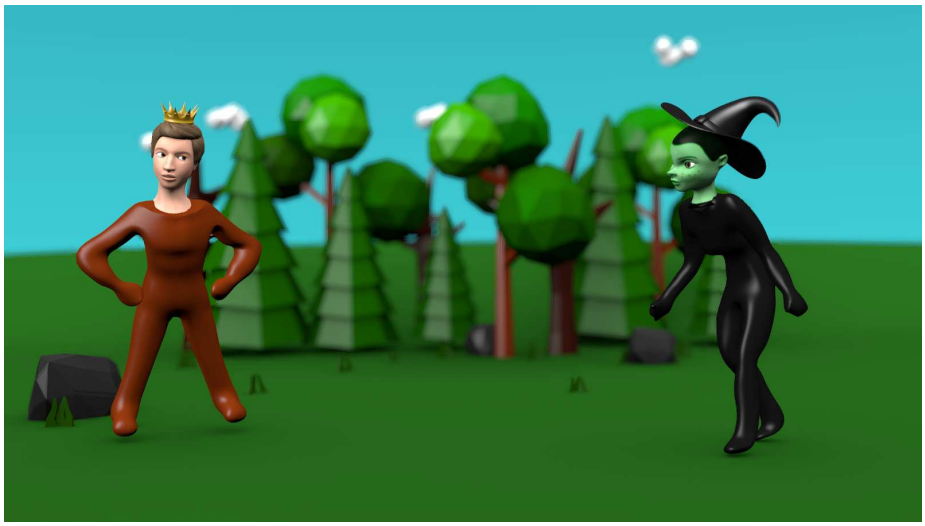


Figure 6: 3D rendering example generated from a recorded session of a make-believe game.

the 3D path reconstruction from the recordings. In our case it is not an issue, as artistic corrections can also be done to improve expressiveness in the final rendering like accentuating some moves or expressions for instance.

7 Conclusion

We have presented *Figurines* a hybrid multimodal framework for recording and playback of storytelling sessions involving tangible interaction with objects. In this framework, tangible objects are figurines enhanced with IMUs and decor elements. Figurines can be articulated, dressed and/or soft puppets. The system records the storytelling sessions using two RGB-D sensors. The overhead RGB-D sensor follows the figurines on the stage using computer vision algorithms in combination with IMUs. Up to two storytellers are monitored with a frontal RGB-D sensor. The system records their facial expressions, upper body motion and voice activity. Output of the system is a synchronous representation of the story: 3D position and orientation of figurines, position of the decor elements and interpretation of the storytellers. This information can be used as input for a 3D rendering system to produce a video animation of the story. Our framework can already be used to create multimodal recording of make-believe games, and we hope this will enhance our understanding of the creative processes at work during immersive storytelling.

In future work, we would like to let users, both children and adults, freely design and print their own story worlds, including sets, props and characters [29] and turn their stories into movies using intelligent tools for 3D animation [30] and cinematography [31]. A variation of the *Figurines* framework has also been used in an experiment to monitor players while solving Chess problem [32]. In this setup, the gathered data are used to infer mental state and chess level of the players.

8 Acknowledgments

The authors would like to thank the experimentation participants who contributed by sharing their stories. We want to thank the SED team for their technical support. Prototyping was done using the Amiqual4Home facilities (ANR-11-EQPX-0002). This work was partly funded by the PERSYVAL-Lab (ANR-11-LABX-0025-01) Labex.

References

[1] (2017) Openpose: A real-time multi-person keypoint detection and multi-threading c++ library. [Online]. Available: https://github.com/CMU-Perceptual-Computing-Lab/openpose

[2] R. Barthes, "An introduction to the structural analysis of narrative," *New literary history*, pp. 237–272, 1975.

[3] K. Newman, "The case for the narrative brain," in *Proceedings of the second Australasian conference on Interactive entertainment*. Creativity & Cognition Studios Press, 2005, pp. 145–149.

[4] J. Van Dijck, "Users like you? theorizing agency in user-generated content," *Media, culture, and society*, vol. 31, no. 1, p. 41, 2009.

[5] B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes, and L.-P. Morency, "The future belongs to the curious: Towards automatic understanding and recognition of curiosity in children," in *Workshop on Child Computer Interaction*, 2016, pp. 16–22.

[6] J. A. Fails, A. Druin, and M. L. Guha, "Interactive storytelling: interacting with people, environment, and technology," *International Journal of Arts and Technology*, vol. 7, no. 1, pp. 112–124, 2014.

[7] S. Benford, B. B. Bederson, K.-P. Åkesson, V. Bayon, A. Druin, P. Hansson, J. P. Hourcade, R. Ingram, H. Neale, C. O'Malley *et al.*, "Designing storytelling technologies to encouraging collaboration between young children," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 2000, pp. 556–563.

[8] D. Harley, J. H. Chu, J. Kwan, and A. Mazalek, "Towards a framework for tangible narratives," in *Proceedings of the TEI'16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction*. ACM, 2016, pp. 62–69.

[9] D. J. Sturman, "Computer puppetry," *Computer Graphics and Applications, IEEE*, vol. 18, no. 1, pp. 38–45, 1998.

[10] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[11] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *CVPR*, 2017.

[12] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016.

[13] M. Camplani, S. Hannuna, M. Mirmehdi, D. Damen, A. Paiement, L. Tao, and T. Burghardt, "Real-time rgb-d tracking with depth scaling kernelised correlation filters and occlusion handling," in *Proceedings of the British Machine Vision Conference (BMVC). pp*, 2015, pp. 145–1.

[14] C. Sylla, S. Gonçalves, P. Brito, P. Branco, and C. Coutinho, "A tangible platform for mixing and remixing narratives," in *Advances in Computer Entertainment*. Springer, 2013, pp. 630–633.

[15] J. H. Chu, P. Clifton, D. Harley, J. Pavao, and A. Mazalek, "Mapping place: Supporting cultural learning through a lukasa-inspired tangible tabletop museum exhibit," in *Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction*. ACM, 2015, pp. 261–268.

[16] W. Yoshizaki, Y. Sugiura, A. C. Chiou, S. Hashimoto, M. Inami, T. Igarashi, Y. Akazawa, K. Kawachi, S. Kagami, and M. Mochimaru, "An actuated physical puppet as an input device for controlling a digital manikin," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 637–646.

[17] A. Mazalek and M. Nitsche, "Tangible interfaces for real-time 3d virtual environments," in *Proceedings of the international conference on Advances in computer entertainment technology*. ACM, 2007, pp. 155–162.

[18] S.-Y. Lin, C.-K. Shie, S.-C. Chen, and Y.-P. Hung, "Action recognition for human-marionette interaction," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 39–48.

[19] R. Slyper, G. Hoffman, and A. Shamir, "Mirror puppeteering: Animating toy robots in front of a webcam," in *Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction*. ACM, 2015, pp. 241–248.

[20] F. Lu, F. Tian, Y. Jiang, X. Cao, W. Luo, G. Li, X. Zhang, G. Dai, and H. Wang, "Shadowstory: creative and collaborative digital storytelling inspired by cultural heritage," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 1919–1928.

[21] O. Mayora, C. Costa, and A. Papliatseyeu, "itheater puppets tangible interactions for storytelling," in *International Conference on Intelligent Technologies for Interactive Entertainment*. Springer, 2009, pp. 110–118.

[22] S. Gupta, S. Jang, and K. Ramani, "Puppetx: a framework for gestural interactions with user constructed playthings," in *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*. ACM, 2014, pp. 73–80.

[23] R. Held, A. Gupta, B. Curless, and M. Agrawala, "3d puppetry: a kinect-based interface for 3d animation." in *UIST*. Citeseer, 2012, pp. 423–434.

[24] A. Gupta, M. Agrawala, B. Curless, and M. Cohen, "Motionmontage: A system to annotate and combine motion takes for 3d animations," in *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI '14. New York, NY, USA: ACM, 2014, pp. 2017–2026. [Online]. Available: http://doi.acm.org/10.1145/2556288.2557218

[25] P. Neto, J. N. Pires, and A. P. Moreira, "3-d position estimation from inertial sensing: minimizing the error from the process of double integration of accelerations," *CoRR*, vol. abs/1311.4572, 2013. [Online]. Available: http://arxiv.org/abs/1311.4572

[26] S. Madgwick, "An efficient orientation filter for inertial and inertial/magnetic sensor arrays," *Report x-io and University of Bristol (UK)*, 2010.

[27] M. Najafian and J. H. Hansen, "Speaker independent diarization for child language environment analysis using deep neural networks," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 114–120.

[28] A. Barbulescu and M. Garcia and A. Begault and M. P. Cani and M. Portaz and A. Viand and R. Dulery and L. Boissieux and P. Heinish and R. Ronfard and D. Vaufreydaz, "A system for creating virtual reality content from make-believe games," in *2017 IEEE Virtual Reality (VR)*, March 2017, pp. 207–208.

[29] M. Skouras, B. Thomaszewski, S. Coros, B. Bickel, and M. Gross, "Computational design of actuated deformable characters," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, p. 82, 2013.

[30] J. Chai and J. K. Hodgins, "Performance animation from low-dimensional control signals," *ACM transactions on Graphics, Proceedings of SIGGRAPH*, pp. 686–696, 2005.

[31] Q. Galvane, R. Ronfard, M. Christie, and N. Szilas, "Narrative-driven camera control for cinematic replay of computer games," in *Proceedings of the Seventh International Conference on Motion in Games*, ser. MIG '14. ACM, 2014, pp. 109–117.

[32] T. Guntz, D. Vaufreydaz, R. Balzarini, and J. Crowley, "Multimodal observation and interpretation of subjects engaged in problem solving," in *1st Behavior, Emotion and Representation: Building Blocks of Interaction Workshop at 5th International Conference on Human-Agent Interaction*. ACM, 2017.

Multimodal Observation and Interpretation of Subjects Engaged in Problem Solving

Thomas Guntz
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LIG, F-38000 Grenoble, France
Thomas.Guntz@inria.fr

Raffaella Balzarini
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LIG, F-38000 Grenoble, France
Raffaella.Balzarini@inria.fr

Dominique Vaufreydaz
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LIG, F-38000 Grenoble, France
Dominique.Vaufreydaz@inria.fr

James Crowley
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LIG, F-38000 Grenoble, France
James.Crowley@inria.fr

ABSTRACT

In this paper we present the first results of a pilot experiment in the capture and interpretation of multimodal signals of human experts engaged in solving challenging chess problems. Our goal is to investigate the extent to which observations of eye-gaze, posture, emotion and other physiological signals can be used to model the cognitive state of subjects, and to explore the integration of multiple sensor modalities to improve the reliability of detection of human displays of awareness and emotion. We observed chess players engaged in problems of increasing difficulty while recording their behavior. Such recordings can be used to estimate a participant's awareness of the current situation and to predict ability to respond effectively to challenging situations. Results show that a multimodal approach is more accurate than a unimodal one. By combining body posture, visual attention and emotion, the multimodal approach can reach up to 93% of accuracy when determining player's chess expertise while unimodal approach reaches 86%. Finally this experiment validates the use of our equipment as a general and reproducible tool for the study of participants engaged in screen-based interaction and/or problem solving.

KEYWORDS

Chess Problem Solving, Eye Tracking, Multimodal Perception, Affective Computing

1 INTRODUCTION

Commercially available sensing technologies are increasingly able to capture and interpret human displays of emotion and awareness through non-verbal channels. However, such sensing technologies tend to be sensitive to environmental conditions (**e.g.** noise, light exposure or occlusion), producing intermittent and unreliable information. Techniques for combining multiple modalities to improve the precision and reliability of modeling of awareness and emotion are an open research problem. Only few researches have been conducted so far on how such signals can be used to inform a system about cognitive processes such as situation awareness, understanding or engagement. For instance, some researches showed that mental states can be inferred from facial expressions and gestures (from head and body) [1, 2].

Willing to increase focus on this area of research, we have constructed an instrument for the capture and interpretation of multimodal signals of humans engaged in solving challenging problems. Our instrument, shown in figure 2, captures eye gaze, fixations, body postures and facial expressions signals from humans engaged in interactive tasks on a touch screen. As a pilot study, we have observed these signals for players engaged in solving chess problems. Recordings are used to estimate subjects' understanding of the current situation and their ability to respond effectively to challenging tasks. Our initial research question for this experiment was:

- *Can our experimental set up be used to capture reliable recordings for such study?*

If successful, this should allow us to a second research question:

- *Can we detect when chess players are challenged beyond their abilities from such measurements?*

In this article, section 2 discusses current methods for capture and interpretation of physiological signs of emotion and awareness. This lays the ground for the design of our experimental setup presented in section 3. Section 4 presents the results from our pilot experiment that was undertaken to validate our installation and evaluate the effectiveness of our approach. We conclude with a discussion on limitations and further directions to be explored in section 5.

2 STATE OF THE ART

Humans display awareness and emotions through a variety of non-verbal channels. It is increasingly possible to record and interpret information from such channels. Thank to progress in related research, notably recently using Deep Learning approaches [3–6], publicly available efficient software can be used to detect and track face orientation using commonly available web cameras. Concentration can be inferred from changes in pupil size [7]. Measurement of physiological signs of emotion can be done by detection of Facial Action Units [8] from both sustained and instantaneous displays (micro-expressions). Heart rate can be measured from the Blood Volume Pulse as observed from facial skin color [9]. Body posture and gesture can be obtained from low-cost RGB sensors with depth information (RGB+D) [10]. Awareness and attention can be inferred from eye-gaze (scan path) and fixation using eye-tracking glasses as well as remote eye tracking devices [11]. This can be directly used to reveal cognitive processes indicative of expertise [12] or situation awareness in human-computer interaction (HCI) systems

* Institute of Engineering Univ. Grenoble Alpes

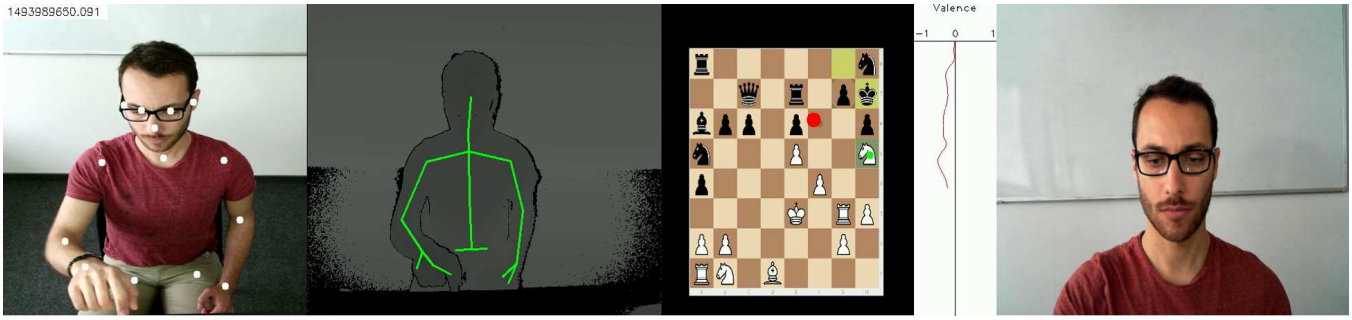


Figure 1: Multimodal view of gathered data. Left to right: RGB (with body joints) and depth view from Kinect 2 sensors, screen record of chess task (red point is current position of gaze, green point is position of last mouse click), plot of current level of positive emotion expression (valence) and frontal view of face from webcam sensor.

[13]. However, the information provided by each of these modalities tends to be intermittent, and thus unreliable. Most investigators seek to combine multiple modalities to improve both reliability and stability [14, 15].

Chess analysis has long been used in Cognitive Science to understand attention and to develop models for task solving. In their study [12, 16], Charness *et al* showed that when engaging in competitive game, chess players display engagement and awareness of the game situation with eye-gaze and fixation. This suggests that the mental models used by players can be at least partially determined from eye gaze, fixation and physiological response. The ability to detect and observe such models during game play can provide new understanding of the cognitive processes that underlay human interaction. Experiments described in this article are the preamble to more advanced research on this topic.

3 EXPERIMENTS

As a pilot study, chess players were asked to solve chess tasks within a fixed, but unknown, time frame. We recorded eye gaze, facial expressions, body postures and physiological reactions of the players as they solved problems of increasing difficulty. The main purpose is to observe changes in their reactions when presented tasks are beyond their level.

3.1 Materials and Participants

3.1.1 *Experimental setup.*

Figure 2 presents the recording setup for our experiment. This setup is a derivative version of the one we use to record children during storytelling sessions [17]. As seen, it is composed of several hardware elements: a 23.8" Touch-Screen computer, a Kinect 2.0 mounted 35cm above the screen focusing on the chess player, a 1080p Webcam for a frontal view, a Tobii Eye-Tracking bar (Pro X2-60 screen-based) and two adjustable USB-LED for lighting condition control. The use of the Touch-Screen during the entire experiment was chosen to provide a gesture-based play resembling play with a physical board. A wooden super-structure is used to rigidly mount the measuring equipment with respect to the screen in order to assure identical sensor placement and orientation for all recordings. This structure have been made using a laser cutter.

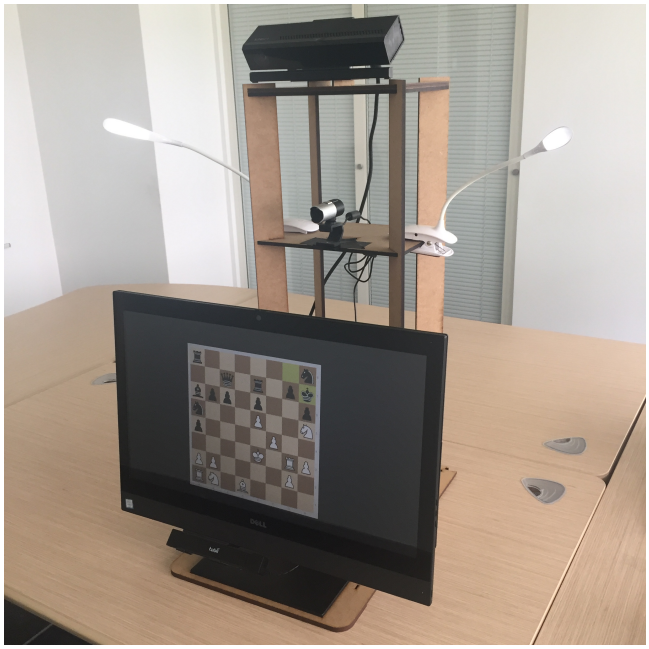


Figure 2: The experimentation equipment used for data collection. On top, a Kinect2 device looking down at the player. In the middle, a webcam to capture the face. At bottom, the touch screen equipped with an eye-tracker presenting the chess game. These views are respectively at left, right and center of figure 1. The wooden structure is rigid to fix position and orientation of all sensors. The lighting conditions are controlled by 2 USB LED lamps on the sides.

Several software systems were used for recording and/or analyzing data. The Lichess Web Platform[1] serves for playing and recording games. Two commercial software provide both online and offline information: Tobii Studio 3.4.7 for acquisition and analyze of eye-gaze; Noldus FaceReader 7.0 for emotion detection. Body postures information were given by two different means: by the Kinect 2.0 SDK and by using our enhanced version of the Realtime Multi-Person Pose Estimation software [4]. Considering

[1] https://en.lichess.org/ (last seen 09/2017)

the state-of-the-art results of the second software, we decided to keep only this one for this experiment. During the study, data were recorded from all sensors (Kinect 2, Webcam, Screen capture, user clicks, Tobii-Bar) using the RGBD Sync SDK[2] from the MobilRGBD project [18]. This framework permits to read recorded and further computed data (gaze fixation, emotion detection, body skeleton position, etc.) for synchronous analysis by associating a timestamp with a millisecond precision to each recorded frame. The same framework can read, analyze and display the same way all gathered or computed data. An example is presented on figure 1 where most of the data are depicted.

3.1.2 *Participants.*

An announcement for our experiment with an invitation to participate was communicated to chess clubs, on the local university campus and within the greater metropolitan area. We received a positive response from the president of one of the top metropolitan area chess clubs, and 21 members volunteered to participate in our pilot experiment. Unfortunately, of these initial 21 participants, 7 recordings were not usable due to poor eye-tracking results and have not been included in our analysis. Indeed, these participants, while reflecting about the game, held their hand above the eye-tracker and disrupted its processing.

The 14 remaining chess players in our study were 7 experts and 7 intermediates level players (20-45 years, 1 female, age: $M = 31.71; SD = 7.57$). Expert players were all active players and with *Elo* ratings[3] ranged from 1759 to 2150 ($M = 1950; SD = 130$). For the intermediate players, the *Elo* ratings ranged from 1399 to 1513 ($M = 1415; SD = 43$) and 6 among them were casual players who were not currently playing in club. We can also give some statistics on the recorded session: the average recording time per participant is 14:58 minutes (MIN = 4:54, MAX = 23:40, SD = 5:26) and the average compressed size of gathered data is 56.12 GiB per session.

3.2 Methods

3.2.1 *Chess Tasks.*

The goal of this experiment was to engage participants into a cognitive process while observing their physiological reactions. Thirteen chess tasks were elaborated by our team in coordination with the president of the chess club. Two kinds of task were selected: *chess openings tasks*, where only 3 to 5 moves were played from the original state; and *N-Check-Mate tasks*, where 1 to 6 moves were required to check-mate the opponent (and finish the game).

Openings. Skilled players are familiar with most of the chess openings and play them intuitively. Intuitive play does not generally require cognitive engagement for reasoning. An important challenge is to detect when a player passes from intuitive reaction to a known opening, to challenging situations. Thus, two uncommon openings were selected to this end: a King's Gambit (3 moves from the initial state) and a Custom Advanced Variation of the Caro-Kann Defense (6 moves from initial state). The goal here is to pull participants out from their comfort zone as much as possible to evoke emotions and physiological reactions. Openings correspond to task number 1 and 2.

N-Check-Mate. Eleven end game tasks were defined. These are similar to the daily chess puzzles that can be found in magazines or on chess websites. Each of these tasks was designed to check-mate the opponent in a number of predefined moves ranging from 1 to 6. Tasks requesting 1 to 3 moves are viewed as easy task whereas 4 to 6 moves tasks require more chess reasoning abilities, etc. Distribution among the 11 tasks differs according to their number of required move and thus to their difficulty: 4 tasks with one move, 4 tasks with two and three moves (2 of each) and 3 tasks with four, five and six moves (1 of each). End games were presented to participants in this order of increasing difficulty while alternating the played color (white/black) between each task.

3.2.2 *Procedure.*

Participants were tested individually in sessions lasting approximately 45 minutes. Each participant was asked to solve the 13 chess tasks and their behaviors were observed and recorded. To avoid biased behavior, no information was given about the recording equipment. Nevertheless, it was necessary to reveal the presence of the eye-tracker bar to participants in order perform a calibration step. After providing informed consent, the Lichess web platform was presented and participants could play a chess game against a weak opponent (*Stockfish*[4] algorithm level 1: lowest level) to gain familiarity with the computer interface. No recording was made during this first game.

Once familiar and comfortable with the platform, the eye-tracking calibration was performed using Tobii Studio software, in which subjects were instructed to sit between 60 and 80cm from the computer screen and to follow a 9-point calibration grid. Participants were requested to avoid large head movement in order to assure good eye-tracking quality. Aside from this distance, no other constraints were instructed to participants.

Each task to solve was individually presented, starting with the openings, followed by the N-Check-Mate tasks. Participants were instructed to solve the task by either playing a few moves from the opening or to check mate the opponent (played by *Stockfish* algorithm level 8: the highest level) in the required number of moves. The number of moves needed for the N-Check-Mate tasks was communicated to the subject. A time frame was imposed for each task. The exact time frame was not announced to the participant, they only knew that they have a couple of minutes to solve the task. This time constraint ranges from 2 minutes for the openings and the easiest N-Check-Mate tasks (1-2 moves) to 5 minutes for the hardest ones (4-5-6 moves). An announcement was made when only one minute was remaining to solve the task. If the participant could not solve the task within the time frame, the task was considered as failed and the participant proceeded to the next task. The experiment is considered finished once all tasks were presented to the participant.

[2] https://github.com/Vaufreyd/RGBDSyncSDK (last seen 09/2017)

[3] The *Elo* system is a method to calculate rating for players based on tournament performance. Ratings vary between 0 and approximately 2850. https://en.wikipedia.org/wiki/Elo_rating_system (last seen 09/2017)

[4] *Stockfish* is an open-source game engine used in many chess software, including Lichess. https://en.wikipedia.org/wiki/Stockfish_(chess) (last seen 09/2017).

3.3 Analysis

3.3.1 *Eye-Gaze.*

Eye movement is highly correlated with focus of attention and engaged cognitive processes [19] in problem solving and human-computer interaction [20]. Other studies [12, 16] show that expertise estimation for chess players can be performed using several eye-tracking metrics such as fixation duration or visit count. In this case, gaze information can be useful to determine information such as:

(1) *What pieces received the most focus of attention from participants?*
(2) *Do participants who succeed to complete a task share the same scan path?*
(3) *Is there significant difference in gaze movements between novices and experts?*

To reach these aims, Areas Of Interests (AOIs) were manually defined for every task. An AOI can be a key piece for the current task (**e.g.** a piece used to check-mate the opponent), the opponent king, destinations cases where pieces have to be moved, etc. Afterward, we compute statistics for every AOI of each task. Among possible metrics, results depicted in this article are based on *Fixation Duration*, *Fixation Count* and *Visit Count*.

Interpretation for these metrics differs according to the task domain. For example, in the domain of web usability, Ehmke et al [21] would interpret long fixation duration on AOI as a difficulty to extract or interpret information from an element. In the field of chess, Reingold and Charness [12, 16] found significant differences in fixation duration between experts and novices.

3.3.2 *Facial emotions.*

Micro-expressions, as defined by Ekman and Fiesen [8] in 1969, are quick facial expressions of emotions that could last up to half a second. These involuntary expressions can provide information about cognitive state of chess players. In our pilot study, the Noldus FaceReader software [22] has been used to classify players' emotions in the form of six universal states proposed by Ekman: happiness, sadness, anger, fear, disgust and surprise (plus one neutral state). These emotional states are commonly defined as regions in a two-dimensional space whose axes are valence and arousal. Valence is commonly taken as an indication of pleasure, whereas arousal describes the degree to which the subject is calm or excited.

In practice, the FaceReader software analyses video by first applying a face detector to identify a unique face followed by a detection of 20 Facial Action Units [8]. Each action unit is assigned a score between 0 and 1 and these are used to determine the state label for emotion. Valance and arousal can be then computed as:

- **Valence**: intensity of positive emotions (*Happy*) minus intensity of negatives emotions (*sadness*, *anger*, *fear* and *disgust*);
- **Arousal**: computed accordingly to activation intensities of the 20 Action Units.

FaceReader was tested on two different datasets: the Radboud Faces Database [23] containing 59 different models and the Karolinska Directed Emotional Faces [24] which regroups 70 individuals. Both dataset display 7 different emotional expressions (plus neutral) on different angles. FaceReader algorithm correclty classified 90% of the 1197 images from Radboud Face Database [25] and 89% of the Karolinska Dataset (4900 images) [22].

3.3.3 *Body Posture.*

Body posture is a rich communication channel for human to human interaction with important potential for human computer interaction [26]. Studies have shown that self-touching behavior is correlated with negative affect as well as frustration in problem solving [27]. Thus, we have investigated a number of indicators for stress from body posture:

- **Body Agitation**: how many joints are varying along x, y and z axis;
- **Body Volume**: space occupied by the 3D bounding box built around joints (see [28]);
- **Self-Touching**: collisions between wrist-elbow segments and the head (see [29]).

These signals are computed from the RGBD streams recorded by the Kinect 2 where a list of body joints is extracted by means of our variant of a body pose detection algorithm [4]. These joints are computed on the RGB streams and projected back to Depth data. Thus, a 3D skeleton of the chess player is reconstructed and can be used as input to compute previous metrics. As one can see on figures 1 at left, from the point of view of the Kinect 2 in our setup (see figure 2), the skeleton information is limited to the upper part of the body, from hips to head.

4 RESULTS

Synchronous data for every feature have been extracted from all sensors. Several tasks, like regression over *Elo* ratings or over the time needed to perform a task, could be addressed using these data. Among them, we chose to analyze a classification problem that can be interpreted by a human:

- *Is it possible, by the use of gaze, body and/or facial emotion features, to detect if a chess player is an expert or not?*

This problem is used as example to obtain a first validation of our data relevancy. It is correlated with whether a chess player is challenging beyond his abilities.

This section presents unimodal and multimodal analysis of extracted features to determine chess expertise of players.

4.1 Unimodal analysis

4.1.1 *Eye-Gaze.*

Two AOIs were defined for each task: one AOI is centered on the very first piece to move in the optimal sequence to successfully achieve the check-mate; and the second one on the destination square where this piece has to be moved. Fixations and visits information of every task are gathered for all participants and results are presented in Figure 3.

As can be clearly seen in this figure, experts have longer and more fixations than intermediates on relevant pieces. Same result is observed for visit count. Similar results can be found in literature [12]. These results are explained by the expert's skill encoding capacity that enables them to quickly focus their attention on relevant piece by a better pattern matching ability.

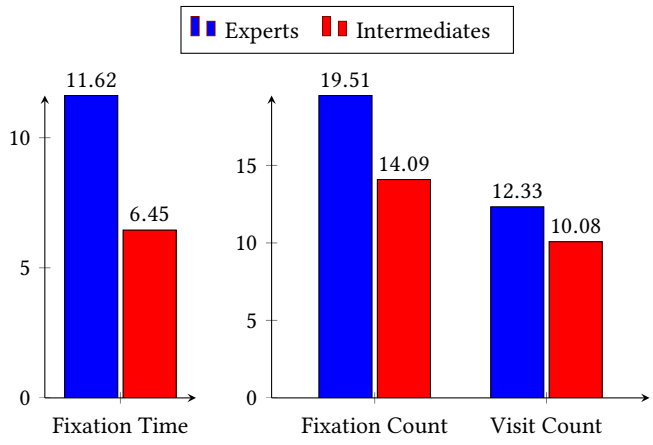


Figure 3: Eye-gaze histograms. Left: Percentage of fixation (in time). Right: number of fixations and number of visits.

More work has to be done on eye-gaze such as analyzing and comparing the scan path order of participants, measuring how fast are participants to identify relevant pieces or analyzing fixation on empty squares.

4.1.2 *Emotions.*

The increasing difficulty in the non-interrupting tasks has caused our participants to express more observable emotions across the experiment. Emotions in a long-task experiment are expressed as peaks in the two-dimensional space (valence, arousal). Thus, standard statistics tend to shrink toward zero as the record becomes longer.

Other approaches should be considered to visualize emotion expressions. One possibility is to consider the number of changes of emotions having the highest intensity (i.e. the current detected facial emotion). As emotion intensities are based on facial unit detection, changes in the main emotion denote underlying changes in facial expression. The result metric is shown on the graph presented in figure 4.

It clearly appears that expression of emotions increase with the difficulty of the problem to solve. For both player classes, there is a peak for the second task (i.e. our uncommon custom advanced variation of the Caro-Kann defense). This opening was surprising for all participants, more than the King's Gambit one (task 1). No participant was familiar with this kind of opening. Moreover, intermediates players present an emotional peak at task number 9, which is the first task to require more than 2 moves to check-mate the opponent, whereas expert's plot shape looks more like the beginning of an exponential curve. An interesting aspect of that plot is the final decrease of intermediate players after task 10, this could be interpreted as a sort of resignation, when players knew that tasks beyond of their skills and could not be resolved.

These primary results suggest that situation understanding and expertise knowledge can be inferred from variation of facial emotions. Although, more detailed analysis, such as activation of Action Units, derivative of emotions or detection if a micro expression occurs right after a move being played should be performed.

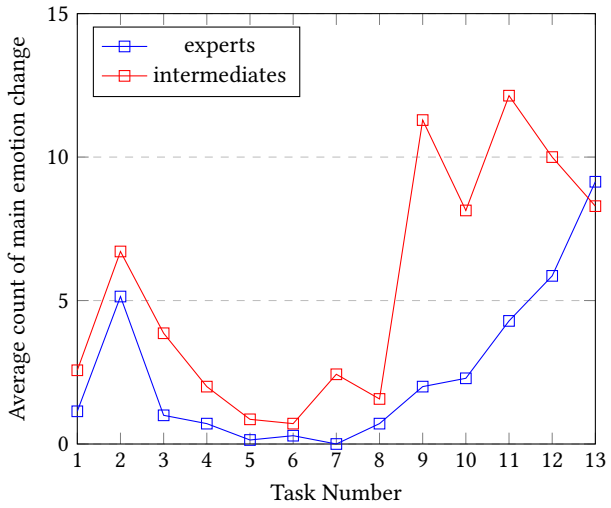


Figure 4: Average count of variation of main detected facial emotion in regard to the task (1-13). Tasks are ranging in an increasing difficulty order.

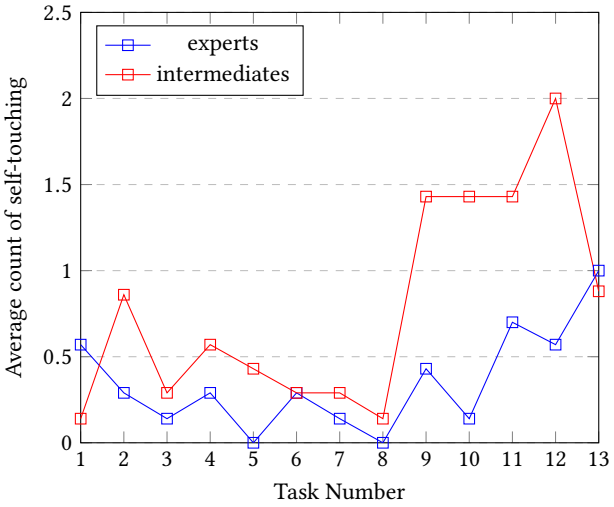


Figure 5: Average count of self-touching in regard to the task (1-13). Tasks are ranging in an increasing difficulty order.

4.1.3 *Body Posture.*

The increasing difficulty of the N-Check-Mate tasks is a stress factor that can be observable according to [27]. Using technique presented in [29] to detect self-touching, we can observe how participants' body reacts to the increasing difficulty of tasks.

The figure 5 presents statistics about self-touching. Shapes of lines are very similar of what is observed for emotions (Figure 4). The same conclusion can be drawn: the number of self-touches increases as tasks get harder and it reveals that this is a relevant feature to consider. However, analysis of volume and agitation features did not reveal interesting information yet. This can be explained either by the nature of the task or by the number of

	G	B	E	G + B	G + E	B + E	G + B + E
Task Dependent ($N = 154$)	62%	58%	78%	58%	79%	79%	78%
All Tasks ($N = 14$)	71%	79%	86%	71%	86%	93%	93%

Table 1: Best accuracy scores from cross-validation for SVMs. First line is Task Dependent approach, the number of sample N is the number of participants (14) times the number of N-Check-Mate tasks (11). Second approach uses only average data of all task for every participant (N=14). Columns are the modality subset chosen to train the SVM (G: Gaze, B: Body, E: Emotions).

analyzed participants. More discussion of this experiment can be found in section 5.

4.2 Multimodal versus unimodal classification

To demonstrate the potential benefit of a multimodal approach, a supervised machine learning algorithm has been used to quantify accuracy of different modalities for classification. Only the data recorded for the 11 N-Check-Mate tasks are considered here. Support Vector Machines (SVM) have been built for each modality and for each possible combination of modalities.

After computing statistical analysis (mean, variance, standard deviation) over our features, two approaches are compared: a task dependent approach on one hand and a skill-only dependent (All Task) on the other hand. First approach considers statistical results for every participant and for every task. That way, one input sample would be the instantiation of one participant for one particular task, given a total number of $14 * 11 = 154$ input samples. Second approach takes the average over all tasks for each participant. Input sample is reduced to participant with average statistics over tasks as features.

Stratified cross-validation procedure has been used on every SVM and for both approaches to compute their accuracy. Results are shown in table 1. First observations of these results show that the task dependent approach presents a far less accuracy score than the second approach. This could be explained by the variation in the length of recordings. Indeed, some participants managed to give an answer in less than 10 seconds. The second hypothesis shows good performance and validates one of our expectation that multimodal system could outperform unimodal ones. Even if these scores are promising, further experiments with more involved participants have to be performed to confirm these primary results.

5 DISCUSSION

This research and primary results (see section 4) show consistency results on unimodal features used to distinguish expert and intermediate chess players. When used together, body posture, visual attention and emotion provide better accuracy using a binary SVM classifier. Although these results appear promising, they are only preliminary: the number of participants (14); the variation of recording duration (from seconds to a couple of minutes depending on the task and playerâĂŹ expertise); and the tasks must all be expanded and developed. Due to the size of our dataset, generalizing this preliminary results is not possible for the moment. Further experiments must be conducted to validate them.

The conditions of the chess tasks should also draw attention. In the experimental configuration, chess players were facing a chess algorithm engine in tasks where they knew the existence of a winning sequence of moves. Moreover, players are seating (see figure 1), some clues like body agitation may provide less information than expected. Participants may not be as engaged as they would have been in a real chess tournament facing a human opponent using an actual chess board. In these particular situations, involving stakes for players, the physiological reactions and emotional expressions are more interesting to observe.

Nevertheless, these experiments reveal that valuable information can be observed from human attention and emotions to determine understanding, awareness and affective response to chess solving problems. Another underlying result is the validation of our setup in monitoring chess players. The problems encountered with the eye-tracker for 7 participants (see section 3.1.2) show that we must change its position to increase eye-tracking accuracy.

6 CONCLUSION

This paper presents results from initial experiments with the capture and interpretation of multi-modal signals of 14 chess players engaged in solving 13 challenging chess tasks. Results show that eye-gaze, body posture and emotions are good features to consider. Support Vector Machine classifiers trained with cross-fold validation revealed that combining several modalities could give better performances (93% of accuracy) than using a unimodal approach (86%). These results encourage us to perform further experiments by increasing the number of participants and integrating more modalities (audio procedural speech, heart rate etc.).

Our equipment is based on off-the-shelf commercially available components as well as open source programs and thus can be easily replicated. In addition to providing a tool for studies of participants engaged in problem solving, this equipment can provide a general tool that can be used to study the effectiveness of affective agents in engaging users and evoking emotions.

ACKNOWLEDGMENTS

This research has been be funded by the French ANR project CEEGE (ANR-15-CE23-0005), and was made possible by the use of equipment provided by ANR Equipement for Excellence Amiqual4Home (ANR-11-EQPX-0002). Access to the facility of the MSH-Alpes SCREEN platform for conducting the research is gratefully acknowledged.

We are grateful to all of the volunteers who generously gave their time to participate in this study and to Lichess webmasters for their help and approval to use their platform for this scientific experience. We would like to thank Isabelle Billard, current chairman of the chess club of Grenoble "*L'Échiquier Grenoblois*" and all members who participated actively in our experiments.

REFERENCES

[1] R. El Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Real-time vision for human-computer interaction*. Springer, 2005, pp. 181–200.

[2] T. Baltrušaitis, D. McDuff, N. Banda, M. Mahmoud, R. El Kaliouby, P. Robinson, and R. Picard, "Real-time inference of mental states from facial expressions and upper body gestures," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 909–914.

[3] T. BaltruÅąaitis, P. Robinson, and L. P. Morency, "Openface: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–10.

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[5] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *CVPR*, 2017.

[6] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016.

[7] D. Kahneman, *Thinking, fast and slow*. Macmillan, 2011.

[8] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.

[9] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE transactions on biomedical engineering*, vol. 58, no. 1, pp. 7–11, 2011.

[10] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[11] R. Stiefelhagen, J. Yang, and A. Waibel, "A model-based gaze tracking system," *International Journal on Artificial Intelligence Tools*, vol. 6, no. 02, pp. 193–209, 1997.

[12] N. Charness, E. M. Reingold, M. Pomplun, and D. M. Stampe, "The perceptual aspect of skilled performance in chess: Evidence from eye movements," *Memory & cognition*, vol. 29, no. 8, pp. 1146–1152, 2001.

[13] L. Paletta, A. Dini, C. Murko, S. Yahyanejad, M. Schwarz, G. Lodron, S. Ladstätter, G. Paar, and R. Velik, "Towards real-time probabilistic evaluation of situation awareness from human gaze in human-robot interaction," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '17. New York, NY, USA: ACM, 2017, pp. 247–248. [Online]. Available: http://doi.acm.org/10.1145/3029798.3038322

[14] T. Giraud, M. Soury, J. Hua, A. Delaborde, M. Tahon, D. A. G. Jauregui, V. Eyharabide, E. Filaire, C. Le Scanff, L. Devillers *et al.*, "Multimodal expressions of stress during a public speaking task: Collection, annotation and global analyses," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 417–422.

[15] M. K. Abadi, J. Staiano, A. Cappelletti, M. Zancanaro, and N. Sebe, "Multimodal engagement classification for affective cinema," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 411–416.

[16] E. M. Reingold and N. Charness, "Perception in chess: Evidence from eye movements," *Cognitive processes in eye guidance*, pp. 325–354, 2005.

[17] M. Portaz, M. Garcia, A. Barbulescu, A. Begault, L. Boissieux, M.-P. Cani, R. Ronfard, and D. Vaufreydaz, "Figurines, a multimodal framework for tangible storytelling," in *WOCCI 2017 - 6th Workshop on Child Computer Interaction at ICMI 2017 - 19th ACM International Conference on Multi-modal Interaction*, Glasgow, United Kingdom, Nov. 2017, author version. [Online]. Available: https://hal.inria.fr/hal-01595775

[18] D. Vaufreydaz and A. Nègre, "MobileRGBD, An Open Benchmark Corpus for mobile RGB-D Related Algorithms," in *13th International Conference on Control, Automation, Robotics and Vision*, Singapour, Singapore, Dec. 2014. [Online]. Available: https://hal.inria.fr/hal-01095667

[19] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford, 2011.

[20] A. Poole and L. J. Ball, "Eye tracking in hci and usability research," *Encyclopedia of human computer interaction*, vol. 1, pp. 211–219, 2006.

[21] C. Ehmke and S. Wilson, "Identifying web usability problems from eye-tracking data," in *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1*. British Computer Society, 2007, pp. 119–128.

[22] M. Den Uyl and H. Van Kuilenburg, "The facereader: Online facial expression recognition," in *Proceedings of measuring behavior*, vol. 30, 2005, pp. 589–590.

[23] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[24] E. Goeleven, R. De Raedt, L. Leyman, and B. Verschuere, "The karolinska directed emotional faces: a validation study," *Cognition and emotion*, vol. 22, no. 6, pp. 1094–1118, 2008.

[25] G. Bijlstra and R. Dotsch, "Facereader 4 emotion classification performance on images from the radboud faces database," 2015.

[26] S. M. Anzalone, S. Boucenna, S. Ivaldi, and M. Chetouani, "Evaluating the engagement with social robots," *International Journal of Social Robotics*, vol. 7, no. 4, pp. 465–478, 2015.

[27] J. A. Harrigan, "Self-touching as an indicator of underlying affect and language processes," *Social Science & Medicine*, vol. 20, no. 11, pp. 1161–1168, 1985.

[28] W. Johal, D. Pellier, C. Adam, H. Fiorino, and S. Pesty, "A cognitive and affective architecture for social human-robot interaction," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. ACM, 2015, pp. 71–72.

[29] J. Aigrain, M. Spodenkiewicz, S. Dubuisson, M. Detyniecki, D. Cohen, and M. Chetouani, "Multimodal stress detection from multiple assessments," *IEEE Transactions on Affective Computing*, 2016.