Perception pour la robotique mobile en environnement humain

Frédéric Lerasle

N° d'ordre:

Habilitation à diriger des recherches

Présentée devant

l'Université Paul Sabatier de Toulouse

par

Frédéric Lerasle
Maître de conférences UPS

Équipe d'accueil : Groupe RAP, LAAS-CNRS
7 avenue Colonel Roche
31077 Toulouse Cedex

Perception pour la robotique mobile

en environnement humain

---

soutenue le 18 Janvier 2008 devant le jury composé de

M. :	Raja	Chatila	Président
MM. :	Jan Olof	Eklundh	Rapporteurs
	Dominique	Meizel	
	Monique	Thonnat	
MM. :	Maurice	Briot	Examinateurs
	Patrice	Dalle	
	Michel	Devy	
	Michel	Dhome	

Remerciements

Les travaux présentés ici ont été effectués au LAAS-CNRS après ma nomination en tant que Maître de conférences à l'Université Paul Sabatier de Toulouse suite à un doctorat réalisé au LASMEA de Clermont-Ferrand. Je remercie les directeurs de ces deux structures, qui m'ont accueilli et m'ont permis de trouver ma voie dans l'enseignement et la recherche : Marc Richetin au LASMEA, puis J.C.Laprie, M.Ghallab et R.Chatila au LAAS-CNRS.

C'est avec grand plaisir que je remercie les rapporteurs de mon mémoire d'habilitation à diriger les recherches pour la caution qu'ils ont bien voulu apporter à mes travaux :

- Jan Olof Ekhlund, Professeur au KTH de Stockholm,
- Dominique Meizel, Professeur à l'ENSIL de Limoges,
- Monique Thonnat, Directrice de recherche INRIA à Sophia-Antipolis.

Leurs conseils, suggestions et questions ont été source d'enrichissement.

Mes remerciements vont également aux examinateurs qui m'ont fait l'honneur et l'amitié de participer à ce jury :

- Maurice Briot, Professeur à l'Université Paul Sabatier et au LAAS-CNRS, à qui je dois beaucoup, tant du point de vue de l'enseignement que de la recherche, pour ses nombreux conseils et sa disponibilité,
- Raja Chatila, Directeur de recherche au LAAS-CNRS, qui par son écoute et son soutien, a toujours encouragé mes activités de recherche et a accepté de présider ce jury,
- Patrice Dalle, Professeur à l'Université Paul Sabatier et l'IRIT, avec qui les (trop rares) échanges scientifiques sont toujours enrichissants,
- Michel Devy, Directeur de recherche au LAAS-CNRS, qui m'a aidé à appréhender les problématiques robotiques et à trouver ma place au sein de RIA,
- Michel Dhome, Directeur de recherche CNRS au LASMEA, qui a encadré mes premiers pas en recherche et a suscité chez moi la "fibre" *Vision par ordinateur* comme l'atteste ce mémoire.

Enfin un grand merci à tous mes collègues toulousains chercheurs, enseignant-chercheurs et administratifs qui, de près ou de loin, ont facilité hier mon intégration malgré mon "accent du nord" et qui, aujourd'hui, contribuent par leur bonne humeur à mon épanouissement au laboratoire et/ou à l'Université.

Résumé

Ce mémoire d'habilitation à diriger les recherches porte sur la perception et la compréhension conjointe de l'espace et du milieu par un robot cognitif autonome. Dans ce contexte, la démarche consiste ici à intégrer des percepts multiples et incertains à tous les niveaux de la perception à partir de capteurs visuels embarqués. Ces travaux se structurent en deux thèmes.

Le premier thème se focalise sur la perception de l'espace pour la navigation autonome en milieu intérieur. Nos travaux antérieurs ont mis l'accent sur une méthodologie complète de détection, reconnaissance et localisation sur amers visuels validée par des expérimentations réelles sur le robot Diligent. Ces amers sont capturés automatiquement par le robot dans les différentes représentations métriques et topologiques de son environnement de travail. La navigation consiste alors à exploiter ces modèles pour se localiser métriquement ou qualitativement, sur la base de données visuelles, éventuellement télémétriques. À terme, ces représentations seront enrichies par des informations sémantiques capturées en interaction avec l'homme.

Cet apprentissage supervisé, la perspective d'un robot sociable, nous ont amené à démarrer le second thème sur la perception par le robot de l'homme pour leur interaction. Nos travaux ont porté sur la détection, le suivi, la reconnaissance de l'homme par vision monoculaire couleur. Parmi ces fonctions, la problématique du suivi est centrale puisque la plupart des tâches robotiques coordonnées avec l'homme nécessite de caractériser la relation d'une plate-forme mobile aux agents humains *a priori* mobiles. Nous avons ainsi prototypé puis intégré plusieurs fonctions de suivi 2D ou 3D de tout ou partie des membres corporels de l'homme par le choix conjoint de stratégies de fusion de données visuelles et de filtrage particulaire répondant aux modalités d'interaction envisagées pour le robot « guide » Rackham et le robot compagnon Jido.

Les prospectives énoncées visent à l'interaction de percepts relative à la perception simultanée par le robot de l'espace et/ou de l'homme. La problématique de l'intelligence ambiante, par l'ajout de robots anthropomorphes type humanoïde dans ces environnements humains, devrait infléchir ces travaux tout en recoupant certaines investigations passées ou actuelles.

Table des matières

Première partie

Activités de recherche

Chapitre 1

Contexte et projet de recherche

1 Préambule

Mes premiers pas en recherche se sont déroulés au LASMEA de Clermont-Ferrand. J'ai préparé, dans ce laboratoire, un doctorat de l'Université Blaise Pascal sous la direction de M.Dhome (DR CNRS) et G.Rives (MCF Université Blaise Pascal) durant la période 1994-1997. Ces travaux s'inscrivent dans le cadre de la vision par ordinateur. En septembre 1997, j'ai intégré le groupe RIA[1] du LAAS-CNRS en tant qu'enseignant-chercheur à l'Université Paul Sabatier de Toulouse III. Une des problématiques abordées aujourd'hui par le pôle RIA est la robotique au service de l'homme avec pour perspective de conférer des capacités interactives et cognitives à des plate-formes autonomes mobiles. Ceci passe par le développement de méthodes et de technologies pour la perception, l'interprétation, le raisonnement et l'apprentissage en interaction avec l'homme. Plus largement, le pôle RIA dispose aujourd'hui d'un important savoir-faire sur l'intégration d'algorithmes dédiés sur des plate-formes autonomes mobiles.

J'ai naturellement trouvé ma place au sein de l'équipe perception du groupe RIA puis depuis septembre 2007 au sein du groupe RAP. Les problématiques abordées sont la construction de modèles de l'environnement depuis ces plate-formes mobiles, la navigation à partir de ces modèles, l'apprentissage et la reconnaissance d'objets pour les robots manipulateurs et, de manière plus générale, sur l'interprétation des scènes rencontrées à partir de données sensorielles acquises par des capteurs extéroceptifs embarqués. Parmi ces capteurs extéroceptifs, l'utilisation de la vision est une tendance forte de la robotique actuelle. Cette tendance est notamment motivée par la richesse de l'information délivrée par un tel capteur.

Ma contribution actuelle est au confluent de la vision par ordinateur et de la robotique même si les problématiques, formalismes et outils utilisés sont souvent communs. Ces deux domaines de recherche sont historiquement très liés, rappelons par exemple que la vision 3D tire ses fondements de la robotique. Côtoyer ces deux communautés permet d'échanger scientifiquement sur les techniques avancées de vision et de les exploiter dans une problématique robotique par définition extrêmement riche et complexe.

Je reste ainsi très attaché à l'intégration des algorithmes développés à bord de véritables robots. Dans le sens des traditions du pôle RIA et des exigences de la robotique, mes travaux de recherche comportent, en règle générale, deux volets : un volet formel et algorithmique qui tire pertinence et validation d'un volet expérimental s'appuyant sur les plate-formes mobiles du

[1]RIA, devenu pôle du LAAS-CNRS depuis peu, comprend trois groupes de recherche dont le groupe Robotique, Action et Perception (RAP).

LAAS-CNRS en milieu intérieur : Diligent, Rackham, Jido et prochainement HRP2 (figure 1.1).

(a) (b) (c) (d)

FIG. 1.1 – Les robots d'intérieur du pôle RIA : Diligent (a), Rackham (b), Jido (c) et HRP2 (d).

Mes contributions passées au sein du laboratoire depuis septembre 1997 portent plus spécifiquement sur deux volets fondamentaux relatifs à la perception depuis un robot mobile, *i.e.* :

1. La perception d'amers visuels en environnement humain pour la navigation au long cours de robots mobiles autonomes.
2. La perception de l'homme pour l'interaction Homme/Robot (H/R).

Ce positionnement scientifique répond clairement au souhait de s'inscrire dans les problématiques et besoins exprimés par le pôle RIA. Néanmoins, un second souhait est de bonifier les travaux réalisés à travers des projets et collaborations industriels ou académiques externes au laboratoire. Ces implications dans des projets ou collaborations externes sont en recouvrement partiel ou total avec mon projet de recherche ; nous le verrons ultérieurement.

Cette partie relative à mes activités de recherche se détaille comme suit. Après cette brève introduction, le chapitre 1 résume, dans la section suivante, les travaux réalisés durant mon doctorat. La problématique du robot assistant, les enjeux sociétaux associés et la finalité de mon projet de recherche sont énoncés dans la section 3. Le chapitre 2 décrit les travaux antérieurs relativement aux deux thématiques énoncées. Le chapitre 3 présente enfin les travaux actuels et les prospectives à moyen et long termes, celles-ci s'inscrivant naturellement dans le projet de recherche énoncé en section 3.

2 Résumé de mes travaux de doctorat

Ces travaux portent sur la capture du mouvement humain à partir du flot vidéo délivré par un système de vision multi-oculaire. Cette problématique

trouve de nombreux champs d'application tels que la médecine rééducative, la biomécanique, l'ergonomie, le sport, la synthèse d'images et bien sûr la robotique.

Certes, le marché propose de nombreux systèmes de capture de mouvement humain. Citons par exemple les systèmes inertiels (Intersense), magnétiques (TRIDENT, Ascension Motion Star) et surtout opto-électroniques à partir de marqueurs artificiels (CODA-3, ELITE, VICON). L'intérêt de ces amers réside dans l'étiquetage et le suivi de primitives corporelles spécifiques car le corps humain n'offre pas de primitives remarquables par manque de texture naturelle liée à la peau. Cependant, le nombre et l'emplacement prédéfinis de ces amers, leur coût, enfin la lourdeur de mise en œuvre, limitent l'usage de tels systèmes.

L'idée originale de ces travaux est de contourner le problème crucial des amers en compensant le manque naturel de texture de la peau par le port d'un justaucorps texturé et d'exploiter les techniques et outils connus de la vision par ordinateur.

Ainsi, l'approche retenue s'inspire des méthodes dites « basées apparence et modèle ». Elle repose sur une **modélisation géométrique 3D et cinématique des membres corporels à suivre**. Le principe est alors de plaquer la texture du justaucorps sur le modèle manipulé lors d'une procédure d'apprentissage préalable au suivi visuel. Le suivi visuel n'est autre qu'une localisation successive sur chaque image du flot vidéo. L'algorithme de localisation repose sur l'interprétation de points caractéristiques dans l'image comme étant les projections perspectives de points 3D liés au modèle texturé articulé et d'un processus itératif fondé sur la méthode de Levenberg-Marquardt pour estimer l'attitude du modèle conforme à l'image analysée. Pour le suivi, une étape de prédiction par filtrage de Kalman est associée à la procédure d'appariements pour restreindre la recherche des points homologues. Enfin, le système multi-oculaire, étalonné hors ligne, permet de combiner les appariements effectués dans plusieurs images prises simultanément de différents points de vue pour en déduire l'attitude du modèle compatible avec toutes les vues.

L'algorithme de localisation déterministe a tout d'abord été validé sur des objets articulés rigides en exploitant les contours extraits dans l'image. La figure 1.2 montre un exemple de localisation par vision monoculaire d'un bras manipulateur.

Concernant la capture du mouvement humain à partir du justaucorps, les expérimentations ont porté sur l'analyse de séquences pré-enregistrées de pédalage avec vélo ergonomique. Pour ce faire, il nous a fallu modéliser la géométrie de la jambe observée à partir de coupes I.R.M ainsi que les degrés

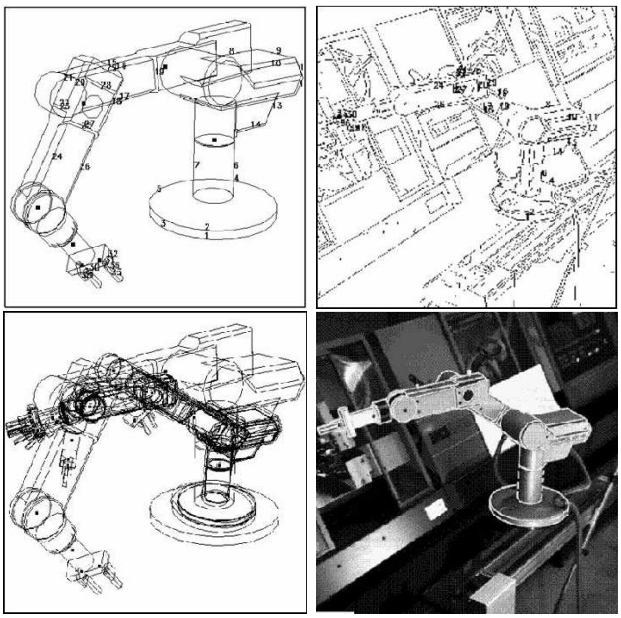
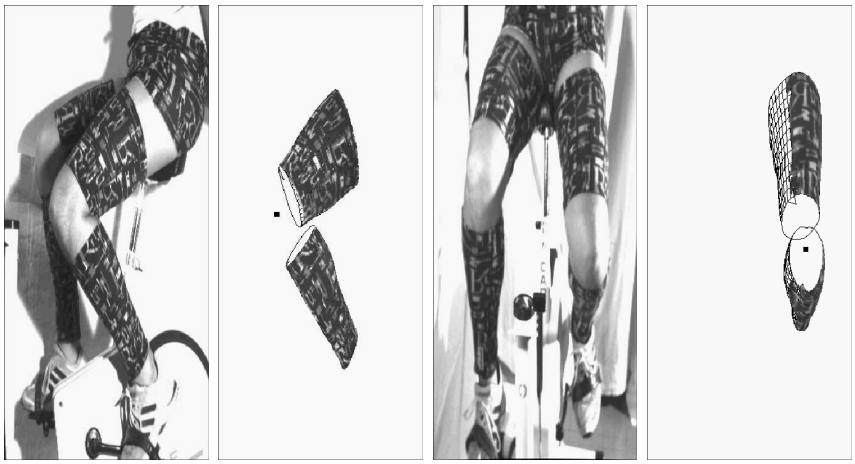


FIG. 1.2 – Localisation par vision monoculaire d'un bras robotique : appariements entre arêtes modèle et segments image appariés -bas-, attitude calculée du modèle par processus itératif et projection du modèle dans cette attitude -bas-.

de liberté (ddls) de l'articulation du genou. Nous avons également ajouté des paramètres de déformation pour prendre en compte la contraction musculaire durant l'effort. La figure 1.3 montre un exemple de localisation par vision binoculaire.

Les résultats obtenus sont encourageants au sens *où*, pour l'articulation du genou, nous retrouvons certaines caractéristiques mises en évidence en recherche médicale. Malgré de nombreuses publications revues [8, 45, 46] et congrès [47, 48, 49, 50], l'approche proposée a néanmoins des limitations certaines. Ainsi, les coûts en temps de calcul sont clairement incompatibles avec une application quasi temps réel. De plus, il serait intéressant de s'affranchir : (1) du port du justaucorps pour considérer des caractéristiques visuelles plus naturelles, (2) de l'examen I.R.M pour exploiter un modèle géométrique frustre donc plus générique. Enfin, il serait opportun d'étendre l'approche au suivi visuel du corps humain tout entier.

Cette problématique est encore aujourd'hui largement abordée dans la communauté Vision comme l'attestent les nombreux travaux actuels sur le sujet [Moeslund 2001]. De mon point de vue, les approches existantes ne sont pas encore complètement compatibles avec les environnements complexes et



(a) (b) (c) (d)

FIG. 1.3 – Localisation par vision binoculaire d'un membre corporel : images délivrées par le système (a)-(c), modèle texturé projeté dans l'attitude calculée pour ces images (b)-(d).

non contraints rencontrés en robotique mobile pour laquelle les ressources CPU embarquées sont par ailleurs souvent limitées. Fort de ce constat, le thème n° 2, détaillé au chapitre 2, inclut entre autre cette problématique de la capture du mouvement humain à partir de capteurs extéroceptifs embarqués sur un robot mobile.

3 Projet de recherche

3.1 Motivations

La Robotique étudie la conception de machines intelligentes capables de perception, de décision, de mouvement, et d'action dans un environnement ouvert ou confiné, dynamique, imparfaitement modélisé, voire inconnu. Son extension à plusieurs agents (robots et/ou humains) censés partager cet environnement, pose le problème de leur cohabitation, voire de leur coopération si ceux-ci exécutent des tâches collaboratives. Il peut s'agir, par exemple, de coopération multi-robots dans un contexte de robotique en environnement naturel ou, ici, de coopération homme/robot dans un contexte de robotique en environnement humain. Les défis sont alors, à court terme, celui de la robotique d'assistance à l'homme et, à plus long terme, celui de la robotique personnelle dans des environnements aménagés par l'homme et donc *a priori* structurés.

Dans ce contexte, nous pensons, à l'instar d'une grande partie de la communauté Robotique, que les robots n'ont pas pour vocation de se substituer à l'homme mais bien d'intégrer ce dernier dans la boucle perception-décision-action. Ainsi, ils perçoivent l'homme et ses activités, leurs actions s'effectuent en partage et/ou en synergie avec lui... en s'appuyant sur des décisions interactives et partagées. La finalité est d'assister l'homme, de le servir ou de l'accompagner dans ses tâches quotidiennes.

Cette perspective vient enrichir la problématique déjà très riche de la machine intelligente. Un tel robot doit plus spécifiquement être doté de capacités de cognition artificielle incluant :

1. La perception et la compréhension de son espace et de son milieu dans le cadre, par exemple, d'une tâche de navigation dans son espace de travail ou de caractérisation de l'état des autres agents du milieu dans le voisinage immédiat du robot.
2. La prise de décision autonome ou partagée, par exemple, pour éviter/céder le passage aux autres agents dans un espace contraint.
3. L'exécution d'actions en interaction sûre et fiable avec les autres agents, par exemple la manipulation conjointe d'objets avec les autres agents.

Il doit enfin pouvoir développer ces capacités à travers son expérience par apprentissage automatique ou supervisé par l'homme. La finalité est de voir le robot adopter un comportement sociable en présence d'humains afin d'en permettre son acceptabilité. Cette acceptabilité va au delà de « simples » capacités perceptuelles embarquées sur lesquelles se focalise mon projet de recherche. Elle sous-tend, en plus, des caractéristiques anthropomorphes, des moyens d'action, de comportement, d'expressivité qui confèrent au robot une apparence familière et compréhensible par l'homme. Ces divers aspects, qui expliquent l'avènement de la robotique humanoïde, ne seront hélas pas considérés dans mon projet de recherche.

Les enjeux et impacts socio-économiques associés sont multiples. Il peut s'agir tout aussi bien :

- **de robots assistants ou auxiliaires de services** aussi bien du grand public que du professionnel d'un domaine donné. Leurs capacités sont définies *a priori* et liées aux services à assurer dans certains lieux publics. Citons par exemple les robots guides dédiés aux musées (Minerva [Thrun 1999], Rhino [Burgard 1999], RoboX [Siegwart 2003], Robovie [Kanda 2007], Rackham [54]) ou aux maisons de retraites (RG [Kulyukin 2004], Pearl [Pineau 2003]).
- **de robots personnels ou compagnons**, assimilables à des robots assistants de seconde génération, destinés à des interactions et tâches

> plus personnelles avec l'homme. Citons ici les robots Papero chez NEC et Qrio chez Sony. À l'instar des ordinateurs personnels, la finalité des robots personnels est d'acquérir de nouvelles capacités et connaissances à l'aide d'un apprentissage ouvert et actif et d'évoluer en constante interaction et coopération avec l'homme. Ces robots, aux capacités évolutives et corrélées aux besoins spécifiques de leurs tuteurs, sont appelés robots cognitifs. Citons ici les plate-formes expérimentales Cog [Fitzpatrick 2003], Biron [Maas 2006], et Jido [Fontmarty 2007].

L'énumération est loin d'être exhaustive. Un état de l'art complet sur les plate-formes sociables et leurs applications est accessible dans [Fong 2003]. Le développement de tels robots pourrait se concrétiser à terme par leur déploiement à large échelle dans les lieux publics ou privés. Cette perspective est une réponse possible à la question de la prise en charge des personnes âgées dans les sociétés modernes... eu égard à leur nombre en constante augmentation pour une population active en diminution. Les robots personnels pourraient plus largement toucher toutes les tranches d'âges et ainsi révolutionner, dans un avenir proche, notre vie quotidienne au même titre que les PDA (pour *Personal Digital Assistants*) et les ordinateurs portables.

Force est de constater cependant que le bilan de la robotique, notamment d'assistance, reste en deçà des attentes, et ce malgré les progrès de l'électronique et de l'informatique. La robotique d'aide aux handicapés, pourtant active depuis une vingtaine d'années, en est une illustration. Certes, nous pouvons nuancer ces propos : certains succès récents ont popularisé le robot mobile assistant auprès du grand public comme le robot humanoïde, tandis que certains projets fédérateurs et de grande envergure ont permis des avancées significatives. Nous pouvons néanmoins affirmer que le robot assistant constitue encore aujourd'hui un défi de recherche très ouvert sur le plan scientifique.

Un premier défi, bien que abordé dès le début de l'aventure robotique, est relatif à la navigation, la navigation étant le premier jalon à poser vers l'autonomie. Schématiquement, naviguer consiste à planifier et contrôler ses déplacements ce qui nécessite sa localisation, la détection et l'évitement des obstacles dans son voisinage. Parmi ces actions, la localisation efficace et sûre constitue un point clé : elle requiert la construction préalable par le robot d'une représentation de l'espace.

Au-delà de la mobilité, un second défi, abordé plus tardivement, est relatif à la communication et l'interaction H/R. Pour communiquer, le robot doit nécessairement partager des concepts communs avec l'homme, et au-delà, comprendre son milieu pour interagir avec les autres agents de l'environnement. Le robot s'appuie ici sur une représentation du milieu à construire.

Une parenthèse est ici nécessaire pour caractériser ces notions d'espace, de milieu, d'environnement ainsi que leurs représentations associées. **Le milieu** fait référence aux concepts de lieux, d'objets et d'actions humaines de l'environnement donc à une description plus sémantique afin de raisonner sur les interactions entre les entités inanimées (lieux, objets) de l'environnement et les entités animés[2] qui partagent l'espace. Nous différencions ici les actions de l'homme sur les objets dans un lieu donné, de ses activités assimilées à des séquences (ou chroniques) de mouvements corporels ou d'actions mettant en jeu possiblement plusieurs objets et/ou lieux. La représentation du milieu capture donc la sémantique de l'environnement. **L'espace** fait référence aux informations métriques et topologiques de l'environnement. Ces « sous-représentations » métriques ou topologiques, exhibent souvent des indices perceptuels saillants, uniques localement et proéminents que l'on appellera amers. Cette définition sous-entend que ces amers peuvent être des entités abstraites de l'environnement. La représentation de **l'environnement** regroupe ces représentations de l'espace et du milieu... qui ne sont pas disjointes. Ainsi, les amers pourront correspondre à des entités physiques de l'environnement, typiquement des objets du milieu. Enfin, les zones ou places de la topologie spatiale pourront correspondre à des lieux du milieu.

Fort de ces définitions, revenons sur les deux défis pré-cités. Ils ont suscité et suscitent encore de nombreux travaux dans la communauté Robotique. Ils ont abouti à la construction et à l'instrumentation de plate-formes robotiques intégrant des capacités plus ou moins avancées de navigation [Burgard 1999, Kulyukin 2004, Siegwart 2003], de communication et d'interaction avec l'homme [Bennewitz 2005, Fitzpatrick 2003, Maas 2006] à partir des capteurs extéroceptifs embarqués. Répertorier et caractériser ces capacités, pour les plate-formes mentionnées précédemment (et plus largement), amènent à deux constats :

1. Ces capacités sont le plus souvent gérées de façon découplée alors qu'elles pourraient se bonifier mutuellement en partageant la connaissance relative à l'espace et au milieu. Ainsi, la perception de lieux et/ou d'objets par le robot lors de sa navigation permet d'inférer des connaissances sur le milieu et faciliter l'interaction H/R. Par ailleurs, la perception par le robot de l'homme (ou de ses actions) durant leur interaction est souvent caractéristique d'un lieu donné ce qui permet une localisation qualitative du robot dans son espace de travail. Il nous semble donc opportun de généraliser l'intégration de percepts multiples dans les diverses représentations et plus largement dans toutes les fonctions perceptuelles embarquées.

[2] Les robots mobiles et surtout ici les hommes.

2. Les capacités d'interaction H/R sont encore balbutiantes en robotique : elles sont intrinsèquement limitées par les capacités dont dispose le robot pour percevoir l'homme. Les données télémétriques, informations pauvres et peu adaptées ici, sont parfois exploitées pour détecter et localiser les humains [Fontmarty 2007, Maas 2006, Siegwart 2003]. Les données visuelles contiennent certes plus d'information mais elles restent encore sous-exploitées dans les mécanismes d'interaction H/R malgré des avancées significatives [Bennewitz 2005, Maas 2006]. Pourquoi ces « réticences » à la vision ? Citons : (i) le champ de vue *a priori* restreint depuis des caméras embarquées, (ii) les temps de traitement souvent prohibitifs eu égard aux ressources CPU embarquées... par définition limitées, (iii) le souci de réactivité, condition nécessaire à l'acceptabilité du robot par l'homme.

Ces capteurs extéroceptifs seront embarqués sur le robot. L'environnement dans lequel évolue le robot est donc supposé non instrumenté. Ainsi, le cadre applicatif actuel se démarque de la surveillance de scènes. Pour la surveillance, certaines contraintes mentionnées précédemment peuvent être relaxées : les informations contextuelles sont possiblement codées en dur, les capteurs instrumentant la scène permettent une perception plus complète de celle-ci, enfin les ressources CPU sont plus facilement extensibles.

Un dernier point d'achoppement concerne le paradigme sous-jacent à la construction des représentations environnementales pré-citées : faut-il doter le robot de représentations internes données *a priori* par un expert, de dimensions nécessairement élevées, ou au contraire laisser ces dernières émerger d'un apprentissage ouvert sur la base des données sensorielles ? Autrefois antagonistes, les deux paradigmes ont aujourd'hui tendance à se rejoindre. Citons deux exemples. Les projets européens CogVis[3] et Cogniron ([PR4], partie II), chacun dans leur domaine - la vision dans le premier cas, la robotique dans le second - visent à concevoir des **systèmes cognitifs**, qui en interaction avec les autres agents, ici les humains, permettent d'enrichir et d'adapter la connaissance interne du robot aux spécificités de l'environnement d'une part, aux besoins de l'utilisateur du robot d'autre part.

3.2 Description

Fort de ces considérations et dans la continuité des travaux passés et actuels du pôle RIA, la finalité de mon projet de recherche est **la perception pour la compréhension conjointe de l'espace et du milieu par un**

[3]Lien URL : www.comp.leeds.ac.uk/vision/cogvis

robot personnel autonome, cognitif, et sociable. Ce robot va classiquement agir en fonction d'un schéma planifié d'actions, s'assurer de la viabilité de celles-ci, de la cohérence des modèles courants, de la présence d'entités contrariant son plan initial, tout en interagissant avec le monde physique grâce à la perception.

La démarche retenue, qui vise clairement à l'intégration de fonctions perceptuelles sur des plate-formes robotiques, s'appuie sur une modélisation probabiliste afin de considérer des percepts multiples et incertains. Au niveau sensoriel, ces percepts seront principalement issus de la vision embarquée sur le robot. Ces fonctions perceptuelles sont à décliner des capacités d'autonomie, de cognition et de sociabilité mises en exergue pour notre robot personnel. Les capacités d'autonomie sont surtout relatives ici à sa mobilité ; elles s'appuient sur la perception et la représentation par le robot de son espace. Les capacités de cognition et de sociabilité caractérisent ici la relation du robot à l'homme. Les capacités de cognition permettent, en lien étroit avec l'homme, de magnifier la connaissance de l'environnement par la capture d'informations sémantiques relatives au milieu. La construction de la représentation associée est subordonnée à la perception par le robot de l'homme ; nous y reviendrons largement. Dans sa relation à l'homme, le robot cognitif ne cherche pas systématiquement ici à le satisfaire contrairement au robot sociable qui vise la mise en commun d'actions ou d'objectifs. Dans la perspective du robot sociable, la perception par le robot de l'homme, plus largement du milieu et de l'espace, est fondamentale. Reprenons et détaillons ces divers aspects.

Intégration sur des plate-formes robotiques — Notre démarche scientifique est intégrative puisque la finalité est d'implanter des fonctions perceptuelles, après prototypage, à bord de véritables robots. Elle représente certes un investissement considérable en temps mais elle donne tout son sens au terme d'agent autonome, cognitif et sociable, qui est par définition en situation dans son environnement. Cette voie implique tout d'abord l'appréhension préalable de l'architecture logicielle du robot. L'architecture mise en œuvre sur les robots du laboratoire est structurée en couches et modules (figure 1.4), ces derniers communiquant *via* un protocole de partage d'informations appelé poster. La démarche est ensuite d'implanter nos fonctions dans la couche fonctionnelle du robot puis d'en évaluer la robustesse dans le cadre de scénarios réalistes prenant en compte explicitement la présence de l'homme au voisinage du robot et/ou en interaction avec lui. Les contraintes de la réalité sont celles du « temps-réel »[4], de l'incertitude et de l'incomplé-

[4]Ressources CPU limitées, temps de réponse bornés pour des besoins de réactivité, etc.

tude des connaissances, et de la diversité des situations rencontrées par le robot mobile.

Conception, intégration et évaluation des algorithmes développés sont donc intimement liées afin de juger (notamment) de leur robustesse :

- à des situations diverses et variées, couloirs *a priori* sur/ou sous-éclairés, zones encombrées et/ou étroites ou autres de son espace de travail ;
- aux occultations ou pertes d'observabilité des entités animées ou inanimées perçues par le robot dans son milieu ou espace.

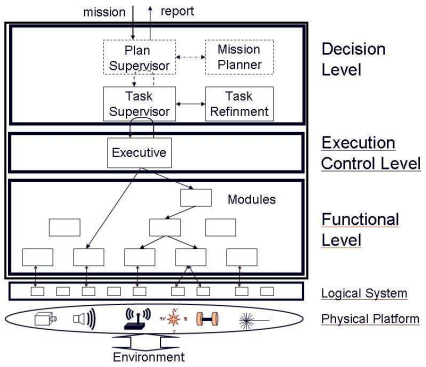


FIG. 1.4 – Architecture logicielle Genom [Alami 1998].

Ces algorithmes s'appuient sur des données sensorielles issues de capteurs extéroceptifs qui sont caractérisés ci-après.

La vision le mode de perception privilégié — S'agissant des capteurs utilisés, nous privilégions actuellement les systèmes de vision monoculaire, multi-oculaire ou stéréoscopique car ils délivrent des informations très riches sur l'apparence et la géométrie des scènes perçues. Les distributions de couleur et/ou de texture, contours, segments, points d'intérêt pour l'apparence ou encore disparité, reconstruction 3D dense ou éparse pour la géométrie sont autant d'indices à fusionner dans la chaîne perceptuelle gérée par le robot. L'utilisation de microphones pour la reconnaissance de la parole vient naturellement compléter la reconnaissance de gestes par vision pour permettre une interaction H/R multimodale. L'utilisation du capteur laser, peu compatible avec la présence humaine, est ici marginale et limitée à de simples coupes à très faibles hauteurs du sol (laser SICK 2D). Il est logiquement dévolu aux tâches relevant de la perception de l'espace : détection de l'espace libre, des obstacles, etc. Enfin, des nouvelles technologies de capteurs (caméras actives[5], RFID), éventuellement communicants, sont également à envisager dans mon projet.

Concernant leur déploiement dans l'environnement, nos travaux anté-

[5]De type *photodetector measurement device* ou *time-of-flight camera.*

rieurs et actuels considèrent des capteurs majoritairement embarqués sur les plate-formes robotisées. L'avènement de l'intelligence ambiante, probablement une des grandes aventures technologiques futures, nous amènera à considérer des réseaux de capteurs pouvant instrumenter une multitude de supports (robots mais aussi mobilier et vêtements). La démarche sera alors de prendre en compte, autant que possible, ces divers flux perceptuels pour répondre aux besoins de robustesse exprimés en préambule.

Intégration de percepts multiples et incertains — Une préoccupation majeure de mon projet est, en effet, d'intégrer des percepts multiples et incertains, ceci à tous les niveaux de la conception des fonctions développées. Au niveau sensoriel, il s'agira de coupler des données extéroceptives homogènes ou hétérogènes lorsqu'elles sont issues de capteurs de natures différentes. Au niveau fonctionnel, il s'agira de coupler les hypothèses de détection, localisation, reconnaissance ou autres inférées par diverses fonctions perceptuelles. Dans les deux cas, nous parlerons de couplage fort ou combinaison, et de couplage faible ou fusion, selon que l'hypothèse finale résulte d'un seul et unique processus de traitement combinant les percepts ou de plusieurs processus exécutés en parallèle et dont on fusionne les résultats associés. Donnons deux illustrations. Le couplage fort de données visuelles et télémétriques dans un processus de localisation du robot donnera lieu à un seul processus d'estimation. Le couplage faible de données visuelles et vocales dans un mécanisme d'interaction H/R multimodale consistera à exécuter deux processus d'interprétation dédiés puis d'en fusionner les hypothèses.

Ces paradigmes de combinaison/fusion sont largement mis en œuvre dans ce projet, notamment pour la perception par le robot de l'homme donc d'entités *a priori* très animées. En réponse, il s'agit de proposer des fonctions perceptuelles réactives, donc incluant des boucles rapides, tout en intégrant, aussi largement que possible, des percepts totalement hétérogènes. Par exemple, un processus de suivi visuel pourra « mieux » pallier les décrochages et variabilités de la cible en intégrant des percepts multiples sensoriels (vision, son, RFID) et fonctionnels (détection de « blobs peau », reconnaissance faciale ou autres).

L'intégration de percepts multiples concerne également la compréhension conjointe de l'espace et du milieu et donc les deux représentations associées. Celles-ci visent à capturer et à connecter les structures métriques et topologiques pour l'espace, conceptuelles pour le milieu. Les informations sous-tendant ces représentations doivent permettre par recoupement de « compenser » les incomplétudes ou incertitudes sur la scène à analyser. Donnons deux illustrations :

- la reconnaissance d'objets permet d'inférer sur la nature du lieu, les actions humaines associées attendues... et réciproquement.
- L'analyse de chroniques, à partir des deux représentations, doit permettre l'interprétation d'activités humaines mettant en jeu plusieurs places, lieux, éventuellement objets.

L'intégration de percepts multiples nous semble également nécessaire dans une démarche descendante dès lors que certains percepts inférés à partir de ces représentations doivent influer sur les processus d'acquisition et de traitement des données sensorielles au niveau bas. Nous parlerons de **stratégies actives de perception**. Ainsi, la réalisation d'une tâche de navigation ne peut pré-supposer l'utilisation d'un capteur unique et adapté indépendamment de tout contexte environnemental : la caractérisation préalable de ce contexte courant (nature du lieu, encombrement ou autre) doit inférer sur l'activation de tel ou tel capteur ou stratégie de perception. De même, la stratégie de perception de l'homme pour l'interaction H/R dépend aussi du contexte environnemental : dans un espace ouvert, la segmentation de l'homme au voisinage immédiat du robot doit plutôt privilégier des informations géométriques donc le(s) capteur(s) 3D embarqué(s). *A contrario*, dans un espace confiné (typiquement un couloir), il semble plus opportun d'exploiter des informations d'apparence donc non nécessairement issues de capteurs 3D.

Modélisation probabiliste — Pour décrire/quantifier l'imprécis et l'incertain associés à ces percepts multiples, nous privilégions une modélisation probabiliste. L'application principale des modèles probabilistes est l'inférence bayésienne qui est la démarche logique pour calculer (ou réviser) la distribution de probabilité pour des variables inconnues sur la base de variables connues (évidence). Cette démarche est régie par l'utilisation de probabilités conditionnelles, desquelles dérive le théorème de Bayes. Dans la perspective bayésienne qui est l'approche privilégiée ici, une probabilité n'est pas interprétée comme le passage à la limite d'une fréquence, mais plutôt comme la traduction numérique d'un état de connaissance, typiquement le degré de confiance accordé à une hypothèse. On parle ici d'interprétation bayésienne, ou subjective, par opposition à fréquentiste. Les filtres de Bayes sont en particulier adaptés aux systèmes dynamiques rencontrés par le robot lors de son interaction avec le monde. Citons ici quelques outils intégrant cette dimension temporelle, outils très répandus et avec de nombreuses variantes : filtre de Kalman [Kalman 1960], filtre particulaire[6] [Arulampalam 2002], modèle de Markov caché [Rabiner 1989] ; nous y reviendrons ultérieurement.

[6]Pour une inférence approchée.

Les fonctions perceptuelles à embarquer sur le robot, s'appuieront donc sur une modélisation probabiliste afin d'intégrer des percepts incertains et multiples tirés pour l'essentiel de la vision. Caractérisons maintenant ces fonctions eu égard des besoins exprimés en termes d'autonomie, de cognition et de sociabilité.

Perception de l'espace pour un robot autonome — À l'instar de l'homme qui ne résume pas l'espace à une simple représentation métrique, nous envisageons diverses représentations pour exhiber la métrique et topologie de l'espace exploré par le robot. Parmi les représentations métriques, citons classiquement les grilles ou cartes d'occupation, et les cartes stochastiques d'indices perceptuels. La construction de ces cartes, souvent référencée par l'acronyme SLAM (*Simultaneous Localization and Mapping*), est un problème, par nature, mal conditionné puisqu'une localisation précise nécessite un modèle de scène bien reconstruit alors qu'une bonne reconstruction de ce modèle requiert une estimation précise de la localisation du robot. Pour briser ce cercle vicieux, la démarche consiste à capturer des amers dans les diverses représentations métriques évoquées. *A contrario*, les représentations topologiques correspondent à une modélisation qualitative discrète de l'espace et s'exprime classiquement sous forme de graphes. Un graphe d'amers ou de zones (places), correspondant éventuellement à des objets ou lieux, rend compte de relations topologiques (connectivité, relation de tout à partie, etc) entre ces différentes entités[7]. Citons le Graphe de Voronoï Généralisé qui est largement utilisé dans ce contexte.

Au final, la représentation spatiale exhibe des percepts multiples en faisant coexister en permanence, et à différents niveaux d'abstraction, différents modèles métriques et topologiques. Cette représentation hybride, destinée aux environnements d'envergure, permet de tirer profit des avantages des deux modèles : compacité et simplicité pour le premier, précision de localisation pour le second. Son exploitation durant ses tâches de navigation permet alors au robot autonome de se localiser en relatif ou en absolu, de manière métrique ou symbolique. Cette représentation est construite de manière ascendante et automatique par le robot, donc sans intervention humaine... *a contrario* de la représentation du milieu.

Perception du milieu pour un robot cognitif — La représentation du milieu doit exhiber la sémantique relative aux concepts de lieux, d'objets et d'actions humaines associées. L'apprentissage et catégorisation automatique d'objets ou de lieux par leurs structures (apparence et/ou géométrie)

[7]Le plan de métro de Paris forme un bon exemple de graphe topologique.

a suscité et suscite encore de nombreux travaux au sein de la communauté Vision, et plus spécifiquement dans notre groupe. Aussi, cette problématique sort du champ de mes investigations même si l'ajout d'une représentation fonctionnelle nous semble une extension pertinente pour mieux catégoriser les objets ou lieux ; nous y reviendrons dans le chapitre 3.

La démarche retenue est ici semi-automatique au sens *où* l'homme intervient à deux niveaux dans la construction de cette représentation. Un expert peut, hors-ligne, figer (« hard-coder ») certaines connaissances sémantiques immuables, comme l'association prévisible entre certaines classes d'objets et de lieux. Il ne peut hélas donner une connaissance exhaustive du monde. La volonté d'un apprentissage ouvert pour le robot cognitif, la perspective du robot personnel, implique l'interaction avec un humain non-expert dont la mission sera : (i) de guider le robot dans une stratégie active de prises de vues des structures locales à caractériser afin de faciliter leur segmentation, (ii) de catégoriser les instances de lieux ou d'objets par la voix. La construction de cette représentation du milieu étant subordonnée au développement de fonctions de perception multi-sensorielle de l'homme par le robot, nos investigations passées et actuelles se sont focalisées en grande partie sur ce dernier point.

Perception de l'espace et du milieu pour un robot sociable — L'exploitation par le robot de ses représentations relatives au milieu et à l'espace, ses capacités perceptuelles de l'homme doivent lui permettre aussi de communiquer et interagir avec les humains partageant l'environnement. Ces fondements nous semblent des pré-requis indispensables à la communication ou l'interaction H/R. Le robot pourra par exemple planifier des tâches symboliques, typiquement exécuter une action à partir d'une commande par le geste et/ou la parole de l'homme *e.g.* « va chercher l'objet X dans le lieu Y ». Plus largement, L'exécution de tâches au voisinage immédiat ou en interaction avec ce dernier visera à satisfaire ce dernier et donc, pour le robot, à agir en acteur sociable.

La figure 1.5 montre à titre d'illustration une ébauche possible de cette représentation environnementale, ses différents niveaux métriques, topologiques, et conceptuels. Cette représentation est inspirée de [Galindo 2005] qui exploite majoritairement des données laser SICK 2D pour modéliser, par télé-opération du robot, un environnement très sommaire.

Mes investigations passées et actuelles s'inscrivent naturellement dans le projet énoncé ci-dessus. Celles-ci s'articulent autour des deux thèmes énoncés en préambule et dont nous pouvons maintenant préciser les contours. Rappelons ces deux thèmes :

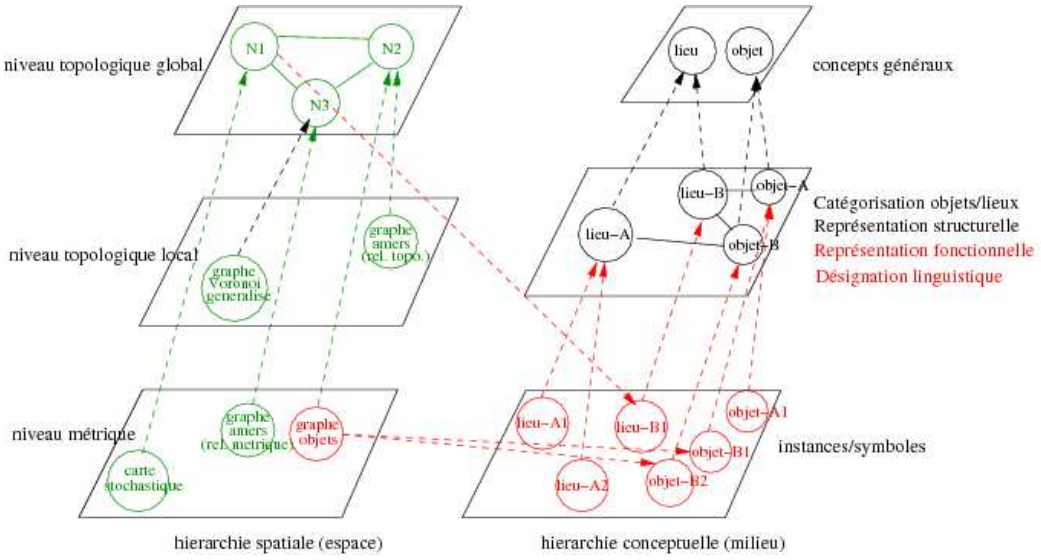


FIG. 1.5 – Ébauche d'une représentation environnementale possible [Galindo 2005] : connaissance *a priori* (en noir), construction automatique (en vert), supervisée (en rouge).

1. **Perception de l'espace pour la navigation autonome.**
2. **Perception de l'homme pour l'interaction H/R.**

Mon implication sur le thème $n°$ 1, en collaboration avec M.Devy (DR CNRS, groupe RAP), est consécutive à mon arrivée au LAAS-CNRS. Mon souhait est aujourd'hui encore de poursuivre ces investigations sur la représentation de l'espace en synergie avec mes collègues et donc de s'appuyer sur leurs expertises et les investigations connexes au thème $n°$ 1. Mes premiers travaux, détaillés au chapitre 2, ont porté sur la perception de l'espace pour la navigation sur amers à partir d'une représentation hybride de l'environnement, *i.e.* métrique et topologique. Leurs extensions actuelles qui font l'objet de travaux dans le cadre de thèses/stages en cours et/ou sujets déposés sont déclinées dans le chapitre 3. Conformément au projet de recherche, il s'agira notamment de compléter la représentation environnementale par une représentation du milieu exhibant des informations à caractère sémantique (relatives aux lieux, objets, actions humaines, etc).

Le thème $n°$ 2 a été abordé plus tardivement (2002) par la thèse de L.Brèthes (TH7, partie II) alors qu'il n'était couvert par aucun permanent du pôle. J'en suis l'initiateur au sein du pôle RIA et, en conséquence, ce thème constitue aujourd'hui ma priorité. Il fait logiquement l'objet d'un investissement plus soutenu d'autant plus que l'interaction avec l'homme, et sa

prise en compte explicite à tous les niveaux du robot-système, prend progressivement une importance grandissante dans le pôle et plus globalement dans la communauté Robotique. Ma contribution ici porte plus spécifiquement sur le développement des interfaces perceptuelles pour l'interaction H/R. Le développement d'architectures complètes[8] dédiées à nos robots assistants ou personnels est traité dans le groupe Robotique et InteractionS (RIS). Mon interlocuteur est ici R.Alami (DR CNRS, groupe RIS). De par l'émergence plutôt récente de ce thème, mes premiers travaux, détaillés au chapitre 2, ont porté essentiellement sur les fonctions suivantes de perception de l'homme : détection, reconnaissance de l'homme, suivi de tout ou partie de ses membres corporels. La prospective à court/moyen termes (thèses en cours ou sujets déposés) et au-delà est présentée dans le chapitre 3. Elle s'inscrit dans la continuité des investigations passées et actuelles et dans l'esprit du projet de recherche énoncé.

Avant de rentrer dans le vif du sujet, signalons que mes publications, encadrements, contrats industriels, projets, collaborations dont il est fait mention dans le texte, sont associés à des index (resp. [x], [THx], [CIx], [PROx], [COx]) permettant leur référencement dans la partie II sur la valorisation de mes travaux de recherche.

[8] *i.e.* impliquant toutes les couches du robot-système (figure 1.4).

Chapitre 2

Travaux antérieurs

1 Préambule

Ce chapitre décrit mes travaux antérieurs dans les deux thèmes énoncés précédemment. La section 2 est relative au thème perception de l'espace pour la navigation tandis que la section 3 est relative au thème perception de l'homme pour l'interaction H/R. Enfin, la section 4 mentionne quelques retombées connexes à ces travaux dans le cadre d'un contrat industriel majeur en terme de temps investi.

2 Perception de l'espace pour la navigation

2.1 Motivations

Pour naviguer de façon autonome, les robots du LAAS-CNRS intègrent plusieurs fonctions sensori-motrices réalisant les actions de planification de trajectoires, localisation, et contrôle du mouvement. Ainsi, le robot Diligent, support expérimental de ces travaux, contrôle ses mouvements grâce à différentes stratégies ou méthodes durant une tâche de navigation : méthode réactive par potentiels attracteur et répulsif, méthode de déformation de bande élastique le long de la trajectoire pré-planifiée. Concernant la localisation, sujet de nos investigations ici, elle s'appuie classiquement sur deux classes de capteurs : (1) les capteurs proprioceptifs qui mesurent des quantités reliées à des dérivées premières ou secondes de la position, (2) les capteurs extéroceptifs qui mesurent l'état d'une relation entre l'entité en mouvement et l'environnement extérieur. Les premiers conduisent à des problèmes de dérive lors d'une navigation au long cours. On privilégie alors les **capteurs extéroceptifs** pour se localiser en relatif ou en absolu à partir de modèles, ici métrique ou topologique, de l'environnement.

Il semble irréaliste d'exploiter la totalité des données sensorielles à la disposition du robot car la quantité d'information en jeu, pour des environnements à large échelle, peut dépasser largement les capacités de calcul de la machine. La stratégie sera donc plutôt d'exploiter, dans ce flot de données, des amers afin de : (1) limiter la complexité de la tâche de localisation, (2) éviter les erreurs constatées lorsque le modèle est seulement constitué de primitives géométriques élémentaires, (3) représenter l'environnement sous une forme plus compacte. Plusieurs travaux « historiques » du pôle RIA ont porté sur la navigation sur amers visuels en milieu extérieur [Betgé-Brezetz 1996, Jung 2004, Vandapel 2000] ou sur la navigation sur amers laser en milieu intérieur [Bulata 1996]. Nos travaux portent ici sur la **navigation sur amers visuels en milieu intérieur à l'aide d'une seule caméra**, cette

problématique sous-tendant des stratégies de perception et de représentations différentes de celles mises en œuvre par le passé. Plus largement, la communauté Robotique exploite de nombreux indices visuels pour caractériser des amers saillants. Citons ici les cartes de saillance [Itti 1998], les bases d'images [Porta 2004], les points d'intérêt [Moreels 2005], les segments de contours [Sim 1999], les invariants projectifs [Branca 2000], etc. Au-delà de ces critères de saillance, il s'agit, conformément au projet énoncé, de prendre en compte d'autres aspects fondamentaux que sont : valeur sémantique, stabilité aux changements de prise de vue, robustesse aux occultations, aptitude à la localisation, etc.

Les fonctions perceptuelles développées sont dévolues au robot Diligent dont le système de supervision permet de faire coexister - voire interférer - plusieurs fonctions sensori-motrices embarquées. En effet, nous ne pouvons pré-supposer l'existence d'une stratégie, unique et optimale dans tout contexte, de réalisation de la tâche de navigation à l'aide de ces fonctions sensori-motrices. Ainsi, diverses stratégies complémentaires sont exécutées suivant les caractéristiques courantes de l'environnement, *e.g.* la traversée d'une zone encombrée d'obstacles imprévus, d'une zone étroite, d'une zone pauvre en amers visuels ou au contraire riche de tels amers, etc. Nous ne détaillerons pas davantage car ces extensions, impliquant certes nos travaux, sortent de notre champ d'investigation. Le lecteur pourra se référer à [Morisset 2002] pour plus de détails.

Dans l'esprit du projet énoncé, les travaux réalisés durant la thèse de J.B Hayet [TH8] ont porté sur les points suivants :

- La définition d'amers visuels pertinents et adaptés aux milieux intérieurs, leur détection et reconnaissance depuis une caméra embarquée.
- La construction d'une représentation hybride (métrique et topologique) et multi-sensorielle (amers visuels et télémétriques) de l'espace.
- La navigation, possiblement multi-sensorielle, du robot à partir de cette représentation.
- L'intégration de ces fonctions perceptuelles sur le robot mobile Diligent et leurs évaluations dans des situations réalistes.

Les sous-sections suivantes reprennent ces différents points puis énumèrent les contributions associées.

2.2 Détection et reconnaissance d'amers visuels

Notre première contribution porte sur la détection, la reconnaissance, la localisation sur amers visuels, ici des quadrangles qui sont majoritairement présents en environnement d'intérieur (affiches, plans, tableaux, portes,...).

Ces entités, souvent positionnées en hauteur dans l'environnement, seront moins sujettes à occultations lorsque l'espace du robot sera « encombré » par plusieurs humains. Enfin, ces amers sont particulièrement saillants, localisables et stables. Leur saillance tient tout à la fois de la structure particulière recherchée *a priori* et de critères sur le contenu texturel de celle-ci ; la localisabilité provient de la structure géométrique de l'amer ; enfin, la stabilité va de pair avec l'hypothèse de la planarité de la structure 3D correspondante.

Le robot évolue dans un milieu intérieur alternant réseaux de couloirs et espaces ouverts avec des caractéristiques très différentes. L'exploitation des amers visuels est logiquement dépendante du contexte environnemental courant et du mode de localisation exécuté. Lorsque le robot dispose, lors de sa navigation, d'une estimée fiable sur sa position spatiale, le principe est alors de se recaler sur les contours estimés par reprojection des amers après prédiction [7]. Sans estimée initiale (stratégie dite de *kidnapping*), deux stratégies de détection d'amers sont possibles suivant les caractéristiques courantes de l'environnement. La première, plutôt dédiée aux espaces ouverts, est basée sur un double processus de relaxation pour apparier des quadruplets de segments à partir d'une image de contours [25, 26]. La figure 2.1 montre des amers extraits dans ce contexte. La seconde stratégie, adaptée aux couloirs, est plus rapide mais moins robuste. Elle traite directement l'image de niveau de gris, dont elle effectue un moyennage contrôlé par les informations structurelles à disposition [27]. Typiquement, pour les couloirs, les points de fuite permettent à faible coût de structurer les images et donc de faciliter la segmentation des amers. La figure 2.2 montre, dans ce contexte, des amers extraits à l'aide de cette stratégie exploitant ligne d'horizon et points de fuite.

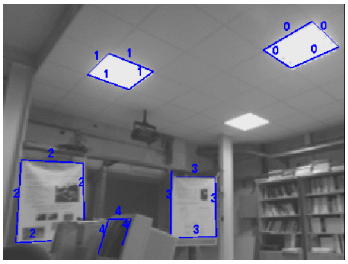


FIG. 2.1 – Détection d'amers en espace ouvert.

L'indexation des amers ainsi détectés s'appuie sur une représentation photogrammétrique de ces derniers. Une rectification homographique est appliquée pour établir une représentation iconique ou imagette des amers invariante aux changements de points de vue et aux paramètres de la caméra.

Pour la reconnaissance, trois caractérisations robustes de ce modèle iconifié sont proposées dans [25] pour calculer la distance d'une imagette inconnue $\mathcal{Q}$ à la classe l liée à l'amer C_l. La première notée $\mathscr{C}$ — classique — repose sur la décomposition régulière du motif rectifié en sous-images et l'utilisation d'une mesure de corrélation partielle robuste ; la seconde $\mathscr{H}_2^f$ considère la distribution spatiale des points d'intérêt détectés sur l'imagette et compare de tels ensembles à l'aide de la distance partielle de Hausdorff ; la troisième

$\mathscr{H}_p^f$, relativement originale, ajoute à la précédente des invariants différentiels normalisés venant pondérer les distances euclidiennes dans le calcul de la distance de Hausdorff précédente.

Une procédure d'apprentissage semi-supervisée permet d'extraire, par Analyse en Composantes Principales (ACP) sur une séquence d'images contenant un ou plusieurs amers potentiellement intéressants, un ensemble réduit de vues iconifiées représentatives de chaque amer (classe). De la distance entre ces vues représentant C_l et l'imagette inconnue $\mathcal{Q}$, nous dérivons une estimation heuristique des probabilités *a posteriori* d'appartenance à chaque classe ainsi qu'une mesure de confiance sur la classification à partir des probabilités des classes perdantes. Une évaluation approfondie du comportement des mesures $\mathscr{C}^f$, $\mathscr{H}_p^f$, $\mathscr{H}_d^f$ vis-à-vis de conditions de prises de vue variables et réalistes vis-à-vis du contexte applicatif a permis de souligner les bonnes performances des représentations $\mathscr{C}^f$ et $\mathscr{H}_p^f$. Nous avons au final privilégié la mesure $\mathscr{H}_d^f$ pour la compacité de la représentation associée. Ces travaux ont donné lieu à de nombreuses publications. Citons en particulier [5] et [27] qui sont jointes en fin du mémoire.

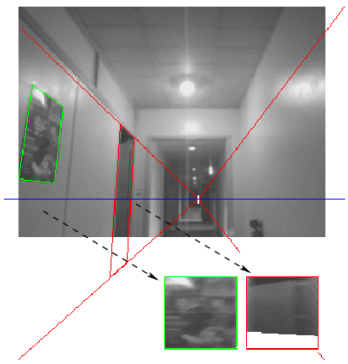


FIG. 2.2 – Détection d'amers en couloir.

2.3 Représentation de l'espace

Pour naviguer, le robot exploite une représentation de son espace qu'il construit automatiquement durant une phase préliminaire d'exploration de l'environnement. Comme évoqué dans le projet, cette représentation, tout en exploitant la détection et reconnaissance d'amers visuels, doit exhiber à différents niveaux d'abstraction des modèles :

- **métriques** *e.g.* grille d'occupation [Galindo 2005], carte stochastique [Pfaff 2006], etc.
- **topologiques** *e.g.* graphe de lieux, d'amers [Booij 2007] ou d'objets [Vasudevan 2006], graphe de Voronoï généralisé [Choset 1997, Victorino 2004], etc.

Cette représentation dite hybride sera également multi-sensorielle au sens *où* elle s'appuiera sur des données perceptuelles hétérogènes à l'instar de [Durieu 1996, Stachniss 2005, Vasudevan 2006]. Nous nous démarquons ainsi des représentations spatiales mono-capteur ; citons, sans être exhaustif, le laser dans [Pfaff 2006, Thrun 1998], la vision dans [Goncalves 2005, Zivkovic 2006] ou les ultrasons dans [Choset 1997]. Une **carte stochastique multi-**

sensorielle associant des indices perceptuels denses type segments laser et épars type amers visuels est ainsi construite et exploitée par les modules de base du robot Diligent. Les amers sont positionnés *a posteriori* dans la carte stochastique en fusionnant les localisations visuelles et laser. La construction de la carte d'amers est abordée par deux voies exploratoires : une approche classique de type SLAM et filtre de Kalman étendu, enfin une approche de reconstruction photogrammétrique d'environnements structurés [Montiel 2001]. La figure 2.3 montre un exemple de carte laser obtenue par SLAM et incluant les amers appris et localisés dans la carte ainsi que leurs zones de visibilité associées. Quatre exemples de représentations iconifiées des amers sont montrés.

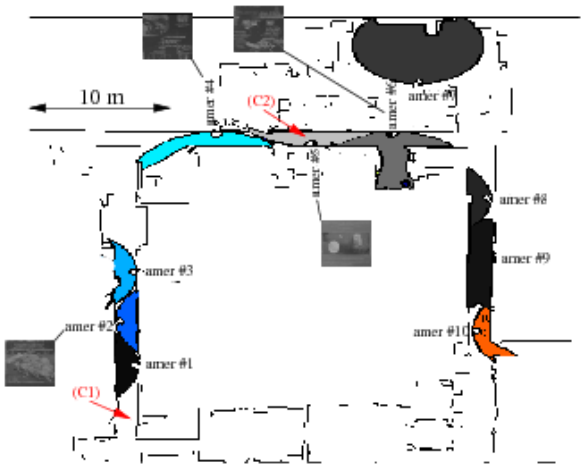


FIG. 2.3 – Une carte métrique multi-sensorielle d'un espace ouvert : segments laser et amers visuels (représentations iconifiées -imagettes- et zones de visibilité associées -couleurs-).

Ces modèles restent malheureusement dépendants de la connaissance métrique de l'environnement, contrairement aux modèles topologiques plus compacts et plus abstraits car construits à partir de raisonnements spatiaux et qualitatifs sur l'environnement. Nos travaux sur la navigation qualitative ont démarré lors d'une collaboration avec l'Université d'Urbana-Champaign (CO3, partie II). Le modèle topologique est ici structuré en un graphe de lieux, modélisé par un **Graphe de Voronoï Généralisé** (GVG), construit à l'aide de capteurs de distance *i.e.* ultrason ou laser. Le GVG est défini comme l'ensemble des points équidistants de deux obstacles (arêtes du graphe) ou de trois ou plus (noeuds du graphe ou *meet points*). Les noeuds représentent classiquement les éléments saillants et structurants de l'environnement, typi-

quement les intersections de couloirs, passages étroits, impasses [Kortenkampf 1994, Choset 1997]. La construction consiste à explorer tous les chemins possibles, dans un environnement de type réseau de couloirs, en mémorisant la structure globale des noeuds détectés. Le contrôle du mouvement du robot sur un arc du graphe consiste à préserver l'égalité des distances aux deux obstacles les plus proches ; nous nous sommes reposés ici sur une méthode éprouvée dans [Victorino 2001].

Nous avons implémenté sur le robot une première stratégie visant à annoter les noeuds du graphe par des amers visuels [7, 30], la détection des *meet points* étant cruciale pour la construction du GVG. La figure 2.4 montre un exemple de (courte) expérimentation dans le couloir C_2 (figure 2.3) : le GVG est ici construit à l'aide de capteurs ultrasonores dont les données sont représentées en vert, de même que les segments laser en rouge ; le robot arrive par la gauche et explore successivement deux noeuds correspondant aux extrémités du couloir. La dérive odométrique est importante mais sans conséquence pour la construction du GVG. Le capteur ultrasonore ayant de nombreuses

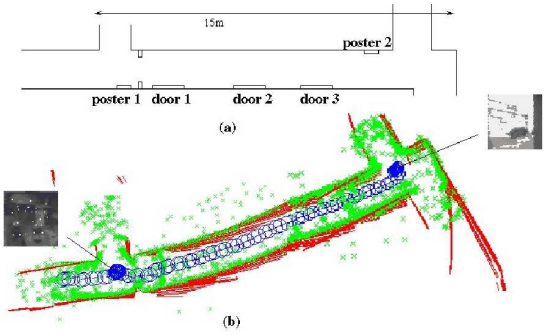


FIG. 2.4 – Annotation des noeuds du GVG relatif au couloir C_2.

limitations, nous avons rapidement privilégié le capteur laser SICK 2D dans la construction du GVG. La présence et la nature d'un noeud du GVG au-devant du robot repose sur une heuristique de prédiction/vérification [DEA4]. Une série d'hypothèses, selon la banque de modèles illustrés en figure 2.5 est utilisée pour reconnaître le noeud approchant. La mise à jour des hypothèses courantes a lieu par détection de changements de la structure topologique du couloir. Ces changements (ou événements), relatifs à la discontinuité dans un mur ou sa terminaison, permettent par la suite d'identifier les noeuds du GVG.

Deux extensions sont proposées au final pour compléter le GVG : (1) l'annotation, par des amers visuels, des arcs du GVG... en plus de ses noeuds, (2) la capture des relations topologiques entre ces amers à partir de l'ana-

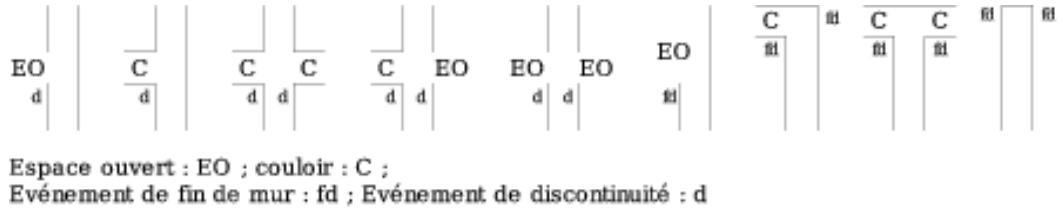


FIG. 2.5 – Principaux types de noeuds rencontrés dans un couloir.

lyse du flot vidéo. Le lecteur pourra, pour plus de détails, se référer à la communication [27] jointe en fin du mémoire.

Les environnements humains, publics ou privés, sont structurés en réseaux de couloirs connectant des espaces ouverts. L'alternative proposée est donc de faire coexister ces modèles dans une représentation hybride de l'espace *i.e.* un ou plusieurs GVG pour les couloirs et des cartes stochastiques ou graphes d'amers pour les espaces ouverts [26]. Nous distinguons ici les cartes stochastiques qui sont des cartes métriques denses d'indices perceptuels, éventuellement hétérogènes, de graphes d'amers (ou d'objets lors de futures investigations) qui considèrent des indices épars avec des informations métriques ou qualitatives entre amers deux à deux. La commutation entre les différents modèles énumérés est notifiée par le superviseur lors du franchissement de portes. Leur détection s'effectue dans un formalisme probabiliste type réseaux bayesiens qui combinent des heuristiques géométriques et photogrammétriques extraites dans les images [DEA2]. Certes, les portes, prises individuellement, sont des amers peu discriminants. Le principe est alors d'exploiter des informations contextuelles *i.e.* des relations de voisinage entre amers pour les différencier lors de leur détection.

2.4 Stratégies de navigation

La navigation au long cours nécessite pour le robot d'exploiter ses différentes représentations spatiales pour se localiser qualitativement ou métriquement. De nombreuses expérimentations ont validé la localisation métrique sur amers visuels du robot Diligent dont l'architecture est illustrée par la figure 2.6, le module « LOC-POST » incluant détection, reconnaissance, et localisation métrique sur amers. Ces expérimentations valident indirectement la localisation qualitative, celle-ci mettant en jeu la détection et reconnaissance des amers... sans estimation précise de la position du robot.

La stratégie de navigation métrique à bord de Diligent se décline comme suit. Sur requête du superviseur, le module de localisation entre dans un

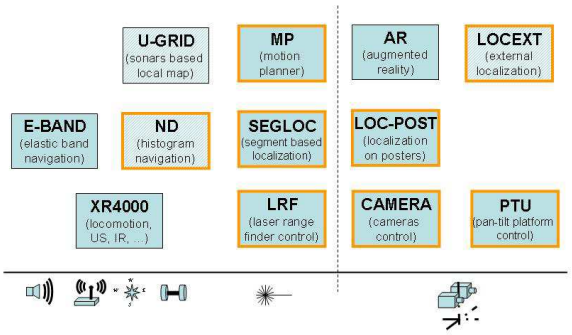


FIG. 2.6 – Architecture du robot Diligent (couche fonctionnelle).

mode de surveillance d'entrée/sortie de zone de visibilité. Lorsqu'une entrée de zone est constatée, une procédure de pointage sur l'amer (grâce à la platine site-azimut embarquée) permet de centrer la caméra sur la position estimée courante de l'amer afin de permettre sa reconnaissance éventuelle. Évaluée hors ligne puis sur le robot, les erreurs de localisation observées sur les amers reconnus sont de l'ordre de quelques centimètres, tandis qu'un ordre de grandeur de la validité des localisations visuelles est spécifié. Ces évaluations sont accessibles dans la communication [5] jointe en fin du mémoire.

La localisation calculée et son incertitude associée sont transmises au superviseur pour intégration éventuelle avec les autres modalités de localisation (modules « SEGLOC », éventuellement « LOCEXT », figure 2.6), conformément à la ligne de conduite énoncée. L'incertitude prend en compte la mesure de confiance sur la reconnaissance et la variance sur la position estimée. L'intégration ou non des localisations extéroceptives dépend de la topologie du lieu couramment perçu. Typiquement, dans un couloir, le superviseur du robot privilégie la localisation visuelle et inhibe la localisation par le laser, ce dernier induisant de fortes imprécisions sur la position du robot dans la direction du couloir. Un espace ouvert est représenté par une carte stochastique multi-sensorielle, voire par un graphe d'amers visuels... si ceux-ci sont omniprésents dans la zone modélisée. Dans le premier cas, la localisation du robot est inférée par fusion de toutes les estimées ou par combinaison des données télémétriques et visuelles dans une seule estimée unifiée. Dans le second cas, la navigation sur le graphe se définit par une séquence d'amers que le robot doit percevoir durant l'exécution de la trajectoire planifiée. Le contrôle

du mouvement du robot est réalisé par des fonctions asservies capteurs à partir d'information 2D - les consignes image étant les amers reconnus - ou 3D impliquant une localisation relative robot/amers. Dans ce dernier cas, l'idée d'une double boucle suivi-localisation, ébauchée dans [34] prend tout son sens. Elle inclut un processus lent de localisation sur le ou les amer(s) courant(s) et un processus rapide de suivi d'amer.

À titre d'illustration, la figure 2.7 montre un exemple de localisation purement visuelle sur le couloir C_1 de longueur 25m (figure 2.3). La figure de gauche indique les traces comparées des positions du robot fournies par localisation visuelle (en rouge) et odométrie (en bleu) qui dérive « naturellement » tout le long du couloir. Les deux autres figures montrent le robot dans son environnement et le point de vue courant depuis sa caméra embarquée.

2.5 Contributions

Les travaux sur ce thème mettent l'accent sur une méthodologie **complète** de détection, reconnaissance et localisation sur amers visuels validée par des expérimentations réelles sur la plate-forme Diligent. Les contributions principales, dans l'esprit du projet énoncé, sont :

- le choix d'amers visuels, quadrangulaires et plans qui nous permet de combiner des propriétés géométriques (forme, rectification homographique) et photogrammétriques (texture, points d'intérêt avec invariants différentiels) garantissant stabilité, localisabilité et saillance. Ces amers naturellement présents dans les environnements humains.
- Le développement de modalités complémentaires de détection de ces quadrangles dédiées respectivement aux couloirs ou espaces ouverts.
- La définition de métriques associées à la représentation des amers permettant au processus de reconnaissance d'être plus robuste aux changements de conditions d'observation (point de vue, éclairement, occultations partielles, etc). Les validations sur données synthétiques et réelles mettent en lumière l'intérêt de ce processus.
- Le développement de processus de localisation multi-sensorielle, dépendant du contexte environnemental, pour la navigation métrique du robot.
- La construction de représentations qualitatives de l'espace à partir de graphes et surtout leur annotation par des amers visuels. Leur exploitation dans le cadre d'une navigation qualitative au long cours sera à approfondir dans le cadre du projet énoncé.

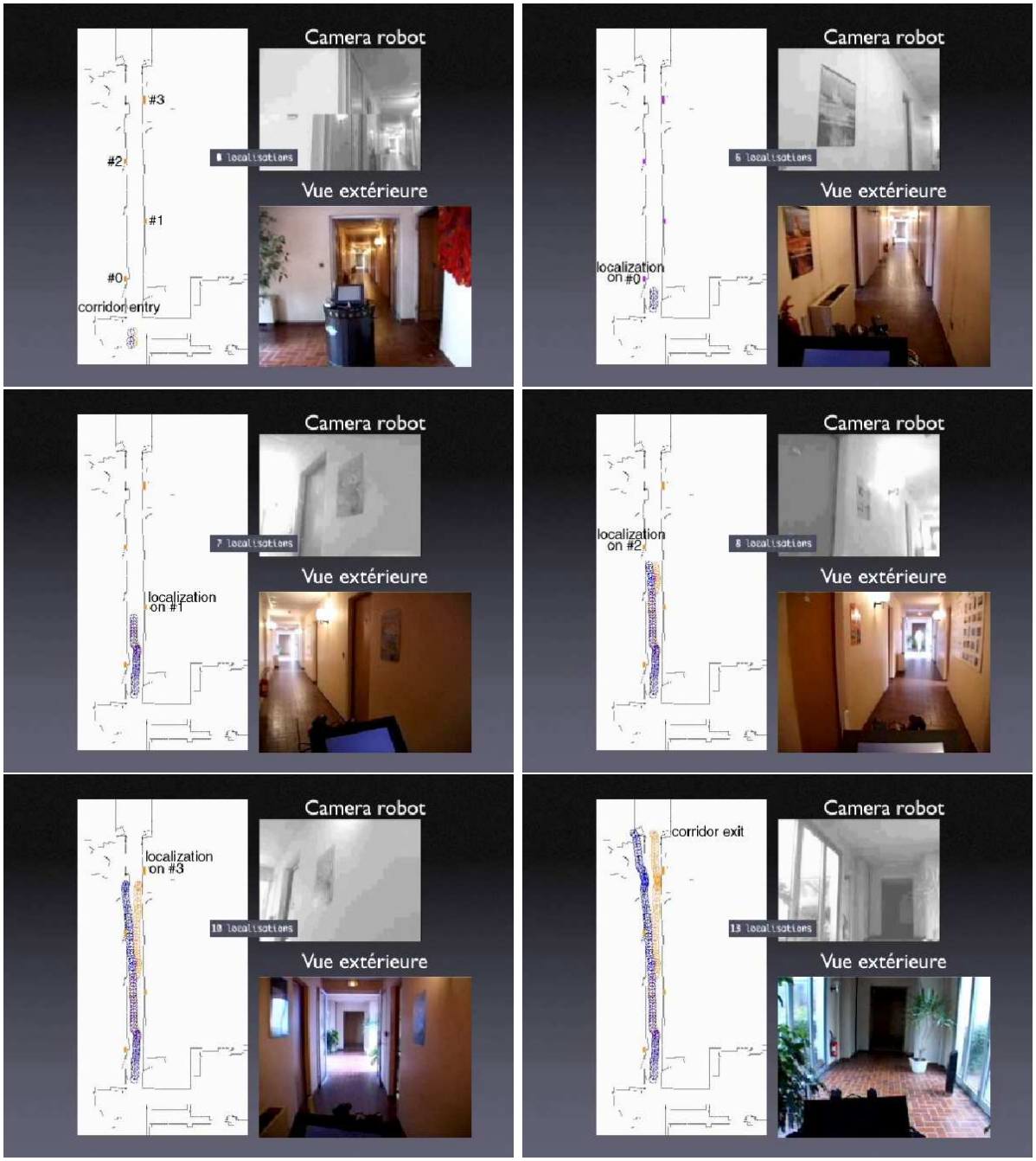


FIG. 2.7 – Recalage en position de Diligent dans le couloir C_1 à partir d'amers visuels.

3 Perception de l'homme pour l'interaction

3.1 Motivations

L'émergence de la robotique d'assistance à l'homme vient aujourd'hui enrichir la problématique générale du robot mobile autonome en mettant l'accent sur des plate-formes expérimentales qui partagent avec l'homme, l'espace, la tâche, la décision. Les défis sous-jacents prennent une place grandissante dans les divers appels à projets robotiques européens ou nationaux[1] et les conférences robotiques consacrées ou récentes telles que la conférence *Human Robot Interaction* (HRI) créée en 2006.

La problématique de la robotique en environnement humain est relativement jeune mais fédératrice au sein du pôle Robotique du LAAS-CNRS car transverse à ses trois groupes Robotique : RAP, RIS, et GEPETTO qui traite de la robotique humanoïde. Ce nouveau défi a été abordé dans le pôle, en 2001 avec le projet RobEA HR+ [PR5]. Mes investigations sur la perception de l'homme ont démarré dans la foulée *via* un stage [DEA6] puis les thèses de P.Menezes [TH6] et L.Brèthes [TH7].

La perception de l'homme me semble plus complexe que la perception de l'espace dont nous avons posé quelques jalons précédemment. Deux raisons (au moins) peuvent être avancées :

1. La géométrie des membres corporels est gauche et déformable tandis que le corps humain n'offre pas d'indices perceptuels remarquables par manque de texture naturelle de la peau. L'apparence vestimentaire, certes discriminante, ne peut être pré-supposée connue dans notre contexte robotique.
2. La perception de l'homme exige beaucoup de réactivité et donc la mise en œuvre de processus rapides à tous les niveaux de la chaîne perceptuelle : acquisition, traitement, analyse, et interprétation des données sensorielles. Cette contrainte est inhérente aux tâches ou actions robotiques qui sous-tendent la perception de l'homme : communication ou interaction H/R, manipulation d'objets et déplacement en présence de l'homme, etc.

La plupart des tâches ou actions pré-citées (et plus largement) ont besoin de caractériser la relation d'une plate-forme mobile aux agents humains *a priori* mobiles. Elles s'appuient majoritairement sur des fonctions perceptuelles dont le but est l'analyse spatio-temporelle , donc dans le flot perceptuel, de l'homme ou de ses membres à partir de la perception embarquée.

[1] *e.g.* les programmes : *advanced robotics* de l'appel d'offre IST *call* 6 ou ANR 2006 « Systèmes Interactifs et Robotiques ».

Nous parlerons de fonctions de suivi ou *trackers*. Forts de ce constat, nos travaux passés sur la perception de l'homme ont porté majoritairement sur le **suivi à partir de vision monoculaire couleur** embarquée sur nos robots assistant Rackham et personnel Jido (figure 1.1). La « panoplie » de *trackers* à prototyper est à mettre en regard des tâches ou actions robotiques associées. Typiquement, une tâche de manipulation d'objets coordonnée avec l'homme implique un suivi 3D de ses mains. L'imitation de gestes par un robot humanoïde justifie le suivi 3D ou capture de mouvement de l'ensemble des ddls du corps humain. *A contrario*, placer/déplacer le robot de façon sociable relativement à l'homme requiert une estimation grossière de la position de l'homme que l'on peut inférer grâce à un suivi dans le plan image. De même, le guidage de l'homme par un robot nécessite un suivi image dès lors que l'objectif est « simplement » de rester en contact visuel avec l'homme durant l'exécution de la mission.

Pour répondre à ces besoins, les deux thèses passées ont porté essentiellement sur le prototypage de *trackers* visuels 2D et 3D. Leur intégration permet d'exhiber des fonctions perceptuelles complémentaires sur le robot-système : très forte réactivité pour le 2D, précision de l'estimation pour le 3D. Ainsi, les travaux menés dans [TH7], thèse co-encadrée avec P.Danès (MCF UPS, groupe RAP), portent sur la détection et le suivi image de personnes, l'interprétation 2D de gestes élémentaires. Les travaux menés dans [TH6] traitent du suivi 3D des membres corporels humains.

Ces travaux partagent néanmoins quelques similitudes dans leurs développements. Rappelons, tout d'abord, qu'ils sont dévolus à une même plateforme mobile censée évoluer dans des environnements dynamiques, évolutifs, et *a priori* encombrés. Les fausses mesures induites par le milieu (présence de plusieurs individus, arrière-plan encombré), la nature de la cible observée (structures articulées, gauches, déformables, et plutôt de révolution) entraînent des singularités dans le lien état-mesure. Typiquement, le filtre de Kalman [Kalman 1960], basé sur les deux premiers moments des distributions, est pris en défaut lorsque ces singularités surviennent [Deutscher 1999]. De nombreux travaux insistent sur l'intêrêt de gérer à chaque instant plusieurs hypothèses sur les paramètres à estimer ; citons par exemple [Deutscher 2001, Sminchisescu 2003]. Ce processus d'estimation doit s'appuyer sur des fonctions d'observation discriminantes donc exploitant différentes sources de mesures. Enfin, rappelons que l'implémentation de ces *trackers* doit répondre à des contraintes temporelles fortes et induire une faible consommation en ressources CPU sur le robot.

3.2 Filtrage particulaire et intégration de données sensorielles

Les techniques de filtrage particulaire sont devenues très populaires depuis 10 ans dans les communautés du Traitement des Signaux et de la Vision par Ordinateur. Elles nous semblent particulièrement adaptées au suivi 2D ou 3D de personnes ou de ses membres corporels dans un contexte robotique. Trois arguments (au moins) peuvent être avancés :

1. Elles permettent de s'affranchir de toute hypothèse restrictive quant aux distributions de probabilités entrant en jeu dans la caractérisation du problème.
2. Le filtrage particulaire offre une grande généricité en terme de stratégie de filtrages [Arulampalam 2002]. Les nombreuses variantes doivent permettre de répondre aux diverses modalités d'interaction envisagées pour le robot. On notera cependant que peu d'évaluations comparatives entre les différentes stratégies de filtrage particulaire sont proposées dans la littérature. La démarche est souvent de comparer au seul algorithme de CONDENSATION[2] [Li 2002, Rui 2001, Torma 2003].
3. Ce formalisme permet une intégration simple, cohérente, et probabilistiquement justifiée des informations issues de différentes sources de mesures. Malgré ce constat, la fusion de données par filtrage particulaire est assez peu exploitée et souvent confinée à un nombre restreint de primitives visuelles. Citons les travaux [Pérez 2004] pour le suivi 2D, et [Sminchisescu 2003] pour le suivi 3D.

Les techniques de filtrage particulaire sont des méthodes de simulation séquentielles de type Monte Carlo permettant l'estimation du vecteur d'état d'un système Markovien non nécessairement linéaire soumis à des excitations aléatoires possiblement non Gaussiennes [Arulampalam 2002]. En tant qu'estimateurs Bayésiens, leur but est d'estimer récursivement la densité de probabilité *a posteriori* $p(x_k|z_{1:k})$ du vecteur d'état x_k à l'instant k conditionné sur l'ensemble des mesures $z_{1:k} = z_1, \ldots, z_k$, une connaissance *a priori* de la distribution du vecteur d'état initial x_0 pouvant être également prise en compte. À chaque instant k, la densité $p(x_k|z_{1:k})$ est approximée au moyen de la distribution ponctuelle $p(x_k|z_{1:k})$ donnée par

$$p(x_k|z_{1:k}) \simeq \sum_{i=1}^{N} w_k^{(i)} \delta(x_k - x_k^{(i)}) \tag{2.1}$$

[2]Pour *Conditional Density Propagation*.

exprimant la sélection d'une valeur – « particule » – $x_k^{(i)}$ avec la probabilité – ou « poids » – $w_k^{(i)}$, $i = 1, \ldots, N$. L'algorithme générique de filtrage particulaire, nommé SIR pour *Sampling Importance Resampling* est résumé dans la communication [4] jointe en fin de mémoire. Les particules $x_k^{(i)}$ évoluent stochastiquement dans le temps. À chaque instant k, disposant de la mesure z_k et de la description particulaire $\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}$ de $p(x_{k-1}|z_{1:k-1})$, la détermination de l'ensemble de particules pondérées $\{x_k^{(i)}, w_k^{(i)}\}$ associé à la densité *a posteriori* $p(x_k|z_{1:k})$ procède en deux étapes. Dans un premier temps, les $x_k^{(i)}$ sont échantillonnés selon la fonction d'importance $q(x_k|x_{k-1}, z_k)$ évaluée en $x_{k-1} = x_{k-1}^{(i)}$. Les poids $w_k^{(i)}$ sont ensuite mis à jour de façon à assurer la cohérence de l'approximation (2.1). Le calcul des poids prend en compte la dynamique $p(x_k|x_{k-1})$ du processus d'état sous-jacent et la vraisemblance $p(z_k|x_k)$ d'un état possible x_k vis à vis de la mesure z_k. Cette vraisemblance est évaluée à partir de la densité de probabilité relative au lien état-observation. Afin de limiter le phénomène de dégénérescence[3], on trouve classiquement un ré-échantillonnage à la fin de chaque cycle.

Nombreuses stratégies de filtres particulaires peuvent être considérées comme des instances de cet algorithme. La CONDENSATION [Isard 1998a] est un cas particulier de l'algorithme SIR puisque la fonction d'importance est relative uniquement à la dynamique du processus d'état ce qui lui confère une structure prédiction / mise à jour comparable à celle du filtre de Kalman. Pour éviter ce positionnement « en aveugle » des particules *i.e.* indépendamment de l'image courante, diverses stratégies basent la fonction d'importance sur la mesure z_k, citons ici la stratégie ICONDENSATION [Isard 1998b] notée encore MSIR pour *Mesure SIR*. Classiquement, une proportion des particules demeure toutefois échantillonnée suivant la dynamique du processus d'état et/ou selon la densité initiale $p(x_0)$.

Les filtres MSIR souffrent d'un problème majeur lié à la définition de la fonction d'importance. En effet, positionner tout ou partie des particules seulement sur la base de l'observation peut conduire à une incompatibilité de ces particules avec leurs particules parentes du point de vue de la dynamique du système. Une alternative intéressante proposée dans [Torma 2003] permet de résoudre cette incohérence en mettant en œuvre des mécanismes basés comme précédemment sur la combinaison du partitionnement de l'espace d'état avec des ré-échantillonnages[4]. Cette dernière admet une structure de type « filtres à particules auxiliaires » comparable à la stratégie *Auxiliary*, introduite dans [Pitt 1999] et notée APF (pour *Auxiliary Particle filter*)

[3] Dégénerescence - selon laquelle après quelques instants, quelques particules du nuage concentrent tous les poids significatifs.

[4] Stratégie notée RBSSHSSIR pour *Rao-Blackwellised History Sampling SIR*.

par la suite. L'APF utilise également des poids auxiliaires pour sélectionner les particules les plus vraisemblables avant leur propagation et donc mieux orienter (*a priori*) les particules vers les zones pertinentes de l'espace d'état. Ces vraisemblances auxiliaires sont possiblement définies sur des mesures visuelles différentes de celles entrant en jeu dans le calcul des poids définitifs.

Comme indiqué précédemment, l'insertion d'un ré-échantillonnage dans le filtre permet la redistribution d'un nuage de particules guidée par une fonction ou un vecteur de poids afin d'obtenir une représentation plus fidèle de la loi *a posteriori*. Ce ré-échantillonnage peut alors être utilisé à différents niveaux du filtre, en particulier lorsque le vecteur d'état peut se séparer en sous-parties échantillonnées successivement. Citons ici les stratégies d'échantillonnage hiérarchisé [Pérez 2004] et partitionné [MacCormick 2000] notée PARTITIONNE par la suite. Cette dernière inclut un ré-échantillonnage dit « pondéré » entre chaque étape de simulation des diverses composantes du vecteur d'état, et ainsi repositionner les particules selon une fonction qui rend compte de la vraisemblance de la partition courante de l'état vis à vis de l'observation. Ces stratégies sont détaillées dans [39].

Les mesures jouent un rôle essentiel dans le fonctionnement des filtres, d'une part dans la définition d'une fonction de vraisemblance des particules, et d'autre part dans la définition d'une fonction d'importance qui détermine la stratégie d'exploration de l'espace d'état. Elles sont extraites des images acquises dans le flot vidéo pour un suivi 2D [TH7] mais également 3D si l'on considère une approche dite *appearance-based* [TH6] ; nous y reviendrons ultérieurement. L'information contenue dans une image étant très riche, des attributs image de natures diverses, typiquement de forme, couleur et mouvement, peuvent être considérés. Comme Pérez *et al.* dans [Pérez 2004], nous les classifierons en attributs persistants ou intermittents selon le contexte applicatif. Les attributs persistants permettent d'obtenir une mesure systématique mais souvent peu discriminante, par exemple un attribut de forme dans un environnement très encombré. *A contrario*, les attributs intermittents dans le flot vidéo, souvent issus de modules de détection, sont par nature discriminants. En marge de ces travaux, il nous semble intéressant d'intégrer plus largement les mesures afin d'augmenter le pouvoir discriminant des fonctions d'importance et de mesure. Ainsi, la fonction d'importance pourra considérer des mesures persistantes tandis que la combinaison/fusion d'attributs dans la fonction de mesure semble pertinente, notamment pour les systèmes adaptatifs qui, par définition, autorisent l'évolution des paramètres de tout ou partie des modèles de mesure. Concernant la fonction de mesure, la fusion de plusieurs attributs par multiplication des vraisemblances associées est assez intuitive et immédiate dès lors que celles-ci sont conditionnellement

indépendantes étant donné l'état. On parle alors de couplage faible. La stratégie de combinaison consiste à calculer dans la fonction de mesure une seule vraisemblance multi-attributs qui sont alors couplés fortement.

Les fonctions d'importance et de mesure implémentées sont basées sur le mouvement, la couleur et/ou la forme. L'attribut **mouvement**, par sa nature intermittente dans notre contexte, est plutôt dédié aux fonctions d'importance, éventuellement aux fonctions de mesure dans une stratégie multi-attributs. Le mouvement image est caractérisé à l'aide du flot optique ou différence absolue entre images successives du flot vidéo.

La **couleur** est un attribut persistant et souvent caractéristique de la cible observée. Nous avons ainsi proposé dans [22] une méthode de segmentation non supervisée des régions peau correspondant aux membres corporels par un algorithme de ligne de partage des eaux. La segmentation en régions caractéristiques permet l'élaboration de fonctions d'importance sur la couleur. La signature colorimétrique d'une, ou plusieurs, région(s) d'intérêt (« ROIs ») relative(s) à la cible est plutôt dédiée aux fonctions de mesure.

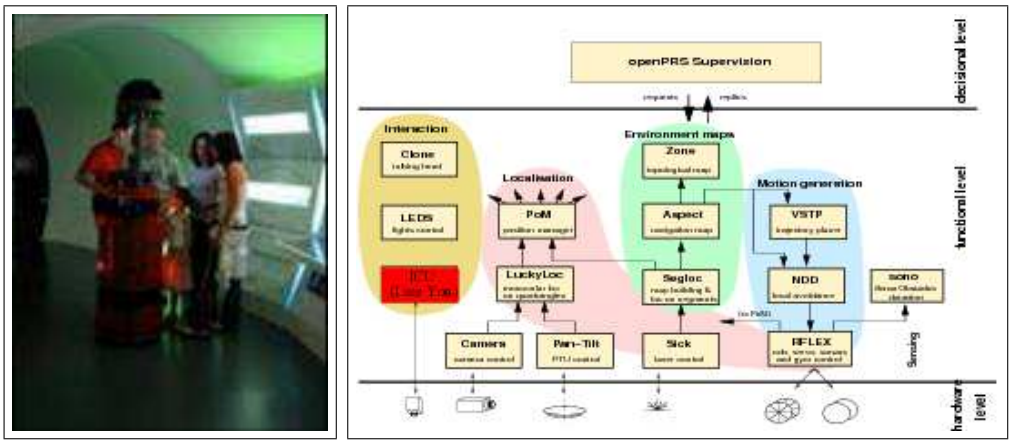
La **forme** est un attribut image persistant lié à la forme caractéristique des membres corporels suivis. Les approches 2D exploitent des modèles de silhouettes rigides ou déformables. La fonction de mesure est classiquement définie par la distance relative du modèle 2D aux contours image. Le principe est transposable à la sous-classe des approches 3D dite *appearance-based où* le modèle 2D résulte de la projection image du modèle 3D de la cible [40]. L'utilisation du seul attribut forme dans la fonction de mesure n'est pas assez discriminant pour des environnements encombrés et justifie l'association de plusieurs attributs. Différents détecteurs permettent de caractériser des fonctions d'importance relative à la forme. Citons quelques exemples : le détecteur multi-échelle de régions circulaires [Bretzner 2002] qui repose sur des invariants différentiels normalisés, le détecteur multi-échelle de visages [Viola 2001] qui exploite des masques de Haar afin de mesurer des contrastes locaux. Une fonction d'importance unifiée consiste en une mixture de mixtures de Gaussiennes centrées sur les différentes régions détectées. Le lecteur trouvera, pour illustration, un sous-ensemble de fonctions d'importance et de mesure fusionnant/combinant ces attributs dans la communication [4] jointe en fin de mémoire.

Les stratégies de filtrage et mesures visuelles implémentées ont permis la spécification puis le prototypage de *trackers* 2D [TH7] ou 3D [TH6], enfin leur intégration puis leur évaluation sur nos plate-formes. La phase de spécification est logiquement guidée par les tâches ou actions robotiques sous-jacentes, les distances H/R mises en jeu et le contexte environnemental propres à leurs réalisations. L'intégration des *trackers* prototypés s'effectue sur deux plate-

formes : (i) le robot guide Rackham pour les fonctions 2D uniquement... ses tâches ou actions en présence de l'homme requérant moins de précision, (ii) le robot manipulateur Jido pour l'ensemble des fonctions 2D et 3D. Détaillons ces différentes fonctions ainsi que leur intégration et évaluation.

3.3 Fonctions 2D

Les fonctions 2D étaient initialement dévolues au robot assistant Rackham, reconverti en « guide de musée », dans le cadre du projet RobEA HR+ [PR5] entre 2001 et 2004. Ce projet, à travers les collaborations initiées avec l'Institut de la Communication Parlée (INPG Grenoble) et l'équipe GRAVIR/IMAG de Grenoble, le savoir faire du pôle RIA en robotique autonome s'est concrétisé par le déploiement de Rackham à la Cité de l'Espace de Toulouse dans le cadre de l'exposition « Mission Biospace ». Ce robot effectuait des séjours de deux semaines tous les trois mois de mai 2004 à février 2005, séjours pendant lesquels il était livré au public sans médiateur ; tout visiteur pouvait ainsi s'adresser à lui pour un renseignement ou être guidé vers le stand de son choix par interaction avec l'écran tactile de Rackham (figure 2.8-(b) haut gauche). La figure 2.8-(b) montre l'architecture logicielle



(a) (b)

FIG. 2.8 – Rackham à la Cité de l'Espace (a), architecture de Rackham et son interface tactile (b).

actuelle de Rackham. Elle exhibe notamment un module Clone dédié à la synthèse vocale et l'animation d'un clone, un module NDD pour l'évitement d'obstacle par laser SICK 2D, enfin notre module de perception de l'homme, appelé ICU (acronyme phonétique de *I see you*). Les séjours à la Cité de l'Espace ont permis de tester intensivement nos fonctions visuelles de détec-

tion de visages (voir ci-après), l'intégration des fonctions de suivi étant hélas prématurée. Elles ont été intégrées plus tardivement et évaluées dans un lieu public type laboratoire.

Néanmoins, la mise en situation de Rackham dans cette exposition, l'observation du comportement des visiteurs au voisinage du robot nous ont guidés dans la déclinaison de ces modalités de suivi visuel permettant au robot d'interpeller, d'interagir *via* ses périphériques, enfin de guider les visiteurs dans le lieu public (figure 2.9). Trois modalités permettent, à nos yeux, de couvrir la majorité des situations H/R rencontrées par notre robot-guide. La modalité $n°$ 1 concerne le suivi de personnes en mouvement dans les plans larges fixes de lieux de passages afin d'interpeller les visiteurs et permettre la mise en action à distance du robot. La modalité $n°$ 2 est relative au suivi de tête dans les gros plans fixes pour l'interaction proximale *via* ses périphériques. La fonction visuelle pour la modalité $n°$ 3 porte sur le suivi de tête et torse dans les plans moyens en mouvement pour permettre au robot-guide de rester en contact visuel avec la personne guidée. La modalité $n°$ 4, plus anecdotique, concerne la reconnaissance de gestes afin de commander à distance le robot alors à l'arrêt ou en mouvement.

(a) (b) (c) (d)

FIG. 2.9 – Situations H/R pour les quatre modalités (a)-(d).

Le vecteur d'état de nos filtres contient les composantes caractérisant la configuration dans le plan image du modèle de la cible. Celles-ci sont supposées évoluer indépendamment suivant des marches aléatoires gaussiennes centrées sur les valeurs estimées de l'état à l'instant précédent. L'initialisation, voire la ré-initialisation (après « décrochage »), des filtres mis en œuvre doivent être automatiques.

Pour le suivi de personnes (modalités $n°$ 1 à 3), diverses stratégies de filtrage impliquant divers attributs visuels sont proposées puis évaluées sur des séquences-types [4]. Celles-ci sont représentatives des situations et artefacts rencontrés par le robot dans chacune de ces modalités : scènes encombrées, changements brusques d'apparence ou de dynamique de la cible, présence de

plusieurs individus, occultations, etc.

Pour chaque modalité, le comportement qualitatif des filtres est illustré dans [4] par des réalisations de suivi sur des séquences mettant en jeu les situations et artefacts mentionnés. Les figures suivantes montrent, pour les modalités $n°$ 1 et $n°$ 3, deux réalisations sur des séquences incluant des occultations significatives. La particule en rouge est la moyenne *a posteriori* de toutes les particules (en bleues).

FIG. 2.10 – Exemple de réalisation sur une séquence incluant un groupe de personnes pour le filtre relatif à la modalité $n°$ 1.

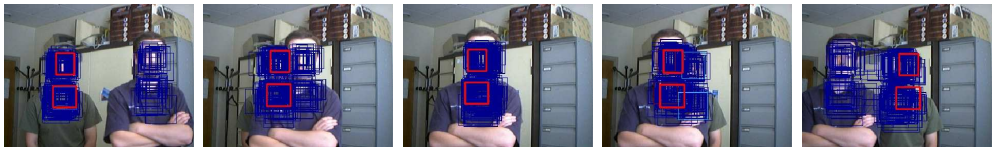


FIG. 2.11 – Exemple de réalisation en présence d'occultations de personnes pour le filtre relatif à la modalité $n°$ 3.

Des évaluations chiffrées sur des jeux de séquences-type sont également proposées et commentées dans [4]. Ces évaluations portent sur trois critères : précision, taux d'échec et temps de traitement. Par exemple, la figure 2.12 illustre, pour la modalité $n°$ 1, les temps de traitement et taux d'échec moyens obtenus à partir de réalisations sur des séquences avec occultations.

Globalement, les fréquences sont de 20 Hz à 50 Hz, sur un Pentium IV 3GHz, pour le nombre de particules couramment utilisé (100 à 200). L'analyse de l'ensemble des évaluations ont permis de caractériser les associations stratégies de filtrage/mesures les plus pertinentes pour chacune des modalités de suivi de personnes [39].

La modalité $n°$ 4 porte sur la reconnaissance de gestes élémentaires. Le geste, comme la parole, apparaît pour les humains comme un moyen spontané de communication. Il est donc légitime d'intégrer des capacités de reconnaissance gestuelle sur nos robots. Une première contribution [TH7] porte sur la reconnaissance de gestes symboliques. Considérant uniquement la main, l'information transmise est contenue dans la configuration de la main et/ou le geste supposé(s) ici fronto-parallèle(s) à la caméra. L'ensemble est modélisé

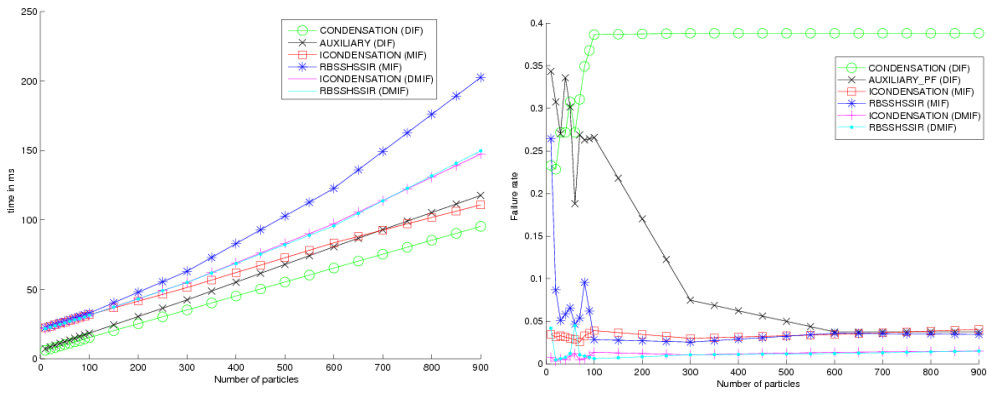


FIG. 2.12 – Temps de traitement et taux d'échec *vs.* nombre de particules.

par un système à sauts Markoviens *où* un geste consiste en l'enchaînement de gestes élémentaires « canoniques », chacun d'eux étant indicé par l'un des couples (configuration, dynamique) constituant une « bibliothèque » définie *a priori*.

Comme [Bretzner 2002,Liu 2004], l'approche retenue ne dissocie plus aussi distinctement les deux phases d'analyse et de reconnaissance car l'estimation et la reconnaissance s'effectuent simultanément. La hiérarchie entre les composantes discrètes et continues du vecteur d'état d'un système à sauts Markoviens permet la définition d'algorithmes de filtrage particulaire simplifiés, où les parties discrètes et continues des particules peuvent être échantillonnées successivement. Citons ici la stratégie *mixed-state* CONDENSATION, introduite par Isard *et al.* [Isard 1998c]. Les composantes continues du vecteur d'état sont relatives à la configuration image de la cible tandis que les composantes discrètes indexent des configurations de la main (représentées par leurs silhouettes) et des trajectoires canoniques inspirées de l'écriture Graffiti développée par Palm Pilot. L'objectif à terme est d'estimer dans le filtre la trajectoire la plus vraisemblable afin de reconnaître les lettres de l'alphabet modélisées par un enchaînement de modèles de dynamiques canoniques. Les configurations de la main permettent, en particulier, de segmenter le noyau, la préparation et la rétraction du geste analysé. Ces travaux préliminaires sur la reconnaissance de gestes fronto-parallèles ont contribué aux publications [15, 21]. Des vidéos de reconnaissance de configurations et/ou de modèles de dynamiques sont accessibles à l'URL `www.laas.fr/~lerasle`. Les dernières évaluations, détaillées dans la communication [15] aboutissent à des taux de reconnaissance avoisinant les 90% pour sept configurations de la main et cinq modèles de dynamiques.

Ces travaux passés et leurs prolongements actuels s'inscrivent aujourd'hui

dans le projet COGNIRON [PR4] qui court jusqu'en 2008. Nos investigations sont discutées, sur un plan formel, dans un groupe de travail traitant de la perception de l'homme. Sur un plan expérimental, ces efforts se focalisent sur le robot personnel Jido et un scénario illustrant le déplacement en présence de l'homme et la manipulation coordonnée d'objets. Le robot Jido intègre aujourd'hui les modalités n° 1 et n° 3 dans leur version ré-actualisée. Nous avons également intégré, pour les besoins du scénario et à des fins de comparaison, le module VooDoo [Knoop 2006] développé initialement par l'Université de Karlsruhe, et relatif au suivi 3D des membres corporels à partir d'une caméra *time-of-flight*.

3.4 Fonctions 3D

Détaillons nos propres développements sur le suivi 3D du corps humain. Ils ont démarré avec la thèse de P.Menezes [TH6] dont les séjours longue durée au LAAS-CNRS s'inscrivent dans une collaboration avec l'Institut des Systèmes Robotiques (ISR) de l'Université de Coimbra [CO1] qui n'est pas partenaire du projet COGNIRON. Ces travaux sont actuellement poursuivis dans le cadre de la thèse de M.Fontmarty [TH5] ; nous y reviendrons au chapitre suivant.

La capture du mouvement humain depuis une caméra monoculaire ou stéréoscopique embarquée sur une plate-forme mobile autonome constitue une problématique complexe et ouverte dans la communauté Robotique... même si la communauté Vision répertorie de nombreux travaux sur cette problématique. À l'instar d'un grand nombre d'entre eux, l'approche retenue est dite *model-based* au sens *où* elle s'appuie sur une représentation 3D frustre de l'enveloppe corporelle tirant partie de considérations biomécaniques et anthropomorphes. Les travaux sur le suivi *model-based* du corps humain sont pléthores ; on distingue classiquement deux classes d'approches :

- par reconstruction 3D *où* le but est d'estimer la configuration spatiale du modèle à partir d'appariements 3D/3D [Delamarre 2001, Urtasum 2004] et d'un système multi-caméras,
- par apparence *où* le but est d'inférer la configuration 3D qui correspond « au mieux » à l'apparence du modèle après projection dans le plan image de chaque caméra [Deutscher 2001, Sidenbladh 2000].

Même si l'apparence du corps humain manque de texture naturelle, il me semble pertinent de mener l'analyse dans le plan image... qui reste par définition très informatif. Nous privilégions donc dans [TH6] une approche *appearance-based* légère et simple en termes à la fois de modélisation et de mise en œuvre (une seule caméra étalonnée).

Ainsi, notre modèle est composé d'un ensemble de quadriques tronquées rigides et articulées représentant les membres corporels. Chaque articulation est caractérisée par un ou plusieurs degrés de liberté modélisant la cinématique de la structure. Les quadriques sont des primitives géométriques assez simples à manipuler, par opposition aux surfaces maillées utilisées par exemple durant ma thèse. De plus, ces primitives anthropomorphes permettent de s'affranchir de tout protocole lourd et contraignant de modélisation. Enfin, la géométrie projective en permet de façon élégante la projection image et la gestion des occultations [Stenger 2001].

La littérature propose alors de nombreux algorithmes pour la gestion des parties cachées après projection. Leur complexité reste liée à la taille des parties projetées et/ou la précision requise pour leur projection [Deutscher 2001, Stenger 2001]. Nous avons implémenté un algorithme dit « de la droite de balayage » ou *sweeping line* dont la complexité dépend uniquement du nombre de quadriques mises en jeu. Le principe est de déduire la visibilité d'un segment projeté à partir du seul test sur son point milieu. Cet algorithme de projection du modèle et gestion de ses parties cachées est largement détaillé dans [20].

Nous privilégions une technique de filtrage particulaire, ici la stratégie APF qui introduit un ré-échantillonnage auxiliaire préalablement à la « propagation ». Chaque particule se voit attribuer une vraisemblance auxiliaire liée à la distribution de couleur peau observée dans les régions correspondant aux projections des extrémités des avants-bras. Les mains sont ici des marqueurs naturels (de par leur couleur caractéristique) mais virtuels car elles ne sont pas représentées dans le modèle 3D. Le calcul des poids définitifs repose sur la fusion des attributs forme, couleur, éventuellement mouvement dans la fonction de mesure associée. Comme [Sminchisescu 2003], cette dernière intègre également des considérations de non collision entre parties et de configurations par défaut lorsque des régions de l'espace des configurations sont ponctuellement non observables.

Une contrainte, relaxée dans [TH5], est la connaissance *a priori* des paramètres de position globale supposés figés et estimés en amont par une fonction de suivi dédiée. Les composantes du vecteur d'état sont donc relatives aux seuls 8 ddls des bras tandis que les expérimentations portent sur des séquences pré-enregistrées dans des environnements quelconques. Les fréquences de traitement sont alors de 1Hz pour 400 particules sur un Pentium IV 3GHz. Ces travaux sur la capture du mouvement ont fait l'objet de publications, citons [18, 17].

La figure 2.13 montre le séquencement de nos modules vision intégrés dans l'architecture du robot compagnon Jido. Ces modules notés TBP, ICU,

HumRec et GEST sont relatifs respectivement à la capture du mouvement, suivi 2D de personnes, reconnaissance de visages et suivi 3D de gestes (c.f. chapitre suivant). Les commutations entre ces modalités s'appuient sur les informations délivrées par les couches fonctionnelle et décisionnelle du robot.

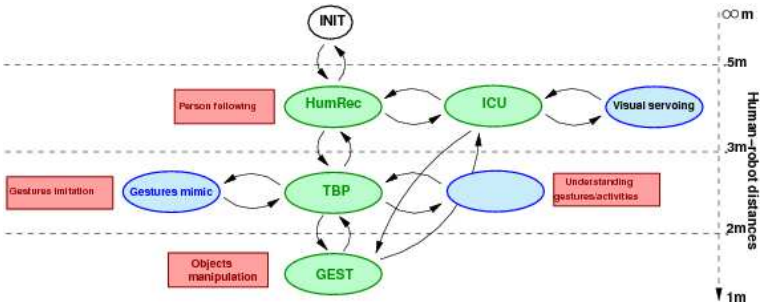


FIG. 2.13 – Séquencement des modules de vision pour le robot personnel Jido.

Une application récente de ces travaux portent sur l'imitation de gestes par un humanoïde à partir de notre système visuel de capture de mouvements. L'animation du robot est réalisée grâce à la plate-forme KineoWorks^{TM} et au modèle HRP2^{TM}. Ces simulations sont détaillées dans [19].

3.5 Contributions

Pour résumer, les contributions passées sur ce thème sont donc les suivantes :

- La caractérisation de fonctions multi-attributs entrant en jeu dans le ré-échantillonnage ou la pondération des particules. La fusion ou combinaison systématique de données sensorielles homogènes dans nos filtres est conforme à notre ligne de conduite et répond au constat suivant : les possibilités d'intégration de données offertes par le cadre général du filtrage particulaire sont assez peu exploitées et souvent confinées à un nombre restreint de ces données.
- Le prototypage de fonctions de suivi de personnes par le choix conjoint de primitives visuelles et de stratégies de filtrage (différentes par leurs structures et leurs étapes d'échantillonnage et ré-échantillonnage) répondant au mieux aux modalités d'interaction envisagés pour le robot assistant Rackham. Ces travaux offrent à ce titre une bonne synthèse des techniques de filtrage particulaire et de leurs performances... alors que la littérature propose des évaluations limitées en termes de com-

paraisons entre stratégies ainsi qu'en nombre et variété des séquences traitées.
- La reconnaissance de gestes fronto-parallèles « statiques » et dynamiques par une stratégie de *mixed-state* CONDENSATION qui permet leur reconnaissance dans la boucle même de suivi. Cette contribution reste à ce jour marginale dans nos travaux.
- La capture de mouvement 3D. La fusion d'attributs permet un suivi robuste aux conditions environnementales malgré l'utilisation d'une seule caméra et d'un modèle anthropomorphe frustre. De part ces choix, les temps de traitement restent compatibles avec le contexte applicatif. La gestion relativement originale des parties cachées du modèle 3D après projection est à mentionner ici. Ces fonctions sont dévolues à terme au robot personnel Jido.
- l'intégration des fonctions visuelles, leur exécution, séquencée par le superviseur du robot Jido ou Rackham, dans le cadre de scénarios robotiques réalistes en souligne leur pertinence et leur complémentarité. À ma connaissance, peu de robots mobiles interactifs intègrent aujourd'hui des capacités visuelles de l'homme aussi avancées.

4 Retombées connexes

Nous avons participé à un projet applicatif PREDIT [CI3] portant sur la perception de l'homme dans le domaine automobile, domaine connexe à la Robotique. Ce projet, impliquant également Siemens VDO et l'ONERA-CERT, consistait à développer un système de supervision optique de l'habitacle d'un véhicule pour détecter la présence du passager, caractériser son éventuelle posture afin de mieux contrôler le déclenchement des systèmes de coussins gonflables ou *airbag*. La société Siemens VDO avait en charge l'intégration et la mise en œuvre de prototype(s), le CERT la conception des capteurs optiques, enfin le LAAS-CNRS (à travers M.Devy et moi-même en chercheurs permanents) le développement d'algorithmes de vision dédiés. Deux approches/systèmes « concurrents » ont été développés en parallèle.

La première approche repose sur un système de stéréovision passive couplant deux caméras CMOS et un éclairage diffus à diodes IR pour éclairer le siège en conditions de faible luminosité. La démarche est alors de mettre en œuvre et évaluer nos algorithmes de stéréocorrélation dense afin de reconstruire le siège passager. Une classification bayésienne à partir d'un vecteur d'attributs 3D permet enfin d'identifier la situation courante.

La seconde approche repose sur un système de lumière structurée associant une caméra CCD et un illuminateur IR. La démarche fût alors de

proposer un processus spécifique d'étalonnage du système, puis de reconstruire en 3D, mais de façon plus éparse que précédemment, le siège passager. Cette reconstruction repose sur un processus de relaxation sur des contraintes inhérentes au système afin de mettre en correspondance les données capteurs (spots image/faisceaux laser). La classification de la situation courante est alors inférée à partir de la distribution volumique des points 3D ainsi reconstruits.

Pour « répartir nos forces » côté LAAS-CNRS, M.Devy s'est focalisé en priorité sur l'approche n° 1 et moi-même sur l'approche n° 2. J'ai travaillé en collaboration avec J.M.Lequellec ingénieur chez Siemens VDO. Les deux prototypes embarqués sur un véhicule s'avèrent robustes aux fortes variations ambiantes d'éclairement ; les temps de traitement dans un environnement PC sont intéressants (resp. 4 Hz et 30 Hz) et ont permis des retombées industrielles dans le milieu automobile.

Ces travaux ont donné lieu à un brevet [9] et de nombreuses communications en congrès [28, 31, 32, 33, 42]. Le lecteur pourra se référer à la revue [6], jointe en fin de mémoire, revue qui décrit les deux systèmes mis en œuvre et compare leurs performances respectives. Signalons enfin que la collaboration avec Siemens VDO a repris en 2006 à travers un contrat équipe-conseil [CI2] sur une problématique de suivi visuel.

Chapitre 3

Travaux actuels et prospectives

1 Préambule

Ma prospective de recherche s'inscrit dans la continuité et complémentarité des travaux antérieurs et actuels en relation étroite avec le projet de recherche énoncé. Je dissocie ici les travaux en cours et la prospective à moyen terme relative aux investigations actuelles et « programmées » dans les cinq prochaines années, de la prospective à plus long terme visant à une projection à plus longue échéance.

La section 2 porte sur les travaux actuels et la prospective à moyen terme. La prospective à long terme s'inscrit dans une réflexion sur les extensions possibles de mes travaux passés et actuels dans l'esprit du projet de recherche énoncé. Quelles orientations, voire inflexions possibles de ces travaux peuvent nourrir des problématiques robotiques encore ouvertes ? À plus longue échéance, quels challenges scientifiques et technologiques futurs sont susceptibles de nourrir la robotique personnelle, indirectement et plus modestement, mes travaux ? La section 3 énumére quelques voies exploratoires permettant de répondre très partiellement à ces questions. Détaillons ces prospectives à courte, moyenne, et longue échéances.

2 Travaux actuels et prospectives à moyen terme

Les travaux actuels et programmés à moyen terme s'appuient sur la même ligne de conduite : l'intégration de percepts multiples et incertains à tous les maillons de la chaîne perceptuelle. Ils n'induisent pas à proprement parlé d'inflexion majeure, ils visent plutôt, en cohérence avec le projet énoncé, à approfondir et proposer des prolongements aux approches et outils développés.

Ainsi, une motivation majeure du projet est une navigation autonome au très long cours. Nos représentations doivent alors capturer des environnements de grande envergure, exhiber certes des amers plus génériques à travers nos modèles métriques et topologiques mais aussi exhiber des informations sémantiques relatives aux objets et lieux de l'environnement. Ces informations sémantiques sont acquises par interaction avec l'homme *via* la parole et/ou la gestuelle donc l'interprétation de signaux audio et visuel. Au-delà de cet exemple, la perception de l'homme sera résolument multi-sensorielle. L'intégration de données visuelles multiples, donc homogènes, abordée précédemment, sera étendue à des données sensorielles hétérogènes donc issues de capteurs de natures diverses.

Forts de ces généralités, nous listons ci-après nos investigations actuelles et futures sur la perception de l'homme par le robot tout en faisant état de

connexions avec la perception de l'espace sur laquelle nous reviendrons un peu plus tard. Dans l'esprit du projet, ces investigations sur la navigation autonome et donc sur les représentations de l'espace et surtout du milieu sont subordonnées à l'avancement des travaux sur la perception de l'homme par le robot. Ces contributions attendues s'inscrivent dans le cadre de thèses, projets ou contrats, en cours ou programmés (partie II).

2.1 Perception de l'homme étant donné la perception de l'espace

Ces contributions, pour certaines effectives, sont déclinées ci-après en parcourant la chaîne perceptuelle de façon ascendante *i.e.* depuis l'acquisition et le traitement des données sensorielles jusqu'à leur interprétation en passant par le suivi qui fait l'objet de toutes nos attentions.

Intégration de données sensorielles homogènes — L'intégration de données sensorielles homogènes, est à approfondir afin d'améliorer encore la robustesse de nos *trackers* aux nombreux artefacts propres aux environnements humains. Une première extension, prévue dans [TH2] et [TH5], sera de considérer des données 3D denses ou éparses dans nos filtres particulaires tout en continuant d'exploiter des données d'apparence. Les capteurs 3D utilisés seront une tête stéréoscopique sur Jido ou HRP2 pour le suivi 3D de personnes [TH2], une tête stéréoscopique ou une caméra monoculaire+caméra active sur Jido (figure 3.4-(a)), pour la capture de mouvement [TH5]. Nous bénéficierons, au passage, de l'expertise acquise durant le projet PREDIT [PR3] quant aux possibles paramétrisations et performances de nos algorithmes de stéréovision dense pour la reconstruction de « scènes humaines ». Pour une reconstruction éparse, le principe sera par exemple de suivre quelques points 3D par leur disparité relative [Moreno 2002] afin de limiter le coût calculatoire. La stratégie retenue est donc à mi-distance entre : (i) les stratégies dites *appearance-based*, *e.g.* [Deutscher 2001, Sminchisescu 2003, Sigal 2004] permettant d'accéder à toute la richesse de l'information visuelle mais hélas peu pertinentes pour inférer du 3D, et (ii) les stratégies par reconstruction dense, *e.g.* [Delamarre 2001, Knoop 2006, Ziegler 2007], qui procurent des contraintes géométriques fortes mais occultent l'information d'apparence visuelle.

Cet « éventail » de primitives autorisera des stratégies flexibles de fusion des mesures associées. Le but sera de s'adapter aux fortes variabilités environnementales en s'appuyant sur les informations contextuelles et sémantiques associées à la localisation du robot donc à partir de la perception

de l'espace. Par exemple, les informations d'apparence semblent plus adaptées pour percevoir l'homme depuis le robot lorsque celui-ci évolue dans un couloir, espace confiné par excellence. *A contrario*, les mesures géométriques semblent plus indiquées pour la perception dans un espace ouvert. Ainsi, la nature du lieu investi, mais aussi les conditions courantes de prises de vues (illumination, encombrement, etc.), la tâche exécutée par le robot, doivent donner lieu à une pondération adaptée des diverses primitives dans la fonction de mesure unifiée. Cette étude sera réalisée dans le cadre de [TH5]. À ma connaissance, la littérature ne propose pas de *trackers* de l'homme ou de ses membres intégrant aussi largement des primitives visuelles, tandis que la fusion adaptative de ces primitives a été abordée dans [Vermaak 2002] pour des attributs d'apparence uniquement.

Ces attributs d'apparence sont actuellement modélisés par des distributions de couleur, texture ou mouvement, probabilités peau, etc caractérisant plutôt des régions. Dans [TH2], nous les mixerons avec des primitives plus locales type points d'intérêt [Serby 2004] connectés par leur topologie [Gabriel 2005]. Ces primitives sont particulièrement adaptées à la gestion des occultations observées pour des environnements fortement peuplés *a priori*. Cette problématique sera abordée dans le cadre du projet CommRob [PR3] dont la finalité est la conception puis le déploiement de chariots assistants autonomes évoluant dans un lieu public type supermarché ou aéroport. Le projet se focalise sur la communication entre les agents (humains, chariots) partageant l'espace. Dans ces environnements hautement dynamiques, la communication inter-robots devrait faciliter leurs tâches de navigation tandis que la communication H/R aura pour but de guider et renseigner, par le geste ou la parole, le visiteur dans l'environnement. Le principe sera de rester en contact visuel (par suivi) avec le visiteur lors de la mission de guidage tandis que l'exécution du mouvement sera contrôlée par asservissement visuel gérant obstacles et occultations persistantes [Folio 2005].

Reconnaissance faciale et caractérisation du regard — La présence de nombreuses personnes au voisinage immédiat du robot induit des commutations intempestives entre cibles, voire des décrochages de nos filtres durant le suivi. Ces comportements, certes sporadiques, ont été observés durant nos expérimentations antérieures menées au laboratoire. La prise en compte (partielle) de l'apparence vestimentaire était alors la parade pour discriminer les individus. Le visage permet aussi cette discrimination. **La reconnaissance faciale**, largement référencée dans la littérature [Hjelmas 2001], fait actuellement l'objet de développements dans [TH2]. L'approche est supervisée, éventuellement non supervisée [Belhumeur 1997], suivant le contexte appli-

catif. Elle s'appuie sur la définition d'images propres générées par ACP ou AFD[1] et une règle de décision dérivée de la distance à l'espace propre associé à chaque classe notée classiquement DFFS (pour *Distance From Face Space*). La contribution principale fut ici d'intégrer, dans un cadre probabiliste, la reconnaissance de visages dans nos filtres. La démarche est détaillée dans la communication [2]. La figure 3.1 illustre une démonstration très récente sur le robot guide Rackham et incluant la reconnaissance de visages. Une exten-

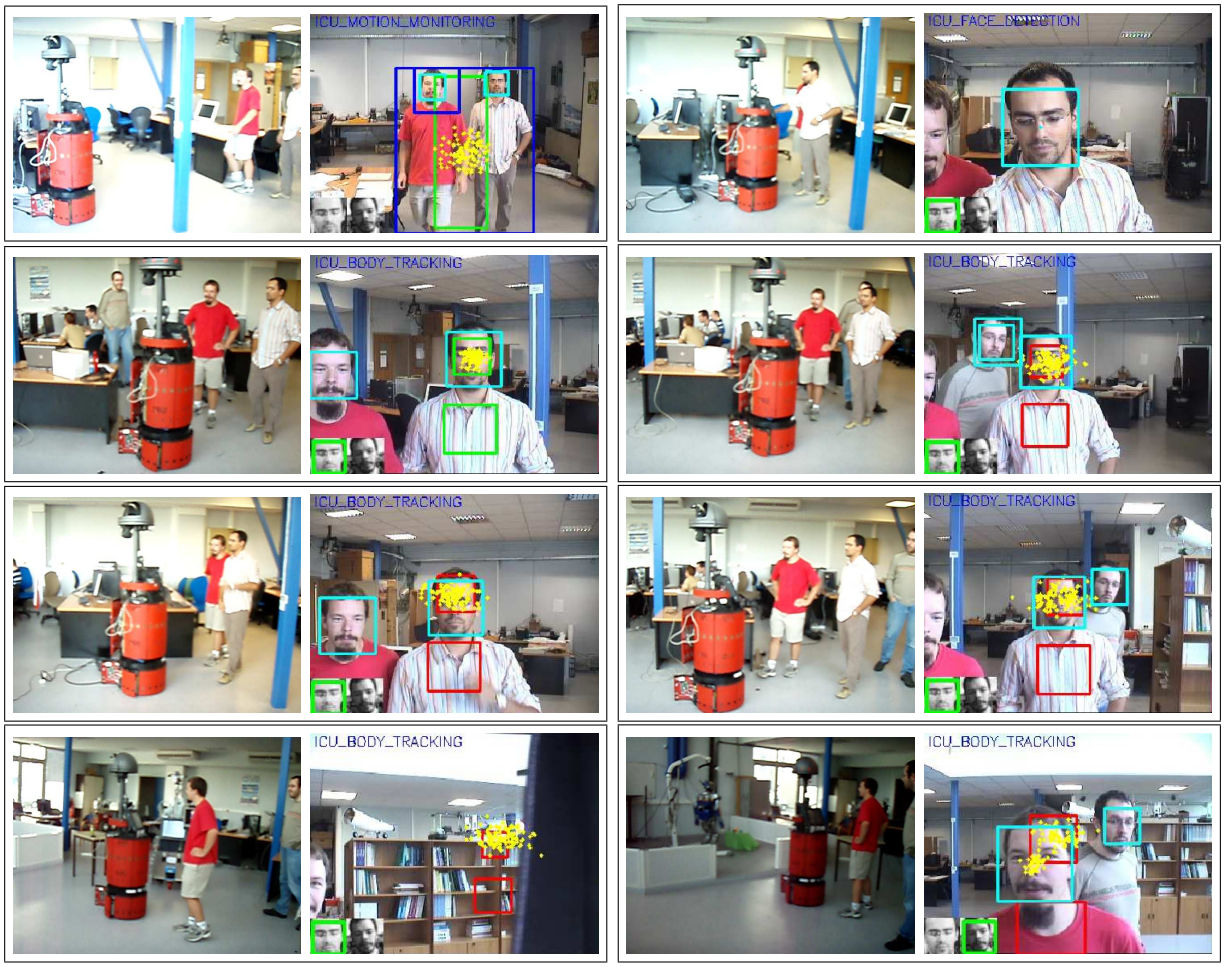


FIG. 3.1 – Séquencement de modalités visuelles pour la mission *go-to-HRP2* avec changement d'interlocuteur en cours de mission : situation courante H/R -gauche-, retour visuel de la caméra embarquée -droite-. Etat courant du suivi avec/sans (en vert/rouge) reconnaissance faciale, particules en jaune, détection faciale des intrus à la mission (en bleu).

[1] Analyse Factorielle Discriminante.

sion attendue sera à terme de **caractériser également la directivité du regard** afin de vérifier l'intentionalité de l'homme durant l'interaction. Le but n'est nullement de suivre le regard, cette problématique étant largement traitée dans la littérature, notamment dans [Fletcher 2005]. L'approche, inspirée de [Li 2006], sera d'entraîner des classifieurs hiérarchiques de visages pour différents sujets et orientations à partir de techniques dites de *boosting*. Il s'agira alors, comme précédemment, d'intégrer les probabilités de classification dans nos *trackers*. Cette stratégie, contrairement au suivi de regard, permettra des détections non nécessairement faciales et donc augmentera les capacités de ré-initialisation automatique des filtres après « décrochages » de la cible. Mentionnons enfin qu'une étude complémentaire sur la classification de visages, par des approches type SVM (*Support Vector Machine*) ou neuronale, est menée à l'IUT Figeac, dans le cadre d'échanges scientifiques avec T.Simon (chercheur associé au LAAS-CNRS, MCF IUT Figeac).

Intégration de données sensorielles hétérogènes — Le contexte applicatif de CommRob nous amène à aller au-delà et investiguer plus largement la reconnaissance de personnes à travers l'intégration de données sensorielles hétérogènes. Ainsi, les travaux menés dans [TH2] étudieront le couplage de signaux radio-fréquences, sonore et vidéo pour la détection, la reconnaissance et le suivi de personnes instrumentées d'étiquettes RFID (figure 3.4-(b)). La caractérisation de nombreux détecteurs à partir de ces sources multiples permettront notamment une ré-initialisation automatique de nos *trackers* sur un éventail de situations H/R plus diverses et variées que par le passé.

Concernant le son, nos travaux seront couplés avec ceux menés dans le groupe par P.Danès et ses doctorants (S.Argentieri, J.Bonal) sur le développement d'un capteur acoustique actif pour la détection et la focalisation sur une source sonore [Argentieri 2005]. L'utilisation de RFIDs (actifs ou passifs), pour l'identification de personnes dans les lieux publics, est en plein essor [Hähnel 2004,Kanda 2007, Matthieu 2005]. Cette technologie est plutôt récente, son couplage avec la vision reste, à ce jour, marginale et évoqué dans quelques travaux exploratoires ; citons [Matthieu 2005] dans un contexte de surveillance à partir de caméras d'ambiance. La programmation de RFIDs, l'intégration de ces signaux radio-fréquence avec des signaux sonores et visuels seront réalisées par T.Germa [TH2] dans le cadre du projet CommRob.

Estimation stochastique — Les investigations sur l'estimation stochastique seront surtout menées par M.Fontmarty dans [TH5], en coopération avec P.Danès co-encadrant de cette thèse. La capture du mouvement humain, donc de la position et des différents ddls de la structure articulaire

humaine, requiert, plus que les autres *trackers*, des stratégies plus élaborées de filtrage particulaire afin de limiter le coût relatif à l'exploration de l'espace des configurations associées. Ce coût croît exponentiellement en fonction du nombre de ddls pour une stratégie CONDENSATION et linéairement en fonction du nombre de partitions pour une stratégie PARTITIONNE. Des évaluations préliminaires, détaillées dans [37], nous ont permis de quantifier l'apport d'une stratégie PARTITIONNE dans ce cadre applicatif.

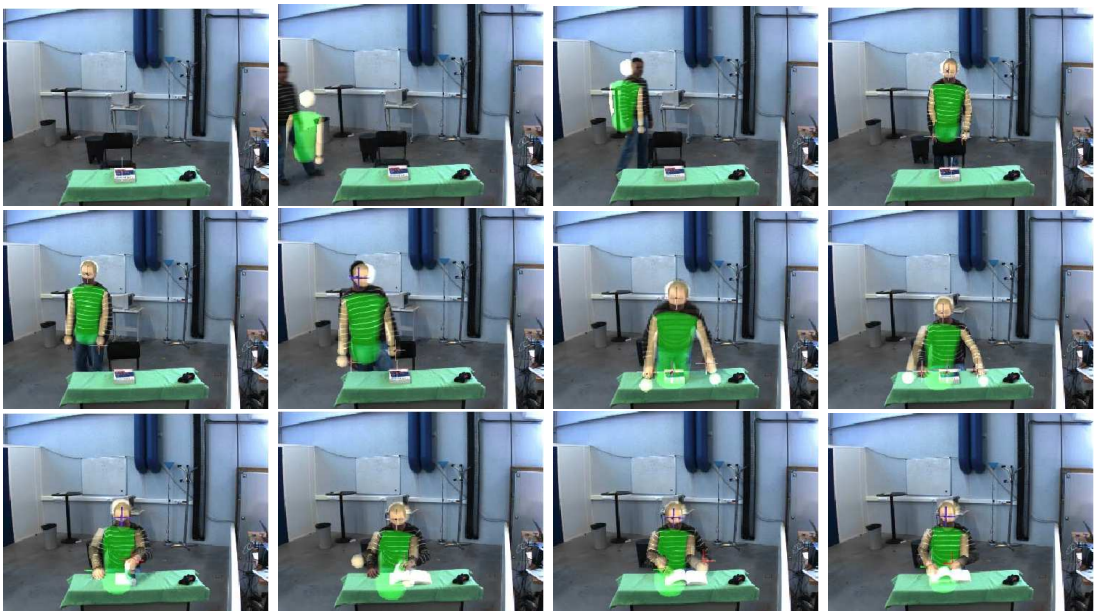


FIG. 3.2 – Capture de mouvement par vision stéréoscopique durant une activité de lecture. Initialisation puis occultations du sujet durant ses mouvements.

Plus récemment, nous avons proposé une stratégie de filtrage combinant les avantages des stratégies ICONDENSATION et hiérarchique par recuit simulé. Cette nouvelle stratégie notée *I-Annealed Particle filter* permet une (ré)-initialisation automatique du filtre, comme illustrée sur la figure 3.2, et une réduction du nombre de particules [14]. Nous espérons réduire, de façon drastique (autour de $\sim 30\%$), leur nombre par des techniques de Quasi Monte-Carlo ou QMC [Guo 2006] qui positionnent les particules grâce à des générateurs de nombres quasi-aléatoires type Sobol ou Halton. Peu de travaux font état de stratégies QMC pour le suivi visuel ; citons [Philomin 2000] pour le suivi 2D... rien, à ma connaissance, pour le suivi 3D.

Dans la veine des travaux passés, la stratégie de filtrage et de fusion sensorielle, offrant le meilleur compromis (robustesse, temps de calcul) sera intégrée sur le robot personnel Jido. La capture de mouvement humain depuis

un robot mobile reste encore très marginale dans la communauté Robotique, le coût calculatoire étant probablement l'obstacle majeur pour cette fonction perceptuelle qui serait d'une utilité majeure pour tout robot interactif. Les développements formels associés à [TH5] s'inscrivent dans le *Research Area* n° 2 de COGNIRON [PR4] tandis que l'intégration sur Jido d'un premier prototype, la comparaison avec le module actuel VooDoo [Knoop 2006], leur éventuel couplage, s'inscrivent dans le *Key-experiment* n° 2.

Signalons enfin que l'achat programmé par le LAAS-CNRS d'un système de **capture de mouvement type VICON** nous intéresse à double titre : (1) pour constituer une vérité terrain et ainsi évaluer la précision des prototypes possibles de capture de mouvement, (2) pour modéliser les corrélations entre ddls de la structure articulaire et ainsi définir une paramétrisation plus compacte de la chaîne cinématique [Urtasum 2004,Wu 2001]. Ce dernier point sera traité en synergie avec le groupe GEPETTO qui traite notamment de la modélisation du langage corporel [Suleiman 2006].

Suivi multi-cibles — Nos travaux sur le suivi mono-cible doivent être étendus à terme au suivi multi-cibles, typiquement pour la navigation réactive[2] ou la manipulation conjointe d'objets. Deux stratégies, distribution de filtres [Qu 2007, Yu 2004] *versus* caractérisation d'un unique filtre [Isard 2001, Zhao 2004], sont proposées dans la littérature. Ces deux stratégies sont à l'étude dans nos travaux dans deux contextes bien distincts.

La thèse de B.Burger porte sur la fusion audio-visuelle pour l'interaction H/R, en particulier pour l'interprétation de commandes à partir de la parole et de gestes symboliques ou déictiques. La reconnaissance de ces gestes spécifiques peut ici s'affranchir d'une capture exhaustive et contraignante de tous les ddls traitée dans [TH5]. Ainsi, dans [TH3], nous travaillons actuellement sur le suivi 3D par vision binoculaire des organes terminaux des membres supérieurs corporels à partir de trois filtres dédiés. Ces **filtres sont distribués et interactifs** lorsque les cibles associées sont proches afin d'éviter toute ambiguïté dans l'association des données visuelles entre filtres. Ce mécanisme interactif, initié sur du suivi 2D multi-personnes [Qu 2007] a été étendu ici au suivi 3D de gestes à partir d'ellipsoïdes déformables modélisant les configurations spatiales des mains. La figure 3.3 illustre ces travaux en cours et leur couplage avec l'interprétation de la parole. Le lecteur pourra se référer aux communications [13, 12], cette dernière étant jointe en annexe.

Cette stratégie montre ses limites pour un nombre supérieur de cibles en interaction. L'alternative est alors **d'instruire un seul et unique vecteur**

[2]Nécessitant d'estimer les déplacements de tous les humains proches afin de planifier localement un chemin adapté.

FIG. 3.3 – Reconnaissance de commandes multimodales à partir de gestes et parole. Suivi 3D par stéréovision embarquée à partir de filtres distribués sur des ellipsoïdes déformables représentant les extrémités corporelles (images bas gauches).

d'état à estimer à partir de l'état relatif à chaque cible, la dimension de ce vecteur étant donc possiblement variable dans le temps. Nous abordons cette formalisation du problème dans la thèse de I.Zuriarrain [TH1] en co-tutelle avec l'Université de Mondragon (correspondant : N.Arana). Le but est : (i) le développement d'un algorithme de surveillance de personnes dans un lieu public à partir de caméras d'ambiance fixes et la caractérisation grossière de leurs attitudes à l'instar de [Boulay 2006], (ii) le câblage de cet algorithme sur FPGA. Ce câblage prend tout son sens dès lors que le nombre de particules croît, en règle générale, exponentiellement en fonction du nombre de cibles. Les travaux les plus proches conceptuellement sont ceux de Uk *et al.* [Uk 2007] mais la parallélisation porte sur deux filtres dédiés au suivi de deux objets. Sur le volet algorithmique, nos premières investigations, dans le cadre d'un stage préparatoire [DEA1], portent sur la fusion probabiliste de cartes de saillance. Ces cartes sont relatives à la détection : (i) de personnes par classifieurs *Adaboost*, (ii) de mouvement par un mélange pondéré et adaptatif de gaussiennes [Stauffer 1999]. L'échantillonnage des particules par un mécanisme de rejet (*rejection sampling*) doit permettre de placer les particules conformément à ces cartes. Cet échantillonnage, original s'il prouve son efficacité, s'inscrit dans une stratégie ICONDENSATION, probablement à amender. Cette thèse constitue la « cheville ouvrière » du projet AMISEG entre l'Université de Mondragon, une *start-up* espagnol et le LAAS-CNRS. Ce projet [PR1] est financé par la communauté de travail pyrénéen entre la région midi-pyrénées et le pays basque espagnol. La finalité du projet est le **prototypage d'une caméra autonome intelligente** *i.e.* disposant de ses propres ressources énergétiques, communiquant *via* un réseau sans fil, enfin intégrant des algorithmes de traitement des images câblés sur FPGA. Ce prototype de capteur (figure 3.4-(c)), par sa facilité de déploiement dans l'environnement, permettra peut-être d'envisager des retombées pour le grand public, par exemple pour la surveillance à domicile de personnes âgées. Notons enfin, que ces travaux, en ouvrant sur l'instrumentation de l'environnement, semblent ici un peu marginaux, ils s'inscrivent néanmoins dans la problématique du réseau de capteurs communicants ébauchée dans notre prospective à long terme.

Fusion visio-auditive pour une interaction H/R multimodale — L'interprétation des gestes et plus largement des activités humaines constitue un enjeu majeur afin d'associer une sémantique aux actions des humains. La démarche générique, consistant à dissocier l'analyse de l'interprétation donne lieu à des investigations complémentaires dans les thèses en cours. L'analyse consiste à suivre en 3D les gestes [TH3] ou à capturer les mouve-

(a) (b) (c)

FIG. 3.4 – Caméra *time-of-light* de type CSEM (a), RFID de la société Caen (b), caméra intelligente de la société Delta-Technologie (c).

ments humains [TH5] pour interpréter les activités. L'interprétation de gestes ou activités, du fait de notre faible expertise actuelle, est à ce jour abordée à travers des collaborations internes ou externes au laboratoire.

Les contributions attendues dans [TH3] portent plus spécifiquement sur la fusion audio-visuelle pour l'interaction H/R multimodale. Ces travaux permettront notamment de construire la représentation du milieu en coopération avec l'homme. Nous nous focalisons : (i) au niveau vision, sur des gestes déictiques pour désigner l'espace et/ou les objets de l'environnement à des fins d'apprentissage, (ii) au niveau parole, sur le langage spatial ; le référent est l'homme, éventuellement les objets pour donner des commandes directionnelles au robot ou diriger son flux vidéo. On notera que l'exécution de ces actions, pour rester cohérente avec la commande vocale, est subordonnée à la localisation préalable de l'homme ou de ses membres supérieurs.

L'interprétation de la parole est traitée par l'équipe SAMOVA/IRIT[3] (I.Ferrane, MCF UPS) qui partage l'encadrement de la thèse de B.Burger [TH3]. Les variables d'observations sont ici constituées de vecteurs acoustiques relatifs aux coefficients cepstraux et à l'énergie du signal audio tandis que la reconnaissance s'appuie sur des modèles de Markov cachés (HMM) structurés hiérarchiquement. Ainsi, le moteur de reconnaissance utilisé repose au niveau bas sur un ensemble de modèles et de connaissance linguistiques (modèles de phonèmes, lexique de mots séquençant ces phonèmes), enfin une grammaire décrivant les séquences de mots du langage considéré. Le système est actuellement évalué sur la base des critères utilisés dans les campagnes NIST[4] de reconnaissance de la parole. Les évaluations portent sur des phrases/mots du langage spatial et incluant des noms courants de lieux ou objets.

[3]Institut de Recherche en Informatique de Toulouse.

[4]Pour *National Institute of Standards and Technologies.*

À l'instar de la parole, nous définirons **une grammaire de gestes**. Les noyaux de ces gestes mais également leurs préparations et leurs rétractions seront modélisées par des modèles graphiques dynamiques qui seront ensuite cascadés pour reconnaître une séquence de gestes par analogie avec la reconnaissance de phonèmes puis de mots et de phrases dans la parole. Nous nous focaliserons ici sur des modèles graphiques de type réseau dynamique bayesien (DBN) [Infantes 2006]. Ces modèles, par leur structure plus flexible que les HMMs, rendent explicites les dépendances entre variables d'état et d'observation et sont particulièrement adaptés à la fusion de plusieurs sources d'informations issues du flot vidéo [Du 2006]. Dans la lignée de nos travaux préliminaires sur la reconnaissance de gestes [TH6], nous exploiterons des informations de natures : (i) locale comme le mouvement des mains mais aussi la directivité du regard, la posture grossière, (ii) globale comme des interactions entre ces vecteurs de communication gestuelle. L'exploitation des DBNs requiert parfois des inférences approchées [Doucet 2000]... que notre connaissance des techniques de Monte-Carlo devrait nous permettre d'appréhender. Ces investigations sur la reconnaissance de gestes seront naturellement étendues à la reconnaissance d'activités élémentaires à partir du système de capture de mouvement humain développé dans [TH5].

Enfin, **les interprétations conjointes des signaux vidéo et audio seront fusionnées** pour confirmer, voire se compléter *e.g.* lors d'une commande vocale « déplace toi ici » ou « observe cet objet ». Le processus d'interprétation pourra exploiter des percepts relatifs à la perception de l'environnement. Ainsi, la nature du lieu, les objets associés, les positions relatives des agents seront également exploités, notamment pour limiter les modèles (HMM, DBN) à exécuter suivant le contexte. Enfin, d'autres informations contextuelles telles que les conditions de prises de vues pour le signal vidéo, le rapport $\frac{signal}{bruit}$ pour le signal audio viendront pondérer les probabilités de reconnaissance associées dans le mécanisme final de fusion. La fusion visio-auditive est largement abordée dans la communauté IHM. Le cadre applicatif parfois restreint et bien ciblé, le contrôle de l'environnement, en permettent (souvent) son instrumentation préalable : gants de données (ou *data glove*), PDA, écran tactile ou autres. L'interaction multimodale depuis un robot mobile, la fusion de signaux audio et visuel, sont plus complexes et, de mon point de vue, plus marginales dans la communauté Robotique [Skubic 2004, Stiefelhagen 2004, Maas 2006].

2.2 Perception de l'espace étant donné la perception de l'homme

Nous souhaitons tout d'abord approfondir et proposer des prolongements aux travaux prometteurs, développés dans [TH8], pour la représentation de l'espace et son exploitation pour la navigation autonome. La représentation de l'espace, par des considérations métriques et topologiques, a été largement abordée depuis dix ans, au moins, dans la communauté Robotique ; citons sans être exhaustif [Thrun 1998, Kuipers 2000, Tomatis 2003, Victorino 2004, Zivkovic 2006]. La prédominance d'informations sémantiques dans ces représentations est relativement récente. Ces informations sont classiquement stockées dans la représentation par codage en dur, par télé-opération du robot [Galindo 2005] ou automatiquement [Stachniss 2005, Vasudevan 2006]... pré-supposant alors un manque probable de pertinence et d'exhaustivité de ces informations au final. Ces informations sont nécessaires à la navigation autonome, mais également à la communication H/R à partir de termes ou concepts communs. Notre stratégie, dédiée à un système cognitif, est donc de capturer ces informations sémantiques, *via* la représentation du milieu, en coopération avec l'homme. Cette stratégie requiert l'intégration préalable de fonctions évoluées de perception de l'homme : suivi visuel de l'homme [TH2] pour l'exploration coordonnée de l'environnement, interprétation de gestes déictiques et du langage spatial [TH3] pour faciliter l'apprentissage du milieu et focaliser sur les entités physiques à fort contenu sémantique. Ces fonctions, qui demandent certes des investigations lourdes, sont incontournables par ailleurs dans la perspective du robot au comportement sociable. La rhétorique ci-après reprend ces différents points : la construction des représentations environnementales du bas au haut niveau (figure 1.5), leur exploitation pour la navigation autonome. Ces investigations seront menées en partie par [TH2] et par un doctorant à identifier dans le cadre d'une thèse qui sera financée, vraisemblablement, par le projet CommRob [PR3].

Des amers plus génériques et des objets dans les représentations — Nous souhaitons étendre nos travaux sur la détection/reconnaissance d'affiches murales à des amers visuels plus génériques, comme d'autres types de primitives planes, segments particuliers (arêtes verticales, plinthes, etc.), et plus largement structures locales discriminantes. La contribution sera forcément marginale tant la littérature propose aujourd'hui un large éventail d'amers visuels pour la navigation autonome [Desouza 2002]. Les amers couramment utilisés reposent sur des nuages très discriminants de points d'intérêt type Harris et surtout SIFT (pour *Scale-Invariant Feature Transform*) [Moreels 2005]. Citons ici les travaux de Goncalvès *et al.* [Goncalves

2005] qui caractérisent ainsi des pans ou places de l'espace. Ces travaux mais aussi nos travaux passés, exploitent des amers, exception faite des portes dans [DEA2], qui encodent peu ou pas d'informations sémantiques.

A l'instar de [Vasudevan 2006], nous souhaitons **incorporer très largement des objets dans nos représentations**. La construction de leur représentation structurelle (géométrie et/ou apparence), leur catégorisation, sont traitées par ailleurs dans le groupe par M.Devy (thèse en cours de M.Cottret et son prolongement). La catégorisation automatique d'objets est une problématique largement abordée dans la communauté Vision [Pons 2006]. Celle-ci, lorsqu'elle doit s'opérer depuis un robot mobile, est probablement plus complexe car les techniques doivent s'adapter à des images « plus naturelles ». Préalablement à leur catégorisation, l'information doit être extraite de son contexte par un mécanisme d'attention. Celui-ci sera ici piloté par l'homme par deux protocoles interactifs et intuitifs, dédiés aux objets manipulables ou non manipulables. Ainsi, pour les objets non manipulables, ce mécanisme exploitera la mobilité du robot dans une stratégie active de prises de vues de ces objets guidée par des gestes intuitifs de désignation [TH3]. Ces gestes, l'intentionalité associée, seront appréhendés grâce à différents percepts : la configuration spatiale bras/main mais aussi la persistence et directivité du regard.

Concernant les objets manipulables, l'homme va chercher intuitivement à présenter ces objets au robot. Il marque cette intentionalité par : son positionnement à distance « sociale » du robot, la présentation de l'objet, la persistance de son regard, éventuellement l'interpellation par la voix. Le mécanisme d'attention s'appuiera ici sur la construction de cartes de saillance « égocentré » de l'espace d'interaction au voisinage immédiat du robot. Il s'agira ici de fusionner des critères de saillance telles que : (i) la profondeur relative H/R à partir d'images de disparité, (ii) la persistance du regard à partir des détecteurs visuels [TH2], (iii) la présence de la voix à partir d'une carte acoustique [Argentieri 2005], etc. La littérature répertorie de nombreux mécanismes d'attention à partir de cartes de saillance combinant apparence, mouvement et/ou profondeur sur la scène observée ; citons par exemple [Itti 1998, Maki 2000, Cottret 2006]. Les travaux de Marfil *et al.* [Marfil 2006] sont à ma connaissance les plus proches de nos investigations futures car ils se focalisent également sur des objets manipulés par l'homme. À l'instar des travaux pré-cités, les critères de saillance restent néanmoins intrinsèques aux objets, alors que l'intentionalité de l'homme pour l'apprentissage supervisé d'objets, plus globalement pour toute tâche collaborative H/R, nous semble incontournable pour initier le mécanisme d'attention du robot. Au final, toutes les instances d'objets ou de lieux appris seront catégorisées par la parole [TH3].

Complétude des représentations — Ces amers et/ou objets viendront alors enrichir les différents modèles spatiaux dont nous disposons : cartes stochastiques denses, éventuellement multi-sensorielles, graphes d'amers et/ou d'objets, GVGs annotés par ces amers ou objets. Pour les modèles géométriques, nous nous appuierons sur les travaux menés par ailleurs au laboratoire sur le SLAM. Nos investigations passées devraient nous aider dans l'appréhension des techniques d'estimation sous-jacentes : filtre de Kalman étendu dans [Jung 2004], estimation stochastique dans [Sola 2005]. L'approfondissement des modèles qualitatifs de l'environnement doit aller au-delà du cas simple de réseaux de couloirs traité dans [TH8] tandis que nous devons faire coexister en permanence et à différents niveaux d'abstraction, des modèles métriques, topologiques, et sémantiques. À l'instar de [Thrun 1998, Zivkovic 2006], un même lieu pourra être classiquement représenté par plusieurs cartes de granularités différentes : fine par une description métrique afin de gérer les faibles dérives odométriques, grossière par une description topologique afin de gérer les problèmes d'alignement globaux liés aux fortes erreurs odométriques. Dans le prolongement des travaux passés, nous chercherons à construire une représentation hiérarchique et multi-sensorielle d'un environnement humain de grande envergure afin de valider, sur des expérimentations poussées, le processus d'exploration de l'environnement par le robot. Nous pensons nous démarquer ici des représentations métriques et topologiques usuelles **en intégrant très largement des informations sémantiques**.

On notera que la construction de la représentation environnementale converge vers un processus mi-automatique, mi-supervisé. Elle répond au double constat suivant : (i) la représentation spatiale, s'appuyant sur des entités ou informations (distances, amers ou autres) difficiles à acquérir, voire abstraites, pour l'homme est construite par le robot seul, (ii) la représentation du milieu, à travers des informations sémantiques abstraites ou du moins plus difficiles à appréhender par le robot, est construite en synergie avec l'homme.

Localisation et navigation au long court — Ces représentations seront exploitées durant la tâche de navigation. **La localisation métrique ou qualitative du robot restera multi-sensorielle**, sur la base de données visuelles et/ou télémétriques durant nos expérimentations au laboratoire. L'utilisation du laser nous semble, par contre, inopportune dans le contexte du projet CommRob [PR3]. La démarche sera d'instrumenter le supermarché par un ensemble d'étiquettes RFID qui permettront la localisation qualitative du robot lors de son passage à proximité. Ces investigations sont actuellement réalisées par A.Roguez durant son stage puis poursuivies dans

le cadre de [TH2]. Indépendamment des capteurs, la finalité sera d'inférer, pour la localisation, une estimée unique mettant en jeu l'ensemble des données sensorielles. Nous intégrerons une composante probabiliste au processus de localisation en l'état complètement déterministe : les techniques Markoviennes [Fox 1999, Kosecka 2004], notamment, devraient pouvoir répondre à nos besoins. Nous pourrons ainsi gérer plusieurs hypothèses quant à la position du robot dans plusieurs lieux de l'environnement, le recalage sur les amers/objets propres à chaque lieu étant assuré par des algorithmes dédiés. Le changement de lieu, donc éventuellement de représentations à considérer, sera notifié par des informations sémantiques telles qu'un passage de porte [DEA2], le changement de l'apparence du sol, etc. La pertinence de ces investigations et expérimentations tiendra dans une navigation au long court que nous espérons la plus robuste possible.

Navigation réactive — Une navigation plus réactive sera probablement à l'étude pour des environnements très peuplés. Nous tirerons alors partie des travaux sur le suivi visuel multi-personnes développé dans [TH1]. Cette modalité, tout en complétant la détection d'obstacles humains par laser SICK 2D, permettra la planification locale de chemins cohérents avec les trajectoires des humains au voisinage immédiat du robot. L'enjeu de ces probables investigations sera le partage harmonieux de l'espace entre le robot et les agents humains lors de leurs déplacements respectifs.

2.3 Retombées connexes attendues

L'expertise acquise à travers nos travaux sur le suivi est actuellement valorisée à travers la thèse CIFRE de G.Gelabert [TH4] et un contrat industriel avec la société ORME [CI1]. Ces travaux de thèse visent deux domaines connexes à la Robotique, à savoir l'automobile et l'aéronautique. Il s'agit respectivement : (1) de caractériser les mouvements d'un mannequin articulé lors de crash-tests automobiles (figure 3.5), (2) d'étudier les déformations d'aubes fan dans un turbo réacteur lors d'essais destructifs.

FIG. 3.5 – Analyse spatio-temporelle des mouvements d'un mannequin en crash-test par vision binoculaire.

Pour ces scènes hautement dynamiques, le but est l'analyse spatio-temporelle *a posteriori* de séquences acquises par un système multi-caméras rapides et synchrones et relevant de ces deux contextes applicatifs. L'analyse sera 3D et viendra compléter le savoir-faire de l'entreprise en matière de suivi 2D. Les travaux menés par le doctorant portent actuellement, en collaboration avec M.Devy, sur l'intégration d'algorithmes d'auto-calibrage dans le logiciel *TrackImage* développé et commercialisé par l'entreprise.

L'application aéronautique s'inscrit dans un projet région [PR1] impliquant également l'IRIT (V.Charvillat, MCF ENSEEIHT) et Snecma Moteurs client potentiel et fournisseur des données de tests ainsi que du cahier des charges. L'analyse s'appuiera sur le suivi multi-oculaire simultané mais indépendant : (1) des bords d'attaque et de fuite de chaque aube grâce aux contours image, (2) de marqueurs (type patches de texture) liés à la surface des aubes, (3) du centre de rotation. Nous évaluerons une approche filtre particulaire mixant des primitives contours et points d'intérêt [Serby 2004], puis un suivi par triangulation d'objets déformables dans la veine des travaux de V.Charvillat [Charvillat 2005].

3 Prospectives à long terme

Cette prospective mentionne quelques prolongements possibles de nos travaux puis ouvre sur l'intelligence ambiante dont l'avènement prévisible pourrait infléchir sensiblement nos travaux.

Intégration sur HRP2 de percepts multiples — Un premier prolongement de nos travaux sera d'**intégrer sur une plate-forme unique** tout ou partie des fonctions robotiques de perception de l'homme et de l'environnement. Les différents modules pourront, à travers la couche fonctionnelle du robot, échanger des percepts multiples et variés sur les lieux, objets, activités ou actions de l'homme couramment perçus ou attendus eu égard du contexte environnemental. À l'instar de [Veloso 2005], nous envisageons d'associer une représentation fonctionnelle aux catégories de lieux ou d'objets exhibées dans la représentation du milieu. La corrélation de tous ces percepts, éventuellement leurs complémentarités, permettront une interprétation plus robuste et plus complète de la scène perçue. Plus globalement, la perception du milieu influera, à tous les niveaux, sur la perception de l'espace (et vice-versa) comme évoqué dans le projet. Nous pensons utiliser ici des mécanismes d'inférence bayésienne dans la veine de [Darnell 1999, Torralba 2003].

S'agissant du support expérimental, nous privilégierons le robot HRP2 (figure 1.1-(d)), support applicatif idéal puisque doté de caractéristiques hu-

maines en termes de morphologie, perception et d'interaction. Il dispose néanmoins d'un nombre restreint de capteurs, ici la vision, avec, qui plus est, un champ de vue relativement réduit par rapport aux robots mobiles plus classiques... ce qui rend sa navigation difficile dans des environnements dynamiques. L'intégration de données sensorielles hétérogènes sera abordée en instrumentant préalablement l'environnement ; nous y reviendrons ci-après.

Vers un environnement multi-agents — Un second prolongement sera de dépasser la compréhension « simple » de situations d'interaction mettant en jeu le robot à l'homme, ou l'homme à son environnement pour interpréter des situations mettant en jeu tous les humains partageant l'environnement au voisinage immédiat du robot. Cette perspective enrichira scientifiquement nos problématiques. La perception de l'environnement est plus complexe car ces usagers, de par leurs actions, apportent de l'incertitude, et de la variabilité dans l'environnement. La perception simultanée de ces multiples agents, par leurs interactions mutuelles, sera également plus complexe. Nous suggérons de synthétiser un vecteur d'état par agent à partir d'informations propres (position, cap, identité, activités en cours, etc) et d'informations contextuelles (objets manipulés ou à proximité, nature du lieu, etc).

Ces percepts multiples devraient aider dans **l'interprétation de situations multi-agents dans cet environnement évolutif**. Pour caractériser de nouvelles situations, le principe sera de factoriser en modèles élémentaires, ou d'adapter en ligne les modèles graphiques probabilistes courants. Il s'agira par exemple de coupler des HMMs par leurs matrices de transition [Oliver 2000] ou de reconfigurer ces modèles en adaptant la structure et les paramètres [Vasquez Govea 2007].

Intelligence ambiante et robotique — La perception plus globale du milieu semble incompatible à terme avec des moyens perceptuels limités et regroupés sur un unique support robotisé... même mobile. Aussi, nous souhaitons inscrire nos travaux futurs dans la problématique de **l'intelligence ambiante**, révolution technologique programmée qui verra les environnements humains s'apparenter à des réseaux de capteurs et d'actionneurs sur une multitude de supports (mobilier, vêtements, etc). Certes, cette problématique de l'intelligence ambiante n'est pas nouvelle : elle est explorée dans de nombreux programmes d'envergure passés ou actuels *e.g.* le $6^{ème}$ programme cadre européen « technologies pour la société de l'information ». L'inclusion de ce robot assistant dans ce réseau de capteurs enrichit la problématique « ambiant » car ces robots seront considérés comme des unités de capteurs mobiles. On parle alors de robotique ubiquiste.

Dans cette perspective, nos investigations sur les capteurs RFID [TH2] pour la localisation des agents, le prototypage prévu d'un réseau de caméras intelligentes [TH1] pour la surveillance de lieux, constituent nos deux premiers jalons. Au-delà, cette ouverture m'intéresse à double titre.

Primo, la nature de la robotique en environnement humain pourrait s'en trouver bouleversée, dans la mesure *où* l'exploitation de capteurs distribués augmenterait les capacités perceptuelles embarquées et donc faciliterait la perception par le robot de l'homme et plus largement de l'environnement. Le traitement de ces divers flux de données sensorielles impliquerait des problèmes de fusion d'informations à large échelle, d'estimation, de maintien de la cohérence des représentations nécessitant probablement des approches nouvelles mais qui devraient s'inscrire dans le projet de recherche énoncé.

Secundo, les problématiques scientifiques et technologiques sous-jacentes sont nombreuses, et dépassent largement le cadre de la robotique. Ce projet, fédérateur pour le groupe, pourrait initier des collaborations avec plusieurs groupes du LAAS-CNRS dont les compétences respectives permettraient de traiter plus largement des divers aspects de la problématique : conception/mise en réseau de capteurs et effecteurs, traitement du signal, sûreté de fonctionnement, etc. Cette perspective serait alors source d'enrichissement sur un plan plus personnel.

Bibliographie

[Alami 1998] R. Alami, R. Chatila, S. Fleury, et F. Ingrand (1998). An architecture for autonomy. *Int. Journal of Robotic Research (IJRR'98)*, 17(4) :315–337.

[Argentieri 2005] S. Argentieri, P. Danès, P. Souères, et P. Lacroix (2005). An experimental testbed for sound source localization with mobile robots using optimized wideband beamformers. Dans *Int. Conf. on Intelligent Robots and Systems (IROS'05)*, pages 2536–2541, Edmonto, Canada.

[Arulampalam 2002] S. Arulampalam, S. Maskell, N. Gordon, et T. Clapp (2002). A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *Trans. on Signal Processing*, 50(2) :174–188.

[Belhumeur 1997] P. Belhumeur, J. Hepanha, et D. Kriegman (1997). Eigenfaces vs. fisherfaces : recognition using class specific linear projection. *Trans. on Pattern Analysis Machine Intelligence (PAMI'97)*, 19(7) :711–720.

[Bennewitz 2005] M. Bennewitz, F. Faber, D. Joho, M. Schreiber, et S. Behnke (2005). Towards a humanoid museum guide robot that interacts with multiple persons. Dans *Int. Conf. on Humanoid Robots (HUMANOID'05)*, pages 418–423, Tsukuba, Japan.

[Betgé-Brezetz 1996] S. Betgé-Brezetz (1996). *Modélisation incrémentale et localisation par amers pour la navigation d'un robot mobile autonome en environnement naturel*. Thèse de doctorat, Université Paul Sabatier de Toulouse.

[Booij 2007] O. Booij, B. Terwijn, Z. Zivkovic, et B. Kröse (2007). Navigation using an appearance based topological map. Dans *Int. Conf. on Robotics and Automation (ICRA'07)*, pages 3927–3932, Roma, Italy.

[Boulay 2006] B. Boulay, F. Bremond, et M. Thonnat (2006). Applying 3D human model in a posture recognition system. *Pattern Recognition Letter, Spwcial Issue on Vision for Crime Detection and Prevention*, 27(15) :1788–1796.

[Branca 2000] A. Branca et al. (2000). Landmark-based Navigation using Projective Invariants. Dans *Int. Symp. on Robotics and Automation (ISRA'00)*, pages 569–574, Monterrey, Mexico.

[Bretzner 2002] L. Bretzner, I. Laptev, et T. Lindeberg (2002). Hand gesture using multi-scale colour features, hierarchical models and particle filtering. Dans *Int. Conf. on Automatic Face and Gesture Recognition (FGR'02)*, pages 405–410, Washington, USA.

[Bulata 1996] H. Bulata et M. Devy (1996). Incremental construction of a landmark-based and topological model of indoor environment by a mobile robot. Dans *Int. Conf. on Robotics and Automation (ICRA'96)*, pages 757–764, Minneapolis, USA.

[Burgard 1999] W. Burgard, A. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, et S. Thrun (1999). Experiences with an interactive museum tour-guide robot. *Artificial Intelligence*, 114 :3–55.

[Charvillat 2005] V. Charvillat et A. Bartoli (2005). Feature-based estimation of radial basis mappings for non-rigid registration. Dans *Workshop on Vision, Modelling and Vizualization(VMV'05)*, pages 195–199, Erlanger, Germany.

[Choset 1997] H. Choset, K. Nagatani, et A. Rizzi (1997). Sensor based planning : using a honing strategy and local map method to implement the generalized voronoi graph. Dans *Graph. SPIE Mobile Robotics*, Pittsburgh, USA.

[Cottret 2006] M. Cottret et M. Devy (2006). Active learning of local structures from attentive and multi-resolution vision. Dans *Int. Conf. on Intelligent Autonomous Systems(IAS'06)*, pages 534–541, Tokyo, Japan.

[Darnell 1999] J. Darnell, A. Irfan, et M. Hayes (1999). Exploiting human actions and objet context for recognition tasks. Dans *Int. Conf. on Computer Vision (ICCV'99)*, pages 80–86, Corfou, Greece.

[Delamarre 2001] Q. Delamarre et O. Faugeras (2001). 3D articulated models and multi view tracking with physical forces. *Computer Vision and Image Understanding (CVIU'01)*, 81 :328–357.

[Desouza 2002] G. Desouza et A. Kak (2002). Vision for mobile robot navigation : A survey. *Trans. on Pattern Analysis Machine Intelligence (PAMI'02)*, 24(2) :237–267.

[Deutscher 2001] J. Deutscher, A. Davison, et I. Reid (2001). Automatic partitioning of high dimensional search spaces associated with articulated body motion capture. Dans *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'01)*, pages 669–676, Hawaï.

[Deutscher 1999] J. Deutscher, B. North, B. Bascle, et A. Blake (1999). Tracking trough singularities and discontinuities by random sampling. Dans *Int. Conf. on Computer Vision (ICCV'99)*, pages 1144–1149, Corfou, Greece.

[Doucet 2000] A. Doucet, N. De Freitas, K. Murphy, et S. Russell (2000). Rao-blackwellised particle filtering for dynamic bayesian networks. Dans *Int. Annual Conf. on Uncertainty in Artificial Intelligence (AAAI'00)*, pages 197–208, Austin, USA.

[Du 2006] Y. Du, F. Chen, W. Xu, et Y. Li (2006). Recognizing interaction activities using dynamic bayesian network. Dans *Int. Conf. on Pattern Recognition (ICPR'06)*, pages 618–621, Hong-Kong, China.

[Durieu 1996] C. Durieu, M. J. Aldon, et D. Meizel (1996). La fusion de données multisensorielles pour la localisation en robotique mobile. *Traitement du Signal (TS'96)*, 13(2) :144–166.

[Fitzpatrick 2003] P. Fitzpatrick, G. Metta, L. Natale, S. Rao, et G. Sandini (2003). Learning about objects through action-initial steps towards artificial cognition. Dans *Int. Conf. on Robotics and Automation (ICRA'03)*, pages 3140–3145, Taipei, Taiwan.

[Fletcher 2005] L. Fletcher, G. Loy, N. Barnes, et A. Zelinsky (2005). A correlating driver gaze with the road scene for driver assistance systems. *Robotics and Autonomous Systems (RAS'05)*, 52 :71–84.

[Folio 2005] D. Folio et V. Cadenat (2005). A controller to avoid both occlusions and obstacles during a vision-based navigation task in a cluttered environment. Dans *Int. Conf. on Decision and Control (CDC'05)*, pages 3898–3903, Seville, Spain.

[Fong 2003] T. Fong, I. Nourbakhsh, et K. Dautenhahn (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems (RAS'03)*, 42 :143–166.

[Fontmarty 2007] M. Fontmarty, T. Germa, B. Burger, L. Marin, et S. Knoop (2007). Implementation of human perception algorithms on a mobile robot. Dans *Int. Symp. on Intelligent Autonomous Vehicles (IAV'07)*, Toulouse, France.

[Fox 1999] D. Fox, W. Burgard, et S. Thrun (1999). Markov localization for mobile robots in dynamics environments. *Journal of Artificial Intelligence Research (JAIR'99)*, 11 :1265–1278.

[Gabriel 2005] P. Gabriel, J. B. Hayet, J. Piater, et J. Verly (2005). Object tracking using color interest points. Dans *Int. Conf. on Advanced Video and Signal-based Surveillance (AVSS'05)*, pages 159–164, Como, Italy.

[Galindo 2005] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J. Fernandez-Madrigal, et J. Gonzalez (2005). Multi-hierarchical semantic maps for mobile robotics. Dans *Int. Conf. on Intelligent Robots and Systems (IROS'05)*, pages 3492–3497, Edmonto, Canada.

[Goncalves 2005] L. Goncalves, E. Di Bernardo, D. Benson, M. Svedman, J. Ostrowski, N. Karlsson, et P. Pirjanian (2005). A visual front-end for simultaneous localization and mapping. Dans *Int. Conf. on Robotics and Automation (ICRA'05)*, pages 44–49, Barcelona, Spain.

[Guo 2006] D. Guo et X. Wang (2006). Quasi-Monte Carlo filtering in nonlinear dynamic systems. *Trans. on Signal Processing*, 54(6) :2087–2098.

[Hjelmas 2001] E. Hjelmas (2001). Face detection : a survey. *Computer Vision and Image Understanding (CVIU'01)*, 83(3) :236–274.

[Hähnel 2004] D. Hähnel, W. Burgard, D. Fox, K. Fishkin, et M. Philipose (2004). Mapping and localization with RFID technology. Dans *Int. Conf. on Robotics and Automation (ICRA'04)*, New Orleans (USA).

[Infantes 2006] G. Infantes, F. Ingrand, et M. Ghallab (2006). Apprentissage de modèle d'activité stochastique pour la planification et le contrôle d'éxécution. Dans *Reconnaissance Des Formes Et Intelligence Artificielle (RFIA'06)*, Tours, France.

[Isard 1998a] M. Isard et A. Blake (1998a). CONDENSATION – conditional density propagation for visual tracking. *Int. Journal on Computer Vision (IJCV'98)*, 29(1) :5–28.

[Isard 1998b] M. Isard et A. Blake (1998b). ICONDENSATION : Unifying low-level and high-level tracking in a stochastic framework. Dans *European Conf. On Computer Vision (ECCV'98)*, pages 893–908, London, UK.

[Isard 2001] M. Isard et J. MacCormick (2001). BraMBLe : a bayesian multiple blob tracker. Dans *Int. Conf. on Computer Vision (ICCV'01)*, volume 1, pages 34–41, Vancouver, Canada.

[Isard 1998c] M. A. Isard et A. Blake (1998c). A mixed-state condensation tracker with automatic model-switching. Dans *Int. Conf. on Computer Vision (ICCV'98)*, pages 107–112, Bombay, India.

[Itti 1998] L. Itti, C. Koch, et E. Niebur (1998). A model of saliency-based visual attention for rapid scene analysis. *Trans. on Pattern Analysis Machine Intelligence (PAMI'98)*, 20(11) :1254–1259.

[Jung 2004] I. K. Jung (2004). *SLAM in 3D Environments using Stereo Vision.* Thèse de doctorat, Institut National Polytechnique de Toulouse.

[Kalman 1960] R. Kalman (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82 :35–45.

[Kanda 2007] T. Kanda, M. Shiomi, L. Perrin, T. Nomura, H. Ishiguro, et N. Hagita (2007). Analysis of people trajectories with ubiquitous sensors in a science museum. Dans *Int. Conf. on Robotics and Automation (ICRA'07)*, pages 4846–4853, Roma, Italy.

[Knoop 2006] S. Knoop, S. Vacek, et R. Dillmann (2006). Sensor fusion for 3D human body tracking with an articulated 3D body model. Dans *Int. Conf. on Robotics and Automation (ICRA'06)*, pages 1686–1691, Orlando, USA.

[Kortenkampf 1994] D. Kortenkampf et T. Weymouth (1994). Topological mapping for mobile robots using a combination of sonar and vision sensing. Dans *National Conf. on Artificial Intelligence (AAAI'94)*, pages 979–984, Washington, USA.

[Kosecka 2004] J. Kosecka et F. Li (2004). Vision based topological markov localization. Dans *Int. Conf. on Robotics and Automation (ICRA'04)*, pages 1481–1486, New Orleans, USA.

[Kuipers 2000] B. Kuipers (2000). The spatial semantic hierarchy. *Artificial Intelligence*.

[Kulyukin 2004] V. Kulyukin, C. Gharpure, J. Nicholson, et S. Pavithran (2004). RFID in robot-assisted indoor navigation for the visually impaired. Dans *Int. Conf. on Intelligent Robots and Systems (IROS'04)*, pages 1979–1984, Sendai, Japan.

[Li 2002] P. Li et T. Zhang (2002). Visual contour based on sequential importance sampling/resampling algorithm. Dans *Int. Conf. on Pattern Recognition (ICPR'02)*, pages 564–568, Quebec, Canada.

[Li 2006] Y. Li, H. Ai, C. Huang, et S. Lao (2006). Robust head tracking with particles based on multiple cues fusion. Dans *ECCV Workshop on HCI*, pages 29–39, Berlin, Germany.

[Liu 2004] Y. Liu et Y. Jia (2004). A robust hand tracking for gesture-based interaction of wearable computers. Dans *Int. Symp. on Wearable Computers (ISWC'04)*, pages 22–29, Hiroshima, Japan.

[Maas 2006] J. Maas, T. Spexard, J. Fritsch, B. Wrede, et G. Sagerer (2006). BIRON, what's the topic ? a multimodal topic tracker for improved human-robot interaction. Dans *Int. Symp. on Robot and Human Interactive Communication (RO-MAN'06)*, Hatfield, UK.

[MacCormick 2000] J. MacCormick et M. Isard (2000). Partitioned sampling, articulated objects, and interface-quality hand tracking. Dans *European Conf. on Computer Vision (ECCV'00)*, pages 3–19, London, UK.

[Maki 2000] A. Maki, P. Nordlund, et J. Eklundh (2000). Attentional scene analysis : Integrating depth and motion. *Computer Vision and Image Understanding (CVIU'00)*, 78(35) :351–373.

[Marfil 2006] R. Marfil, R. Vazquez-Martin, L. Molina-Tanco, A. Bandera, et F. Sandoval (2006). Fast attentional mechanism for a social robot. Dans *Workshop on Human Robot Interaction (HRI'06)*, Palermo, Italy.

[Matthieu 2005] A. Matthieu, J. Crowley, V. Devin, et G. Privat (2005). Localisation intra-bâtiment multi-technologies : RFID, wifi et vision. Dans *Journées Francophones : Mobilité et Ubiquité (UbiMob'05)*, Grenoble, France.

[Moeslund 2001] T. Moeslund et E. Granum (2001). A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding (CVIU'01)*, 81(3) :231–268.

[Montiel 2001] S. Montiel et A. Zisserman (2001). Automated architectural acquisition from a camera undergoing planar motion. Dans *Virtual Augmented Architecture (VAA'01)*, pages 207–218, Dublin, Ireland.

[Moreels 2005] P. Moreels et P. Perona (2005). Evaluation of features detectors and descriptors based on 3D objects. Dans *Int. Conf. on Computer Vision (ICCV'05)*, pages 800–807, Beijing, China.

[Moreno 2002] F. Moreno, A. Tarrida, J. Andrade, et A. Sanfeliu (2002). 3D real-time head tracking fusing color histograms and stereovision. Dans *Int. Conf. on Pattern Recognition (ICPR'02)*, volume 1, pages 368–371, Quebec, Canada.

[Morisset 2002] B. Morisset (2002). *Vers un robot au comportement robuste. Apprendre à combiner des modalités sensorimotrices complémentaires.* Thèse de doctorat, Université Paul Sabatier de Toulouse.

[Oliver 2000] N. Oliver, B. Rosario, et A. Pentland (2000). A bayesian computer vision system for modeling human interactions. *Trans. on Pattern Analysis Machine Intelligence (PAMI'00)*, 22(8) :831–843.

[Pfaff 2006] P. Pfaff, W. Burgard, et D. Fox (2006). Robust monte-carlo localization using adaptative likelihood models. Dans *European Robotics Symp. (EUROS'06)*, pages 181–194, Palermo, Italia.

[Philomin 2000] V. Philomin, R. Duraiswami, et L. S. Davis (2000). Quasi-random sampling for condensation. Dans *European Conf. on Computer Vision (ECCV'00)*, pages 134–149, Dublin, Ireland.

[Pineau 2003] J. Pineau, M. Montemerlo, M. Pollack, N. Roy, et S. Thrun (2003). Towards robotic assistants in nursing homes : challenges and results. *Robotics and Autonomous Systems (RAS'03)*, 42 :271–281.

[Pitt 1999] M. K. Pitt et N. Shephard (1999). Filtering via simulation : Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446) :590–599.

[Pons 2006] J. Pons, M. Hebert, C. Schmid, et A. Zisserman (2006). *Towards Category-level Object Recognition*. Springer-Verlag.

[Porta 2004] J. Porta et B. Kröse (2004). Appearance-based concurrent map building and localization using multi-hypotheses tracker. Dans *Int. Conf. on Intelligent Robots and Systems (IROS'04)*, pages 3424–3429, Sendai, Japan.

[Pérez 2004] P. Pérez, J. Vermaak, et A. Blake (2004). Data fusion for visual tracking with particles. *IEEE*, 92(3) :495–513.

[Qu 2007] W. Qu, D. Schonfeld, et M. Mohamed (2007). Distributed bayesian multiple-target tracking in crowded environments using multiple collaborative cameras. *EURASIP Journal on Advances in Signal Processing*.

[Rabiner 1989] L. Rabiner (1989). A tutorial on hidden markov models and selected applications in speech recognition. *IEEE*, 77(2) :257–286.

[Rui 2001] Y. Rui et Y. Chen (2001). Better proposal distributions : Object tracking using unscented particle filter. Dans *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'01)*, pages 786–793, Hawai.

[Serby 2004] D. Serby, E. Meier, et L. Van Gool (2004). Probabilistic object tracking using multiple features. Dans *Int. Conf. on Pattern Recognition (ICPR'04)*, Cambridge, UK.

[Sidenbladh 2000] H. Sidenbladh, M. J. Black, et D. J. Fleet (2000). Stochastic tracking of 3D human figures using 2D image motion. Dans *European Conf. on Computer Vision (ECCV'00)*, pages 702–718, Dublin, Ireland.

[Siegwart 2003] R. Siegwart et al. (2003). ROBOX at expo 0.2 : a large scale installation of personal robots. *Robotics and Autonomous Systems (RAS'03)*, 42 :203–222.

[Sigal 2004] L. Sigal, S. Bhatia, S. Roth, M. Black, et M. Isard (2004). Tracking loose-limbed people. Dans *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'04)*, Washington, USA.

[Sim 1999] R. Sim et G. Dudek (1999). Learning Visual Landmark for Pose Estimation. Dans *Int. Conf. on Robotics and Automation (ICRA'99)*, pages 1972–1978, Detroit, USA.

[Skubic 2004] M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, et W. Adams (2004). Spatial language for human-robot dialogs. *Trans. on Systems, Man, and Cybernetics*, 34.

[Sminchisescu 2003] C. Sminchisescu et B. Triggs (2003). Estimating articulated human motion with covariance scaled sampling. *Int. Journal on Robotic Research (IJRR'03)*, 6(22) :371–393.

[Sola 2005] J. Sola, A. Monin, M. Devy, et T. Lemaire (2005). Undelayed initialization in bearing only SLAM. Dans *Int. Conf. on Intelligent Robots and Systems (IROS'05)*, Edmonto, Canada.

[Stachniss 2005] C. Stachniss, O. Martinez-Mozos, A. Rottman, et W. Burgard (2005). Semantic labeling of places. Dans *Int. Symp. of Robotics Research (ISRR'05)*, San Francisco, USA.

[Stauffer 1999] C. Stauffer et W. Grimson (1999). Adaptative background mixture models for real-time tracking. Dans *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'99)*, volume 2, pages 22–46, Fort Collins.

[Stenger 2001] B. Stenger, P. Mendonça, et R. Cipolla (2001). Model-based hand tracking using an unscented kalman filter. Dans *British Machine Vision Conference (BMVC'01)*, volume 1, pages 63–72, Manchester, UK.

[Stiefelhagen 2004] R. Stiefelhagen, C. Fügen, P. Gieselmann, H. Holzapfel, K. Nickel, et A. Waibel (2004). Natural human-robot interaction using speech, head pose and gestures. Dans *Int. Conf. on Intelligent Robots and Systems (IROS'04)*, pages 2422–2427, Sendai, Japan.

[Suleiman 2006] W. Suleiman, A. Monin, et J. Laumond (2006). Synthesizing and modeling human locomotion using system identification. Dans *Int. Conf. on Intelligent Robots and Systems (IROS'06)*, pages 1972–1977, Beijing, China.

[Thrun 1999] S. Thrun, M. Bennewitz, W. Burgard, A. Cremers, F. Dellaert, D. Fox, D. Halnel, C. Rosenberg, N. Roy, J. Schulte, et D. Schulz (1999). MINERVA : A second generation mobile tour-guide robot. Dans *Int. Conf. on Robotics and Automation (ICRA'99)*, pages 14–26, Detroit, USA.

[Thrun 1998] S. Thrun, J. Gutmann, D. Fox, W. Burgard, et B. Kuipers (1998). Integrating topological and metric maps for mobile robot navigation : A statistical approach. Dans *National Conf. on Artificial Intelligence (AAAI'98)*, pages 989–995, Madison, USA.

[Tomatis 2003] N. Tomatis, I. Nourbakhsh, et R. Siegwart (2003). Hybrid simultaneous localization and map building : A natural integration of topological and metric. *Robotics and Autonomous Systems (RAS'03)*, 44 :3–14.

[Torma 2003] P. Torma et C. Szepesvári (2003). Sequential importance sampling for visual tracking reconsidered. Dans *AI and Statistics*, pages 198–205, Key West, USA.

[Torralba 2003] A. Torralba, K. Murphy, W. Freeman, et M. Rubin (2003). Context-based vision system for place and object recognition. Dans *Int. Conf. on Computer Vision (ICCV'03)*, pages 80–86, Nice, France.

[Uk 2007] C. Uk, S. Hun, P. Dai, et J. Wook (2007). Multiple objects tracking circuit using particle filters with multiple features. Dans *Int. Conf. on Robotics and Automation (ICRA'07)*, pages 4639–4644, Roma, Italy.

[Urtasum 2004] R. Urtasum et P. Fua (2004). 3D human body tracking using deterministic temporal motion models. Dans *Europ. Conf. on Computer Vision (ECCV'04)*, Prague, Czech Republic.

[Vandapel 2000] N. Vandapel (2000). *Perception et sélection d'amers en environnement polaire pour la navigation d'un robot mobile.* Thèse de doctorat, Institut National Polytechnique de Toulouse.

[Vasquez Govea 2007] A. Vasquez Govea (2007). *Incremental Learning for Motion Prediction of Pedestrians and Vehicles.* Thèse de doctorat, Institut National Polytechnique de Grenoble.

[Vasudevan 2006] S. Vasudevan, V. Nguyen, et R. Siegwart (2006). Towards a cognitive probabilistic representation of space for mobile robot. Dans *Int. Conf. on Intelligent Robots and Systems (IROS'06)*, Beijing, China.

[Veloso 2005] M. Veloso, F. Hundelshausen, et P. Rybski (2005). Learning visual object definitions by observing human activities. Dans *Int. Conf. on Humanoid Robots (HUMANOID'05)*, pages 148–153, Tsukuba, Japan.

[Vermaak 2002] J. Vermaak, P. Pérez, M. Gangnet, et A. Blake (2002). Towards improved observation models for visual tracking : Selective adaptation. Dans *European Conf. on Computer Vision (ECCV'02)*, pages 645–660, Copenhagen, Denmark.

[Victorino 2004] A. Victorino et P. Rives (2004). An hybrid representation well-adapted to the exploration of large scale indoors environments. Dans *Int. Conf. on Robotics and Automation (ICRA'04)*, pages 2930–2935, New Orleans, USA.

[Victorino 2001] A. Victorino, P. Rives, et J. J. Borelly (2001). Mobile robot navigation using a sensor-based control strategy. Dans *Int. Conf. on Robotics and Automation (ICRA'01)*, Seoul, Korea.

[Viola 2001] P. Viola et M. Jones (2001). Rapid object detection using a boosted cascade of simple features. Dans *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'01)*, Hawaii.

[Wu 2001] Y. Wu, J. Y. Lin, et T. S. Huang (2001). Capturing natural hand articulation. Dans *Int. Conf. on Computer Vision (ICCV'01)*, pages 426–432, Vancouver, Canada.

[Yu 2004] T. Yu et Y. Wu (2004). Collaborative tracking of multiple targets. Dans *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'04)*, pages 834–841, Washington, USA.

[Zhao 2004] T. Zhao et R. Nevatia (2004). Tracking multiple humans in crowded environment. Dans *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'04)*, Washington, USA.

[Ziegler 2007] J. Ziegler, K. Nickel, et R. Stiefelhagen (2007). Tracking of the articulated upper body on multi-view stereo image sequences. Dans *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'07)*, Minneapolis, USA.

[Zivkovic 2006] Z. Zivkovic, B. Bakker, et B. Kröse (2006). Hierarchical map building and planning based on graph partitioning. Dans *Int. Conf. on Robotics and Automation (ICRA'06)*, pages 803–809, Orlando, USA.

Deuxième partie

Valorisation de la recherche

1 Animation scientifique

L'animation scientifique englobe ici les activités d'encadrement (thèses, stages, etc) ainsi que mes implications diverses dans l'organisation de manifestations, de groupes de travail et autres activités de lecture critique. Ce chapitre reprend ces divers points.

1.1 Activités d'encadrement

Conscient que la formation à et par la recherche est un des rôles fondamentaux de l'enseignant-chercheur, j'ai encadré ou co-encadré des thèses, stages de fin d'études M2R (ex-DEA), M2P (ex-DESS), ou ingénieur. **Ces activités d'encadrement m'ont permis d'être titulaire de la Prime d'Encadrement Doctoral et de Recherche (PEDR) depuis 2000.** Ces activités, détaillées ci-après, se résument à :

- 8 encadrements de thèse passés ou en cours,
- 9 encadrements de M2R,
- 7 encadrements de stages M2P ou ingénieurs,
- nombreux tutorats de stages en entreprises.

1.1.1 Direction de thèses de doctorat

[TH1] Iker Zurriarain (09/2007-...). **Implémentation d'algorithmes de suivi multi-personnes sur FPGA pour caméras intelligentes.** Thèse de l'Université de Mondragon (Espagne) en co-direction avec Nestor Arana (MCF Université de Mondragon, 50%). Thèse financée par le projet AMISEG. Publication associée :[10].

[TH2] Thierry Germa (11/2006-...). **Navigation et perception multimodale de l'homme en environnement actif.** Thèse de l'Université Paul Sabatier en co-direction avec Patrick Danès (MCF UPS, 50%). Thèse financée par le projet européen CommRob. Publications associées : [2, 15, 16, 38].

[TH3] Brice Burger (10/2006-...). **Fusion de données audiovisuelles pour l'interaction Homme/Robot.** Thèse MENRT de l'Université Paul Sabatier en co-direction avec Isabelle Ferrané (MCF IRIT/ UPS, 50%). Publications associées : [12, 13].

[TH4] Guillaume Gelabert (05/2006-...). **Auto-calibrage d'un système multi-caméras rapides synchrones et analyse du mouvement 3D. Applications automobile et aéronautique.** Bourse CIFRE avec la société ORME de Toulouse. Thèse de l'Université Paul Sabatier en co-direction avec Michel Devy (DR LAAS-CNRS, 50%). Travaux

incluant des clauses de confidentialité.

[TH5] Mathias Fontmarty (2005-2008). **Vision et filtrage particulaire pour le suivi tridimensionnel de mouvement humain. Applications à la Robotique**. Thèse MENRT de l'Université Paul Sabatier en co-direction avec Patrick Danès (MCF UPS, 50%). Soutenance en décembre 2008.
- rapporteurs : M.O.Berger (CR HDR LORIA Nancy) et P.Pérez (DR INRIA Rennes).
- publications associées : [3, 14, 37, 36, 11].
- devenir : ATER INSA Toulouse.

[TH6] Paulo Menezes (2002-2006). **Multiple-cue Visual Tracking for Human-Robot Interaction.** Thèse de l'Université de Coimbra (Portugal) en co-direction avec J.Dias (PR ISR Coimbra, 20%). Soutenue le 13 juillet 2007. Mention très honorable.
- rapporteurs : A.Cazals (PR Université Polytechnique de Catalogne) et M.Devy.
- publications associées : [1, 17, 19, 20, 24, 40, 41].
- devenir : enseignant-chercheur à ISR Coimbra, Portugal.

[TH7] Ludovic Brèthes (2002-2005). **Suivi visuel par filtrage particulaire. Application à l'interaction Homme/Robot.** Thèse MENRT de l'Université Paul Sabatier en co-direction avec Patrick Danès (50%). Soutenue le 13 décembre 2005. Mention très honorable.
- rapporteurs : P.Pérez (DR INRIA Rennes) et F.Jurie (CR HDR INRIA Grenoble)
- publications associées : [4, 21, 22, 24, 39, 41].
- devenir : création de la *start-up* noomeo, Castres.

[TH8] Jean-Bernard Hayet (1999-2003). **Contribution à la navigation d'un robot mobile sur amers visuels texturés dans un environnement structuré.** Thèse BDI de l'Université Paul Sabatier en co-direction avec Michel Devy (50%). Soutenue le 29 janvier 2003. Mention très honorable.
- rapporteurs : P.Rives (DR INRIA Sophia) et M.Dhome (DR CNRS Clermont-Ferrand)
- publications associées : [5, 7, 25, 26, 27, 30, 34, 35, 43].
- devenir : chercheur au CIMAT, Université de Guanajuato, Mexique.

En 2002, j'ai ponctuellement participé à l'encadrement de la thèse de Nestor Arana (MCF Université Mondragon) alors doctorant dans le groupe RIA du LAAS-CNRS. Cette implication s'est concrétisée par une publication [29].

1.1.2 Direction de Mastère 2 Recherche (M2R)

[DEA1] Iker Zurriarain, Ecole Polytechnique de Mondragon (2007). **Suivi multi-personnes à partir de caméras de surveillance.**

[DEA2] Sandro Maury, UPS (2006). **Détection probabiliste de portes par vision monoculaire à partir d'un robot mobile.** Co-direction avec Michel Devy.

[DEA3] Mathias Fontmarty, INSAT (2005). **Suivi d'un bras humain par *Unscented Kalman Filter*.** Co-direction avec Patrick Danès.

[DEA4] Florent Lanterna, INSAT (2004). **Contributions à la navigation d'un robot mobile en environnement d'intérieur.** Co-direction avec Michel Devy.

[DEA5] Claudia Esteves, UPS (2003). **Construction d'une carte topologique pour la navigation qualitative en environnement d'intérieur.** Co-direction avec Michel Devy.

[DEA6] Yann Rotrou, INSAT (2002). **Reconnaissance gestuelle par vision pour l'interaction Homme/Robot.** Co-direction avec Patrick Danès.

[DEA7] Pascal Belaubre, INSAT (2001). **Modélisation d'objet à partir d'un système de lumière structurée monté sur un bras robotique.** Co-direction avec Michel Devy.

[DEA8] Jean-Bernard Hayet, ENSTA (1999). **Localisation d'un robot mobile autonome en milieu intérieur à partir d'une caméra embarquée.**

[DEA9] Sylvain Durand, INSAT (1998). **Détection et localisation d'objets mobiles dans une scène structurée.** Co-direction avec Michel Devy.

1.1.3 Encadrement et tutorat de stages de fin d'études

Je distingue ici l'encadrement de stages ingénieur effectués au laboratoire et donnant lieu à un suivi régulier de tutorat de stages en entreprises *où* mon implication était très ponctuelle. Ces stages étaient d'une durée de 5 à 6 mois.

- **Encadrement de stages ingénieur**
  - sur le thème perception de l'environnement pour la navigation : A.Roguez (2007), L.Prunet (2004), S.Mas et A.Cuniasse (2003), J.Carbajo et B.Zwick (2002).
  - sur le thème perception de l'homme pour l'interaction H/R : M. Destephes (2007), T.Germa (2006), A.Yvernès (2004).
- **Tutorat de stages industriels :** Depuis 1998, trois à quatre tutorats par

an de :

- stages de fin d'études d'étudiants de Mastère 2 Professionnel (ex-DESS) mention Systèmes Intelligents, Intelligence Artificielle, Reconnaissance des Formes et Robotique ou Productique.
- stages de Mastère 1 (ex-Maîtrise) de l'Institut Universitaire Professionnel (IUP) Systèmes Intelligents (IUP SI).

1.2 Fonctions d'intérêt général

1.2.1 Organisations de manifestations scientifiques

- 2007 : membre du comité de programme du congrès doctoral *Electronics, Control, Modelling, Measurement and Signals* (*ECMS'07*). Cette école s'est déroulée du 21 au 23 mai 2007 à Libérec. Toutes les informations sont accessibles à l'URL `www.mechatronika.cz/ecms2007`
- 2006 : membre du comité scientifique du *workshop* européen *Visual-based Human/Robot Interaction* le 18 mars à Palerme (Italie). Ce *workshop* s'est déroulé durant le premier symposium européen de Robotique (*EUROS'06*) du 15 au 18 mars. Lien URL : `paloma.isr.uc.pt/~hri06`.
- 2004 : membre du comité d'organisation du congrès Reconnaissance des Formes et Intelligence Artificielle (*RFIA'04*) du 28 au 31 janvier à Toulouse. Il constitue le congrès francophone majeur en Vision, se déroule tous les deux ans, et regroupe environ 350 chercheurs de la communauté Vision. Lien URL : `www.laas.fr/rfia2004`.
- 2001 : Membre du comité de programme du congrès francophone ORASIS. Ce congrès regroupait 120 chercheurs de la communauté Vision par ordinateur du 5 au 8 juin 2001 à Cahors.
  Lien URL `www.irit.fr/ORASIS2001`.
- 2001 : coordinateur du comité d'organisation du Symposium International on Intelligent Robotic Systems (*SIRS'01*) du 16 au 18 juin à Toulouse. Ce symposium regroupait 70 chercheurs de la communauté Robotique. Lien URL : `www.laas.fr/sirs2001`.

1.2.2 Groupes de travail

- *Research Activity 2 (RA2)*, projet COGNIRON : participation au groupe de travail *Detection and Understanding of human Activity* incluant 5 partenaires européens du projet COGNIRON.
- *Workpackage* 4 (WP4), projet CommRob : responsable/coordinateur du groupe de travail *Human Motion Interpretation* incluant 3 partenaires européens du projet CommRob.

- GDR nationaux GT5 Robotique et ISIS Vision : participation à journées thématiques en tant qu'orateur ou simple auditeur.
- SigVision, pôle RIA du laboratoire : participation aux réunions bimensuelles fédérant une quinzaine de chercheurs du laboratoire sur le thème *perception*.
- Thème n° 2, groupe RAP du laboratoire : animation/coordination au quotidien de ce thème impliquant 3 permanents, 4 doctorants, 2 à 4 stagiaires par an et porteur de ce thème dans les différents projets associés du laboratoire.

1.2.3 Activités de lecture critique

- *IEEE Transactions on Robotics (TRO)*,
- *IEEE International Journal of Pattern Recognition (PR)*,
- *IEEE International Conference on Robotics and Automation (ICRA)*,
- *IEEE/RSJ International Conference on Intelligent RObots and Systems (IROS)*,
- *IEEE International Conference on Computer Vision (ICCV)*,
- *IEEE Transactions on Instrumentation and Measurement Journal*,
- *International Symposium on Intelligent Robotic Systems (SIRS)*,
- *IEEE International Conference on 3D Digital Imaging and Modeling (3DIM)*,
- *International Joint Conference on Artificial Intelligence (IJCAI)*,
- *International Conference on Model-based Imaging, Rendering, Analysis and Graphical Effects (MIRAGE)*,
- Journal Européen des Systèmes Automatisés *(JESA)*,
- *IEEE International Conference on Intelligent Transportation Systems*,
- Congrès francophone de Vision *(ORASIS)*,
- Programme Robotique et Entités Artificielles *(RobEA)*,
- Projet région Languedoc-Roussillon.

2 Contrats, collaborations et projets de recherche

Conscient que le transfert technologique est une composante majeure de nos activités, mon implication auprès du milieu industriel toulousain a démarré dès 1998 avec le projet PREDIT [CI3]. Les contacts et relations noués ou entretenus depuis avec le milieu industriel local ont permis récemment la signature de deux contrats. Mes implications dans ces différents contrats, collaborations et projets de recherche du pôle RIA sont détaillées ci-après par ordre chronologique inverse. Elles se résument depuis septembre 1997 à

des responsabilités/participations dans :
- 2 projets européens de recherche,
- 3 contrats industriels,
- 1 projet national de recherche,
- 1 projet trans-pyrénéen,
- 1 projet régional de recherche,
- 3 collaborations internationales.

Contrats industriels

[CI1] 04/2006-2008 : **co-responsable avec M.Devy du contrat avec la société ORME de Labège.** Mes nombreuses visites en tant que tuteur de stagiaires (section 1.1.3) dans la société ORME ont initiées une collaboration scientifique qui s'est concrétisée par un contrat de recherche avec cette PME locale dirigée conjointement par L.Oriat et M.L.Meyer. La thèse CIFRE de G.Gelabert [TH4] s'inscrit dans ce contrat. J'ai donc naturellement contribué à son élaboration tandis que je participe actuellement à sa gestion scientifique en tant que co-encadrant de la thèse CIFRE. Ce projet s'inscrit dans le cadre de l'évolution du logiciel *TrackImage* développé par la société ORME. Cette évolution porte sur l'intégration de nouveaux algorithmes dédiés à l'auto-calibrage d'un système multi caméras synchrones et l'analyse 3D de scènes très dynamiques à partir de ce système.

[CI2] 2006-2007 : **responsable d'un contrat équipe-conseil entre le LAAS-CNRS et Siemens VDO de Toulouse**. Le but est de donner des avis et conseils scientifiques sur la problématique suivante : détection et le suivi de véhicules routiers par vision embarquée. Ce contrat, limité à des prestations intellectuelles, se matérialise par des échanges scientifiques sur l'état de l'art dans le domaine. Sa description se veut ici sommaire car il contient une clause de confidentialité.

[CI3] 1998-2000 : **participation au contrat PREDIT[5] avec Siemens VDO de Toulouse et le CERT[6] de l'ONERA.**. Le but du projet est de développer un module de perception visant à l'amélioration de la sécurité des passagers d'une automobile. La perception doit ainsi caractériser l'état courant du siège passager afin de mieux contrôler le déclenchement des systèmes de coussins gonflables. Notre contribution couvre deux volets : (1) la mise au point d'un capteur 3D, de type caméra associée à une source de lumière structurée, et (2) l'évaluation

[5]Programme national de recherche et d'innovation dans les transports terrestres.
[6]Centre d'Études et de Recherches de Toulouse.

pour cette application, de nos algorithmes de stéréovision dense par corrélation. J'ai participé à la gestion scientifique de ce projet, notamment en focalisant mes investigations sur le premier volet.

Projets de recherche

[PR1] 2007-2009 : **participation au projet AMISEG financé par la communauté de travail des pyrénées entre la région Midi-Pyrénée et le pays basque espagnol.** Ce projet, coordonné par N.Arana (Université de Mondragon), se focalise sur le prototypage d'un réseau de caméras vidéo intelligentes autonomes pour la surveillance de personnes. Ces capteurs devront intégrer leurs propres ressources CPU (cartes FPGA), éventuellement énergétiques (batteries solaires), et devront pouvoir échanger des informations entre eux *via* un réseau sans fil. Les partenaires du projet sont l'Université de Mondragon (correspondant : N.Arana) pour le câblage des algorithmes de vision sur FPGA et Innovae Vision une *start-up* de San Sebastian (correspondant : P.Ayala) pour les spécifications et l'étude de marché. Les algorithmes de vision sont développés au LAAS-CNRS par I.Zurriarran [TH1], dans le cadre de sa thèse en co-tutelle avec l'Université de Mondragon.

[PR2] 2007-2009 : **responsable d'un projet régional de transfert technologique.** Ce projet s'inscrit dans l'appel annuel à projets de recherche pour le transfert de technologies de la région Midi-Pyrénées. Il est en forte corrélation avec la thèse CIFRE de G.Gelabert [TH4] effectuée au sein de la société ORME. Le projet se focalise sur le calibrage du banc multi-caméras dédié aux essais puis sur l'analyse spatio-temporelle des déformations des aubes fan d'un réacteur sous des contraintes mécaniques spécifiques obtenues par simulation d'incidents[7]. Les partenaires du projet sont l'IRIT (correspondant : V.Charvillat), les sociétés ORME et Snecma Moteurs client potentiel et fournisseur des données de tests et du cahier des charges.

[PR3] 2007-2009 : **participation au projet européen CommRob[8].** Ce projet, coordonné par G.Novak (Université de Vienne) s'inscrit dans la ligne d'action *Advanced Robotics* du domaine STREP de la Priorité Technologies de la Société de l'Information du $6^{ème}$ PCRD. Il implique des transferts technologiques vers les industries partenaires du projet. La finalité du projet est la conception et la mise en œuvre de chariots

[7] typiquement la rétention d'aubes ou l'ingestion de volatiles.

[8] Pour plus de détails, consulter l'URL suivante : commrob.zenon.gr.

mobiles autonomes évoluant dans un lieu public type supermarché ou aéroport. Les partenaires du projet sont :
- Vienna University of Technology (TUW), Vienne, Autriche.
- Kunglika Tekniska Högskolan (KTH), Stockholm, Suède.
- Forschungszentrum Informatik (FZI), Karlsruhe, Allemagne.
- Zenon industry, Athènes, Grèce.
- Electronics Trading & Production e. K. industry (ETP), Karlsruhe, Allemagne.

En coordination avec deux chercheurs du groupe, je me suis impliqué dans la définition du projet et la rédaction du dossier associé. Le projet est structuré autour de huit *Work Packages* (WPs). Nous contribuerons essentiellement sur le WP n° 3 (*Advanced robot behaviour and navigation*), en particulier la navigation à partir d'amers visuels et de RFIDs, et surtout le WP n° 4 (*Perception for HRI*) que je coordonne avec mes collègues. La thèse de T.Germa [TH2], portant majoritairement sur le WP n° 4, sera financée sur les fonds propres de ce projet.

[PR4] 2004-2008 : **participation au projet européen COGNIRON**[9]**.** Ce projet, coordonné par R.Chatila (DR LAAS-CNRS, groupe RIS), s'inscrit dans la ligne d'action *Beyond Robotics* du domaine *Future and Emerging Technologies* (FET) de la Priorité Technologies de la Société de l'Information du $6^{ème}$ PCRD. L'objectif scientifique du projet est de conférer des capacités cognitives aux robots à travers l'étude et le développement de méthodes et de technologies pour la perception, l'interprétation, le raisonnement et l'apprentissage en interaction avec l'homme. Le robot n'est pas considéré ici comme une machine pré-programmée mais comme une créature artificielle, dont les capacités se développent dans un processus continu d'acquisition de nouvelles connaissances et compétences. Les partenaires du LAAS-CNRS sont :
- L'Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Suisse.
- Fraunhofer Institut fur Produktionstechnik und Automatisierung (IPA, Stuttgart, Allemagne.
- Kunglika Tekniska Högskolan (KTH), Stockholm, Suède.
- Universität Bielefeld (UniBi), Bielefeld, Allemagne.
- University of Hertfordshire (UH), Hatfield, Royaume-Uni.
- Universität Karlsruhe-TH (UniKarl), Karlsruhe, Allemagne.
- Vrije Universiteit Brussel (VUB), Bruxelles, Belgique.
- Gesellschaft für Produktionssyteme (GPS), Stuttgart, Allemagne.

Le projet est structuré autour de six groupes de travail et la réalisation de trois démonstrations fédératrices. Nos travaux s'inscrivent dans le

[9]Pour plus de détails, consulter l'URL `www.cogniron.org`

groupe de travail $n°$ 2 (*detection and understanding of human activity*) et la démonstration $n°$ 2 (*the curious robot*). Dans ce projet, je suis impliqué à travers :
- L'encadrement de doctorants [TH2, TH5, TH7] ou stagiaires (T.Germa et A.Yvernès) pour les besoins plus ponctuels du projet.
- La participation aux différentes réunions de travail propres au RA2 et KE2,
- La rédaction des documents d'avancement relatifs à RA2 [51, 52] ou KE2 [53].
- implication dans l'école d'hiver (janvier 2008).

[PR5] 2001-2004 : **participation au projet national RobEA HR+ en collaboration avec la Cité de l'Espace de Toulouse.** Ce projet, coordonné par R.Alami (DR LAAS-CNRS), s'inscrit dans le programme Robotique et Entités Artificielles (RobEA) du CNRS. Les partenaires sont l'Institut de la Communication Parlée (INPG Grenoble) et l'équipe GRAVIR/IMAG de Grenoble. Le but du projet est l'étude et la conception d'un robot assistant intégrant des capacités de perception, de décision, de navigation et surtout d'interaction avec les visiteurs de l'exposition « Mission Biospace » située à la Cité de l'Espace. Notre contribution a porté sur la conception et l'intégration de deux modules sur ce robot (figure 2.8) : (1) le module « LuckyLoc » de localisation visuelle sur amers colorimétriques développé lors d'un stage [DEA3], (2) le module « ICU » de détection et suivi image de personnes développé dans [TH6] et [TH7]. Les travaux menés dans ce projet sont synthétisés dans les actes des journées RobEA de mars 2005 [54].

Collaborations internationales

Indépendamment des projets ou contrats mentionnés, j'entretiens ou ai entretenu des collaborations avec les institutions suivantes :

[CO1] 2003-2005 : **collaboration avec ISR de l'Université de Coimbra (Portugal), projet GRICES du CNRS $n°$ 14155.** Correspondant à Coimbra : J.Dias. Je suis responsable côté LAAS-CNRS de cette collaboration avec l'Institut des Systèmes et Robotique (ISR). Cette collaboration donne lieu à des échanges scientifiques via des séjours croisés sur le thème perception de l'homme pour l'interaction H/R. La thèse au sein de notre groupe de P.Menezes [TH6] s'inscrit dans cette collaboration pour laquelle je me suis impliqué dans son élaboration puis sa gestion scientifique.

[CO2] 1999-2003 : **collaboration avec Faculté de Salamanca (Mexiqu-**

e), projet ECOS-Nord $n°$ **M99M01.** Correspondants à Salamanca : J.Rivas et V.Ayala. Cette coopération qui porte sur la navigation de robots mobiles à usage agricole s'est traduit par des échanges scientifiques via des séjours croisés. La dispense d'enseignements à l'Université de Guanajuato (partie III) a permis le recrutement en thèse de deux étudiants mexicains G.Avina et J.Goncalvez. Impliqué dans la gestion scientifique de cette coopération.

[CO3] 1999-2001 : **collaboration avec Institut Beckman de Urbana-Champaign, (USA).** Correspondant à Urbana-Champaign : S.Hutchinson. Cette coopération donna lieu à des échanges scientifiques via des séjours croisés sur le thème planification de stratégies de navigation et perception. Cette coopération s'est concrétisée par l'intégration sur une plate-forme commune des travaux de J.B.Hayet [TH8] et P.Ranganathan étudiant de Urbana-Champaign et la rédaction d'une publication dans une revue internationale [7]. Je suis intervenu dans la gestion scientifique de cette coopération.

3 Publications et rapports

Les activités de recherche réalisées durant mon doctorat au LASMEA puis en tant qu'enseignant-chercheur au LAAS-CNRS m'ont permis à ce jour de réaliser 51 publications qui se décomposent comme suit :

- 1 contribution à ouvrage collectif international,
- 8 revues internationales,
- 1 revue nationale,
- 1 brevet,
- 28 congrès internationaux,
- 10 congrès nationaux.

Ces publications ainsi que quelques rapports, liés à mes activités contractuelles sont énumérés ci-après. On notera que les deux années postérieures à ma mobilité géographique et thématique, en septembre 1997, sont logiquement peu publiantes.

Publications réalisées depuis ma nomination MCF

Contribution à ouvrage :

1. Towards an Interactive Humanoid Companion with Visual Tracking Modalities. P.Menezez, F.Lerasle, J.Dias et T.Germa. Chapter of book titled Humanoid Robots. Edited by International Journal of Advanced Robotic Systems (*ARS'07*), pp 48-78. ISBN 978-3-902613-07-3.

Revues internationales :

2. Video-based Face Recognition and Tracking from a Robot Companion. T.Germa, F.Lerasle, T.Simon. International Journal of Pattern recognition and Artificial Intelligence. Article accepté et sous presse.
3. Evaluations of Particle Filter based Human Motion visual Trackers for Home Environment Surveillance. M.Fontmarty, P.Danès, F.Lerasle. International Journal of Pattern Recognition and Artificial Intelligence. Article en révision.
4. Particle Filtering Strategies for Data Fusion dedicated to Visual Tracking from a mobile Robot. L.Brèthes, F.Lerasle, P.Danès, M.Fontmarty. Journal Machine Vision and Applications (*MVA'08*). A paraître. Lien URL : *http ://dx.doi.org/10.1007/s00138-008-0174-7.*
5. A Visual Landmark Framework for Mobile Robot Navigation. J.B.Hayet, F.Lerasle, M.Devy. Journal of Image and Vision Computing (*IVC'07*), vol 25, numéro 8, pp 1341-1351, August 2007.
6. Comparison of Structured Light and Stereovision Sensors for New Airbag Generations. S.Boverie, M.Devy, F.Lerasle. Control Engineering Practice Journal (*CEP'03*), vol 11, num 12, pp 1413-1421, December 2003.
7. Topological Navigation and Qualitative Localization for Indoor Environment using Multisensory Perception. P.Ranganathan, J.B.Hayet, M.Devy, S.Hutchinson, F.Lerasle. Robotics and Autonomous Systems Journal (*RAS'02*), vol 41, num 2, pp 137-144, November 2002.
8. Tracking of Human Limbs by Multi-ocular Vision[10]. F.Lerasle, G.Rives, M.Dhome. Computer Vision and Image Understanding Journal (*CVIU'99*). Volume 75, num 3, pp 229-246. September 1999.

Brevet :

9. Dispositif optoélectronique permettant la détection et la classification des occupants de l'habitacle d'une automobile utilisant une source lumineuse unique. Brevet n° fr 0012987, 2000. Co-déposants : Siemens Automotive, CERT, co-auteurs LAAS-CNRS : M.Devy et F.Lerasle.

Articles parus dans congrès internationaux avec comités de lecture :

10. An MCMC-based Particle Filter for multiple Person Tracking. I.Zuriarrain, F.Lerasle, N.Arana, M.Devy. IEEE International Conference on Pattern Recognition (*ICPR'08*), 4 pages, Tampa, USA, December 2008. A paraître.

[10]Article portant sur mes travaux de thèse mais rédigé au LAAS-CNRS.

11. Towards Real-time Markerless Human Motion Capture from Ambiance Cameras using and hybrid Particle Filter. M.Fontmarty, F.Lerasle, P.Danès. IEEE International Conference on Image Processing (*ICIP'08*), San Diego, USA, October 2008.
12. Mutual Assistance between Speech and Vision for Human-Robot Interaction. B.Burger, F.Lerasle, I.Ferrané. IEEE/RSJ International Conference on Intelligent Robots and Systems (*IROS'08*), Nice, France, October 2008.
13. Multimodal Interaction Abilities for a Robot Companion. B.Burger, I.Ferrané, F.Lerasle. International Conference on Vision Systems (*ICVS'08*), 10 pages, Santorini, Greece, Mai 2008.
14. Data Fusion within a Modified Annealed Particle Filter dedicated to Human Motion Capture. M.Fontmarty, F.Lerasle, P.Danès. IEEE/RSJ International Conference on Intelligent Robots and Systems (*IROS'07*), San Diego, USA, October 2007.
15. Human/Robot Visual Interaction for a Tour-Guide Robot. T.Germa, F.Lerasle, P.Danès, L.Brèthes. IEEE/RSJ International Conference on Intelligent Robots and Systems (*IROS'07*), San Diego, USA, October 2007.
16. Data Fusion and Eigenface based Tracking dedicated to a Tour-Guide Robot. T.Germa, L. Brèthes, F.Lerasle, T.Simon. International Conference on Vision Systems (*ICVS'07*), 10 pages, Bielefeld, March 2007.
17. Visual Tracking Modalities for a Companion Robot. P.Menezes, F.Lerasle, J.Dias. IEEE/RSJ International Conference on Intelligent Robots and Systems (*IROS'06*), pp 5363-5368, Beijing, China, October 2006.
18. Data Fusion for 3D Gestures using a Camera mounted on a Robot. P.Menezes, F.Lerasle, J.Dias. IEEE International Conference on Pattern Recognition (*ICPR'06*), 4 pages, Hong-Kong, August 2006.
19. A Single Camera Motion Capture System dedicated to Gestures Imitation. P.Menezes, F.Lerasle, J.Dias, R.Chatila. IEEE/RAS International Conference on Humanoid Robots, pp 430-435, Tsukuba, Japan, December 2005. **Article sélectionné pour parution dans *Advanced Robotic Systems book* [1].**
20. Appearance-based Tracking of 3D Articulated Structures. P.Menezes, F.Lerasle, J.Dias, R.Chatila. International Symposium on Robotics - (*ISR'05*), Tokyo, Japan, November 2005.
21. Data Fusion for Visual Tracking dedicated to Human-Robot Interaction. L.Brèthes, F.Lerasle, P.Danès. IEEE International Conference on

Robotics and Automation (*ICRA'05*), pp 2087-2092, Barcelona, Spain, April 2005.

22. Face Tracking and Hand Gesture Recognition for Human-Robot Interaction. L.Brèthes, P.Menezes, F.Lerasle, J.B.Hayet. IEEE International Conference on Robotics and Automation (*ICRA'04*), pp 1901-1906, New Orleans, USA, May 2004.
23. Environment Modeling for Topological Navigation using Visual Landmarks and Range Data. F.Lerasle, J.Carbajo, M.Devy, J.B.Hayet. IEEE International Conference on Robotics and Automation (*ICRA'03*), pp 1330-1335, Taiwan, May 2003.
24. Visual Tracking of Silhouettes for Human-Robot Interaction. P.Menezes, L.Brèthes, F.Lerasle, P.Danès, J.Dias. International Conference on Advanced Robotics (*ICAR'03*), pp 971-976, Coimbra, Portugal, June 2003.
25. Visual Landmarks Detection and Recognition for Mobile Robot Navigation. J.B.Hayet, F.Lerasle, M.Devy. IEEE International Conference on Computer Vision and Pattern Recognition (*CVPR'03*), pp 313-318, Madison, USA, June 2003.
26. A Visual Landmark Framework for Indoor Mobile Robot Navigation. J.B.Hayet, F.Lerasle, M.Devy. IEEE International Conference on Robotics and Automation (*ICRA'02*), vol 4, pp 3942-3947, Washington, USA, May 2002.
27. Qualitative Modeling of Indoor Environments from Visual Landmarks and Range Data. J.B.Hayet, C.Esteves, M.Devy, F.Lerasle. IEEE/RSJ International Conference on Intelligent Robots and Systems (*IROS'02*), pp 631-636, Lausanne, Switzerland, September 2002.
28. 3D Perception for New Airbag Generation. S.Boverie, M.Devy, F.Lerasle. World Congress of the International Federation of Automatic Control (*IFAC'02*), 6 pages, Barcelona, Spain, July 2002.
29. A Smart Sensor based Visual Landmarks Detection for Indoor Robot Navigation. N.Arana, F.Lerasle, M.Briot, C.Lemaire, J.B.Hayet. IEEE International Conference on Pattern Recognition (*ICPR'02*), 4 pages, Quebec, Canada, August 2002.
30. Topological Navigation and Qualitative Localization for Indoor Environment using Multisensory Perception. P.Ranganathan, J.B.Hayet, M.Devy, S.Hutchinson, F.Lerasle. International Symposium on Intelligent Robotic Systems (*SIRS'01*), pp 497-506, Toulouse, July 2001.

31. Projected Light Beams Tracking for Efficient 3D Reconstruction. F.Lerasle, P.Danès. IEEE International Conference on Image Processing (*ICIP'01*), pp 951-951, Thessaloniki, Greece, October 2001.
32. Relaxation vs Maximal Cliques Search for Projected Beams Labeling in a Structured Light Sensor. F.Lerasle, J.M.Lequellec, M.Devy. IEEE International Conference on Pattern Recognition (*ICPR'00*), pp 782-785, Barcelona, Spain, September 2000.
33. Car Cockpit 3D Reconstruction by a Structured Light Sensor. J.M.Lequellec, F.Lerasle. IEEE Intelligent Vehicle Symposium. (*IV'00*), pp 87-92, Deadborn, USA, October 2000.
34. Visual Localization of a Mobile Robot in Indoor Environments using Planar Landmarks. V.Ayala, J.B.Hayet, F.Lerasle, M.Devy. IEEE/RSJ International Conference on Intelligent Robots and Systems (*IROS'00*), vol 1, pp 275-280, Takamatsu, Japan, November 2000.
35. Planar Landmarks to Localize a Mobile Robot. J.B.Hayet, F.Lerasle, M.Devy. International Symposium on Intelligent Robotic Systems - (*SIRS'00*), pp 163-169, Berkshire, UK, July 2000.

Articles parus dans congrès nationaux avec comités de lecture :

36. Une stratégie hybride de filtrage particulaire pour la capture de mouvement depuis un robot mobile. M.Fontmarty, F.Lerasle, P.Danès. Congrès francophone Reconnaissance des Formes et Intelligence Artificielle (*RFIA-'08*), Amiens, Janvier 2008.
37. Filtrage particulaire pour la capture de mouvement dédiée à l'interaction homme-robot. M.Fontmarty, F.Lerasle, P.Danès, P.Menezes. Congrès francophone ORASIS, 8 pages, Strasbourg, Juin 2007.
38. Suivi et identification de personnes par un robot guide. T.Germa, L.Brèthes, F.Lerasle, T.Simon. Congrès francophone ORASIS, 8 pages, Strasbourg, Juin 2007.
39. Stratégies de filtrage particulaire pour le suivi visuel de personnes : description et évaluation. L.Brèthes, P.Danès, F.Lerasle. Congrès francophone Reconnaissance des Formes et Intelligence Artificielle (*RFIA'06*), 10 pages, Tours, Janvier 2006.
40. Suivi visuel de structures articulées 3D par filtrage particulaire. P.Menezes, F.Lerasle, J.Dias, R.Chatila. Congrès francophone ORASIS, 8 pages, Fournol, Mai 2005.
41. Segmentation couleur et condensation pour le suivi et la reconnaissance de gestes humains. L.Brèthes, P.Menezes, F.Lerasle, M.Briot. Congrès francophone Reconnaissance des Formes et Intelligence Artificielle (*RFIA'04*), vol 2, pp 967-975, Toulouse, Janvier 2004.

42. Supervision de l'habitacle d'un véhicule par lumière structurée. J.M.Lequellec, F.Lerasle, S.Boverie. Congrès francophone Reconnaissance des Formes et Intelligence Artificielle (*RFIA'02*), vol 1, pp 87-95, Angers, Janvier 2002.
43. De l'utilisation d'amers plans pour la navigation d'un robot mobile en milieu intérieur. J.B.Hayet, F.Lerasle, M.Devy. Congrès francophone ORASIS, pp 193-202, Cahors, Juin 2001.

Publications réalisées durant mon doctorat

Mémoire de thèse :

44. « Vers le suivi du geste sportif par vision artificielle ». Thèse préparée au LASMEA de l'Université Blaise Pascal. Soutenue le 13 janvier 1997.

Revues :

45. Suivi du corps humain par vision monoculaire. F.Lerasle, G.Rives, M.Dhome, A.Yassine. Traitement du Signal (*TS'97*), vol 16, num 6, pp 675-685, 1997.
46. Leg Cycling Tracking by Dynamic Vision. F.Lerasle, G.Rives, M.Dhome, J.M.Garcié, E.Van Praagh. Journal of Biomechanics, vol 30, num 8, pp 837-840, 1997.

Articles parus dans congrès internationaux avec comités de lecture :

47. Human Body Tracking by Multi-ocular Vision. F.Lerasle, G.Rives, M.Dhome. Scandinavian Conference on Image Analysis (*SCIA'97*), vol 1, pp 285-292, Lappeenranta, Finland, June 1997.
48. Human Body Tracking by Monocular Vision. F.Lerasle, G.Rives, M.Dhome. European Conference on Computer Vision (*ECCV'96*), vol 2, pp 518-527, Cambridge, UK, April 1996.

Articles parus dans congrès nationaux avec comités de lecture :

49. Suivi du corps humain par vision multi-oculaire. F.Lerasle, G.Rives, M.Dhome. Reconnaissance des Formes et Intelligence Artificielle (*RFIA'98*), vol 2, pp 193-200, Clermont-Fd, Janvier 1998.
50. Suivi du corps humain par vision monoculaire. F.Lerasle, G.Rives, M.Dhome, A.Yassine. Reconnaissance des Formes et Intelligence Artificielle (*RFIA'96*), vol 2, pp 859-868, Rennes, Janvier 1996. **Article sélectionné pour parution dans Traitement du signal [45].**

Quelques rapports de projets et contrats réalisés

51. *Detection and Understanding of Human Activity.* Rapports de Contrat, COGNIRON, Project FP6-IST-002020, 2006, 110 pages. Document à diffusion limitée.
52. *Detection and Understanding of Human Activity.* Rapports de Contrat, COGNIRON, Project FP6-IST-002020, 2005, 70 pages. Document à diffusion limitée.
53. *Mid-Term Key-experiment Specification and Implementation Status.* Rapports de Contrat, COGNIRON, Project FP6-IST-002020, 2006, 30 pages.
54. *HR+ : Towards an Interactive Autonomous Robot.* Journées RobEA, pp 39-45, Montpellier, Mars 2005.
55. Système intelligent de supervision optique de l'habitacle d'un véhicule. Rapport de contrat, PREDIT II. 37 pages. Novembre 2000. Document à diffusion limitée.
56. Architecture pour la planification, le contrôle et l'interaction RObotique de SErvice. Rapport de projet ROSE. 17 pages. Décembre 1999.

Troisième partie

Activités d'enseignement

1 Enseignements dispensés

Préambule

Depuis 13 ans, j'enseigne dans la section 61, section regroupant les domaines de l'Automatique, Informatique Industrielle et Traitement du Signal. Ces enseignements ont été réalisés en tant que vacataire à l'Université Blaise Pascal de Clermont-Ferrand (1994-1997) puis de Maître de conférences à l'Université Paul Sabatier (UPS) de Toulouse depuis 1997. Cette mutation géographique a logiquement impliqué la remise à plat de tous mes enseignements.

Le tableau 1.1 synthétise l'ensemble des enseignements réalisés durant la période 1994-2008. Son analyse amène à quelques commentaires généraux. Sur le plan pédagogique, mon recrutement à l'UPS était motivé par le départ à la retraite, quelques années plus tard, de M.Briot ; celui-ci était alors le seul intervenant en image de l'équipe pédagogique section 61. Durant la période 1997-2000, j'ai assuré majoritairement des enseignements d'Automatique et d'Informatique Industrielle. Ces enseignements, dispensés par le passé (pour la plupart), étaient alors non couverts par mes collègues enseignants. J'ai également assuré quelques enseignements de perception et imagerie numérique. A partir de 2000, la retraite de M.Briot et surtout la création de nouveaux diplômes ou modules relevant de l'imagerie numérique ont fait émerger des besoins dans les thématiques perception pour la robotique, vision, traitement et analyse des images. Pour cette nouvelle charge de service, j'ai alors : (1) réactualisé les enseignements existants, (2) défini le contenu et mis en place les nouveaux enseignements associés. Aujourd'hui, j'interviens dans la plupart des formations de l'UPS relevant des sections 34, 61 et/ou 27, qui dispensent des enseignements en imagerie numérique et/ou robotique.

Proches de mes activités de recherche, ces enseignements me permettent aujourd'hui d'apprécier la dualité entre enseignement et recherche propre au statut d'enseignant-chercheur. Ainsi, les cours dispensés, par leur réactualisation permanente, tirent profit des expériences et connaissances acquises durant mes activités de recherche. Réciproquement, la préparation de cours permet souvent de bonifier des connaissances nécessaires par ailleurs pour mes activités de recherche. La figure 1.6 illustre de façon grossière les recouvrements thématiques entre les volets recherche et enseignement matérialisé par mes quatre cours principaux actuels : formation des images, traitement des images, vision industrielle, perception pour la Robotique.

TAB. 1.1 – Évolution des enseignements dispensés depuis 1994 (% – *cycle universitaire*)

Années universitaires	1994-1997	1997-1999	1999-2000	2000-2001	2001-2002	2002-2006	2006-2007
Info. Industrielle	$40\%, 2^e$cycle	$5\%, 2^e$	-	-	-	-	-
Info. générale	-	$40\%, 1^e - 2^e$	$10\% - 2^e$	–	–	–	–
Automatique	$60\% - 2^e$cycle	$50\% - 2^e$	$40\% - 2^e$	$10\% - 2^e$	$10\% - 2^e$	–	–
Perception Rob.	-	$5\% - 2^e$	$25\% - 2, 3^e$	$25\% - 2, 3^e$	$25\% - 2, 3^e$	$30\% - 2, 3^e$	$15\% - 2^e, 20\% - 3^e$
Trait./anlyse d'images	-	-	$15\% - 2^e$	$20\% - 2^e$	$20\% - 2^e$	$30\% - 2^e$	$25\%, 2^e$
Formation des images	-	-	-	$15\% - 2^e$	$15\% - 2^e$	$15\% - 2^e$	$15\%, 2^e$
Vision industrielle	-	-	-	$20\% - 3^e$	$15\% - 3^e$	$15\% - 3^e$	$15\%, 3^e$
Divers (projets, TERs)	-	-	10%	10%	15%	10%	10%

Autres interventions : CNAM, INSA de Toulouse (INSAT), Ecole des Mines d'Albi-Carmaux (EMAC), Faculté de Salamanca (Mexique).

Convaincu qu'un enseignement d'image doit être par définition illustré (imagé !), **tous les cours sont dispensés à l'aide de transparents.** Les transparents sont regroupés dans un polycopié fourni aux étudiants en début de cours. Son contenu est complété en séance par la prise de notes et développements au tableau.

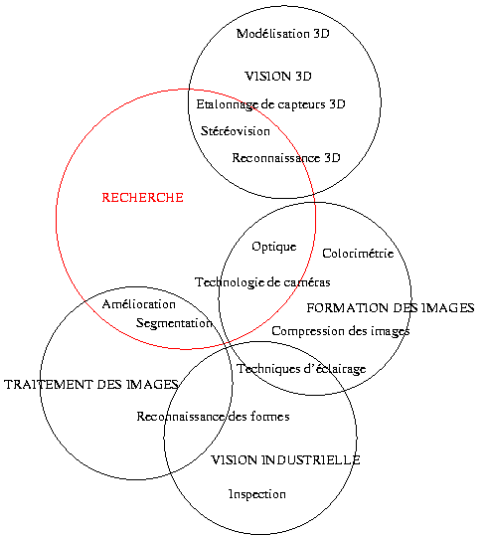


FIG. 1.6 – Recouvrement thématiques entre recherche et enseignement.

Les interventions sont très différentes de par : (1) leurs natures *i.e.* cours, travaux dirigés (TD), travaux pratiques (TP), bureaux d'études (BE) ou travaux d'étude et de recherche (TER), (2) les publics concernés *i.e.* second et troisième cycles universitaires, formations en Électronique, Électrotechnique et Automatique (EEA), formations universitaires « classiques » (Licence et Mastère), professionnalisées (IUP) et écoles d'ingénieurs. Les enseignements assurés depuis 1994 sont détaillés ci-après par ordre chronologique et thématique.

Vacations durant mon doctorat

Les vacations assurées durant mes trois années de doctorat à l'Université Blaise Pascal ont permis de confirmer mon intérêt pour l'enseignement et d'acquérir une première expérience pédagogique. Mes premières vacations datent de 1994 et portent sur l'encadrement de travaux pratiques d'automatique linéaire continue et discrète, de systèmes à événements discrets en maîtrise EEA. Durant les années suivantes, j'ai renouvelé cette expérience en maîtrise EEA mais également au CNAM et en école d'ingénieurs universitaire (CUST, option Génie Physique). Enfin, pour cette dernière formation, j'ai assuré des travaux dirigés d'automatique linéaire et discrète. Les vacations assurées durant ces trois années représentent environ 210 heures équivalent TD.

Services de Maître de conférences UPS

Les enseignements d'Automatique et Informatique Industrielle ont été dispensés majoritairement durant les trois années qui suivirent mon recrutement

(1997-2000). Sans lien direct avec mes activités de recherche, je ne détaillerai pas davantage. Le départ de M.Briot en 2000, la création de nouveaux modules ou diplômes relatifs à l'imagerie numérique m'ont amené à abandonner peu à peu les enseignements d'automatique et informatique industrielle. En effet, de par mes activités de recherche, j'étais le seul enseignant de l'équipe pédagogique à pouvoir prendre en charge ce service. Ces enseignements, regroupés sous les mots-clés image et perception pour la robotique, sont décrits ci-après.

- Licence IUP Systèmes Intelligents (SI) : **Cours et TD sur la formation et compression des images numériques.** Cet enseignement, dispensé initialement par M.Briot, décrit : (1) les éléments intervenants dans la mise en œuvre d'une chaîne d'acquisition d'images, (2) les étapes et facteurs intervenant dans la mise en forme d'un signal vidéo de qualité, (3) son stockage par compression avec/sans pertes. Les connaissances acquises par les étudiants sont validées grâce à des exercices corrigés en séance.
  Le contenu de ce cours/TD a subi de nombreuses refontes, en particulier pour prendre en compte les innovations technologiques liées aux capteurs d'images. Ces ré-actualisations successives, souvent discutées avec M.Briot m'ont permis de rédiger un polycopié assez complet. L'objectif est aujourd'hui d'en faire un ouvrage pédagogique de collection [S1] accessible aux milieux universitaire et industriel. Cette volonté est motivée par le constat suivant : les ouvrages de collection sur le traitement et analyse des images sont nombreux dans la communauté Image mais ils restent assez laconiques quant aux outils de prises de vues et à leurs mises en œuvre. Notre ouvrage permettrait modestement de combler cette lacune.
  Le support [S2] distribué aux étudiants est agencé sous forme d'une centaine de transparents qui tirent partie du polycopié [S1]. Ce cours/TD, d'un volume de 24h, est complété par un TP illustrant l'acquisition des images et la colorimétrie sous le logiciel industriel Aphelion. Signalons enfin que la dispense de cet enseignement me permet de côtoyer les étudiants de cette Licence, formation dont j'ai la responsabilité depuis septembre 1999 (section 3).
- Maîtrise EEA, parcours Traitement De l'Information (TDI) : **Cours, TP et TER de traitement et analyse des images numériques.** M.Briot, G.Flouzat et moi-même avons défini le contenu et mis en place un module d'image en adéquation avec ce nouveau parcours de Maîtrise créé en 1999. Depuis 2000, j'assure la moitié des enseignements rattachés à ce module, l'autre moitié étant dispensée par G.Flouzat. L'éventail de mes interventions est assez large et balaie la chaîne de

perception depuis l'acquisition jusqu'à l'interprétation de l'image : acquisition, compression, filtrage, amélioration et restauration, segmentation, analyse de texture, morphologie mathématique, reconnaissance des formes. Le support de cours [S3] relatif au traitement, analyse et interprétation des images, regroupe 80 transparents présentés en séance. Les travaux pratiques mis en place illustrent certaines notions de cours puis des séances de TER s'articulent autour de la résolution de petits problèmes tirant partie des nombreux champs applicatifs de l'imagerie numérique (inspection, télédétection, analyse de matériau,...). Les séances que j'encadre s'appuient sur le logiciel industriel Aphelion tandis que les étudiants découvrent un second logiciel avec G.Flouzat.
La dernière habilitation du diplôme en 2004 a donné lieu à une refonte du contenu de cet enseignement et sa restructuration en deux modules de 48 h chacun. J'assure la responsabilité pédagogique de ces deux unités d'enseignement depuis 2004 et la responsabilité pédagogique des travaux pratiques associés à cet enseignement depuis son origine (section 3).

- Maîtrise IUP Instrumentation, Capteurs et Mesures (ICM) : **Cours et TP de traitement et analyse des images numériques.** L'équipe pédagogique $34^{ème}$ section, autour de A.Bernès (PR UPS), a créée en 2004 ce nouvel IUP. Durant l'année 2005-2006, il m'a été demandé de dispenser un cours d'introduction à l'imagerie numérique (volume horaire : 8 h) au premier semestre de la Maîtrise IUP ICM. Cette initiation s'appuie sur le cours de capteurs et instrumentation, vu par ailleurs dans le cursus, pour présenter le synoptique classique d'une chaîne de perception et quelques applications de l'imagerie numérique. Ce cours est illustré par deux travaux pratiques sur la métrologie. Cette introduction permet aux étudiants de suivre, au second semestre, le module d'imagerie numérique du Mastère IS2I.
- Maîtrise IUP SI : **Cours et BE de Perception 3D.** J'ai créé et mis en place cet enseignement qui couvre les aspects suivants : capteurs 3D, modélisation 3D, localisation/reconnaissance 3D. Je distribue un support [S4] qui regroupe une soixantaine de transparents résumant les notions importantes. Ce cours, dispensé sur 14 h, est illustré par deux BE autour de 10 postes de stéréovision dédiés aux étudiants. Le but est d'appréhender les grandes étapes de la stéréovision : acquisition de données stéréo, étalonnage du système, reconstruction 3D. Les étudiants sont ici utilisateurs des outils mis à leur disposition ("toolbox" Matlab, interface Cygwin, visualisateur 3D).
- DESS Intelligence Artificielle, Reconn. des formes, et Robotique (IRR) : **Cours et BE de Perception pour la Robotique.** Cet enseignement

reprend les grandes lignes de l'enseignement dispensé en Maîtrise IUP SI. Concernant le cours, il s'appuie sur un horaire plus conséquent 20 h) et permet un approfondissement différent. J'assure la moitié du cours, l'autre moitié étant dispensée par M.Devy. Je reproduis ici les BE de Maîtrise IUP SI afin d'illustrer les concepts de la stéréovision.

- DESS Productique : **Cours et BE de vision industrielle.** J'ai hérité de cet enseignement suite au départ de M.Briot en 2000. J'ai souhaité ré-orienter le contenu du cours et BE associés afin de coller au mieux aux spécificités de la formation et aux applications de vision industrielle associées. Ainsi, le cours d'un volume de 16 h reproduit de façon synthétique certains outils et concepts d'imagerie numérique dèjà mentionnés et met l'accent sur les techniques de traitement des images adaptées à l'inspection visuelle en milieu industriel (environnement contrôlé, contraintes temps réel, etc). Dans ce cadre, les types et techniques d'éclairage sont par exemple introduites. Le cours se termine par quelques applications de vision industrielle. Le support [S5] distribué aux étudiants regroupe 90 transparents présentés en séance. Les trois bureaux d'études mis en place se focalisent sur le contrôle qualité de pièces manufacturées (métrologie, vérification de présence, tri,...).
- **DESS Télédétection et Imagerie Numérique (TIN)** : **Cours et BE de vision 3D.** J'ai créé et mis en place cet enseignement qui introduit quelques concepts de vision 3D puis se focalise sur la stéréovision. Un support de cours regroupant une trentaine de transparents rétroprojetés durant les séances est distribué au préalable aux étudiants. Le cours, dispensé sur 9 h est complété par des séances de BE qui s'articulent autour de la réalisation d'un mini projet encadré sur 12 h. Les séances s'organisent autour de programmes Matlab "à trous" permettant aux étudiants de concevoir un programme complet de stéréovision dense.
- DESS SI : **Projet annuel de fin d'étude.** Ce projet se déroule sur plusieurs mois et implique tous les étudiants de la formation. En début d'année universitaire, l'objet du projet est discuté, éventuellement redéfini, en présence de l'équipe pédagogique et des étudiants. Je suis amené à participer à ce projet (conception, suivi, évaluation) lorsque le cahier des charges implique la mise en œuvre et l'exploitation d'un système de vision.

Enseignements hors UPS

- Ecole des Mines d'Albi-Carmaux (EMAC) depuis 2004 : S.Leroux (IR EMAC) et moi-même avons défini et dispensons actuellement un module optionnel de reconnaissance des formes pour l'imagerie numérique aux étudiants de troisième et quatrième année de l'EMAC. Ces enseignements s'articulent autour de séances de cours puis de travaux pratiques. Une première partie, dispensée par S.Leroux, porte sur des notions générales puis la segmentation en contours. Mon intervention porte sur la segmentation région, la présentation de descripteurs associés, enfin la reconnaissance des formes. Pour illustrer cette seconde partie, j'ai mis en place des travaux pratiques spécifiques qui s'appuient sur la boîte à outils reconnaissance (*recognition toolkit*) du logiciel Aphelion. Le support de cours [S6] distribué aux étudiants comporte une trentaine de transparents pour un volume horaire total de 20 h. Le contenu de ce module optionnel est modifié depuis 2006 et porte sur la vision industrielle.
- INSA de Toulouse (INSAT) en 2005 : T.Sentenac (MCF EMAC) et moi-même avons mis en place et assuré un cours d'introduction à l'imagerie numérique aux étudiants de quatrième année GEII de l'INSA de Toulouse. Ce cours devait s'inscrire dans un module optionnel Électronique et Multimédia. Une première partie, dispensée par mon collègue, était consacrée à des notions générales et l'acquisition des images. J'ai complété ce cours par une introduction à la stéréovision, puis une présentation des techniques de compression d'images, enfin une description des normes de compression vidéo en vigueur. Le support de cours distribué aux étudiants comportait une quarantaine de transparents pour un volume horaire total de 20 h.
- Faculté de Salamanca (Mexique) en 2001 : Lors d'une mission relevant d'une collaboration de recherche (CO2, partie II), j'ai été chargé d'assurer un cours destiné aux étudiants de Mastère (bac+5) de la Faculté d'Ingéniérie en Mécanique, Électrotechnique et Électronique (FIMEE) de Salamanca située dans l'état de Guanajuato. Ce cours intitulé *3D vision*, était dispensé en anglais. Il s'appuyait sur le cours dispensé en DESS TIN. Le support de cours [S7] distribué aux étudiants comporte 70 transparents pour un volume horaire de 15 h de cours.

2 Supports de cours

Sont mentionnés ci-après quelques supports de cours liés à mes activités pédagogiques. Le premier concerne la rédaction d'un ouvrage pédagogique en cours de rédaction, les suivants sont des polycopiés de transparents présentés en séances aux étudiants.

[S1] Formation et compression des images. M.Briot et F.Lerasle. 150 pages. Ouvrage pédagogique en cours de rédaction.

[S2] Formation et compression des images - Support de cours. 100 transparents.

[S3] Traitement et analyse des images - Support de cours. 80 transparents.

[S4] Perception pour la robotique - Support de cours. 60 transparents.

[S5] Vision industrielle - Support de cours. 90 transparents.

[S6] Reconnaissance des formes - Support de cours. 30 transparents.

[S7] *3D Vision* - Support de cours en anglais. 70 transparents.

3 Fonctions d'intérêt général et responsabilités pédagogiques

Préambule

Mes responsabilités pédagogiques portent essentiellement sur une Licence IUP et deux modules de mastère tandis que je suis membre de diverses commissions et conseils. Ces différentes charges d'intérêt général assurées au titre de l'enseignement sont détaillées ci-après.

Responsabilités pédagogiques

1. à l'IUP SI (1999-...) : **Co-responsabilité d'année avec P.Joly (PR UPS). La charge de travail hebdomadaire moyenne est évaluée à 2 heures par co-responsable.** Cet IUP, créé par M.Briot en 1992 et l'équipe pédagogique du DESS IRR, dispense une formation pluridisciplinaire à l'intersection de l'informatique, de l'automatique et de l'informatique industrielle. Elle permet à l'élève ingénieur-maître d'acquérir des compétences scientifiques[11], spécialisées[12], enfin générales d'ouverture vers l'entreprise. Elle s'appuie sur des collaborations

[11] Mathématiques, automatique, informatique.

[12] Intelligence artificielle, interface homme-machine, robotique.

avec le secteur industriel et les laboratoires de recherche publics et privés et le secteur industriel qui dispense certains enseignements dans la formation IUP.

Le cursus de l'IUP se déroule sur trois ou quatre années : les étudiants démarrent au niveau bac+2, éventuellement bac+3 (Licence SI), et sortent au niveau bac+4 ou bac+5 (DESS SI) après admission par l'équipe pédagogique. Les deux dernières années IUP sont validées par un stage industriel de 5 ou 6 mois. La formation recrute sur dossiers puis auditions pour : (1) les étudiants bac+2 donc titulaires d'un DEUG, DUT, BTS, CPGE, (2) les étudiants bac+1 donc ayant validés une première année d'enseignement supérieur (DEUG ou CPGE).

Depuis 1999, je suis responsable de la Licence qui comporte chaque année une quarantaine d'étudiants alors que la direction de l'IUP est assurée par M.Taix (MCF UPS) depuis 2000. La responsabilité de la Licence est probablement la plus contraignante des quatre années du cursus de par le volume horaire à gérer (850 h/an). Le rôle du responsable est fondamental dans le fonctionnement de l'année qui lui est confiée. Il doit en effet s'acquitter de tâches diverses et variées telles que :

- l'organisation des enseignements, définition du planning annuel et suivi de son déroulement avec l'aide du secrétariat IUP (Mme Sitbon).
- l'organisation du contrôle des connaissances et jurys associés.
- la coordination entre les intervenants de l'équipe pédagogique constituée d'une quarantaine d'enseignants permanents ou temporaires (moniteurs, ATER), industriels ou vacataires.
- l'organisation de réunions bilans en présence des étudiants et de l'équipe pédagogique afin d'améliorer son fonctionnement.
- la participation au conseil de perfectionnement.
- la définition de la nouvelle maquette et syllabus lors de nouvelles campagnes d'habilitation.

Au delà de la Licence, je m'implique dans le fonctionnement global de l'IUP. Ainsi, chaque année, je participe au recrutement des étudiants en première et seconde année IUP. A chaque campagne d'habilitation, je participe aux réflexions portant sur la refonte des programmes de traitement des signaux et images, reconnaissance des formes et robotique ; thématiques qui sont communes aux quatre années de l'IUP.

2. Maîtrise EEA, parcours Traitement De l'Information (2004-...) :
Responsabilité de modules. Créée en 1999 par M.Courdesses (PR UPS), cette formation a un double objectif : (1) donner aux étudiants

une maîtrise des outils et techniques de l'acquisition, du traitement et de l'analyse de l'information, que ce soit sous la forme de signaux ou d'images, (2) sensibiliser les étudiants à leur utilisation pratique dans un grand nombre d'applications. Dès 1999, j'ai participé à la définition et à la mise en place des enseignements relatifs au module image (responsable : G.Flouzat, PR UPS). J'avais alors la responsabilité des travaux pratiques de ce module.

L'évolution de son programme d'une part, la semestrialisation d'autre part, nous ont amenés à refondre les enseignements de cette formation lors du renouvellement de son habilitation en 2004. Cette Maîtrise, transformée à cette occasion en Mastère Information Signal Image et Instrumentation (IS2I), totalise actuellement 13 Unités d'Enseignement parmi lesquelles deux UEs d'imagerie numérique sous ma responsabilité depuis 2004. A ce titre, mon action porte notamment sur l'organisation, le planning, le contrôle des connaissances associés à ces enseignements, ainsi que la définition de leur contenu. Le volume horaire est de 48 h par module.

Fonctions électives et nominatives

Commissions de spécialistes - J'ai participé aux commissions de spécialistes suivantes :

- Membre suppléant de la commission de spécialistes section 61 de l'Université Paul Sabatier (1999-2001 et 2001-2004)
- Membre titulaire et du bureau de cette même commission (2004-2007)
- Membre suppléant de la commission de spécialistes section 61 de l'Université de Perpignan (2004-2007)

Commission Enseignement-Recherche - Durant la période 2003-2005, j'ai participé à la commission Enseignement-Recherche (CER) du LAAS-CNRS. Cette commission, dont les membres sont nommés par la direction du laboratoire, a été mise en place en février 2003. La CER a pour mission d'aborder les problèmes spécifiques des enseignants-chercheurs (réforme LMD, délégations au CNRS, dossiers PEDR, définition des profils de postes enseignant-chercheurs,...) et de proposer des moyens d'action pour accroître la synergie enseignement-recherche. Mon mandat s'est achevé en décembre 2005 avec le renouvellement de cette commission par la direction.

Conseil de perfectionnement de l'IUP SI - Depuis 1999, je participe au conseil de perfectionnement de l'IUP en tant que membre nommé

par la direction de l'IUP SI. Ce conseil, composé à parité d'universitaires et d'industriels, se réunit régulièrement afin de définir la politique générale de l'IUP. Cette instance délivre également le diplôme d'ingénieur-maître au niveau bac+4 (Maîtrise SI), sur proposition d'un jury d'enseignants.

Quatrième partie

Les cinq publications jugées essentielles

Available online at www.sciencedirect.com
ScienceDirect
Image and Vision Computing 25 (2007) 1341–1351
www.elsevier.com/locate/imavis

A visual landmark framework for mobile robot navigation

J.B. Hayet *, F. Lerasle, M. Devy

LAAS-CNRS, 7 avenue Colonel Roche, 31077 Toulouse Cedex, France

Received 19 August 2005; received in revised form 22 August 2006; accepted 23 August 2006

Abstract

This article describes visual functions dedicated to the extraction and recognition of visual landmarks, here planar quadrangles detected by a single camera. Landmarks are extracted among edge segments through a relaxation scheme, used to apply geometrical, topological and appearance constraints on sets of segments. Once extracted, such a landmark is characterized by invariant attributes so that recognition is made possible from a large range of viewpoints.

Landmarks are represented by an icon which is built using the homography between the current viewpoint and a reference shape (a square). When detected again, the landmark is recognized by using a distance between icons. We propose a comparison of several of these metrics and an evaluation on actual and synthetic images that shows the validity of our approach. Results issued from experiments of a mobile robot navigating in an indoor environment are finally presented.

Keywords: Visual landmarks; Visual navigation; Object recognition

1. Introduction

Vision has become a major element in mobile robot navigation and many strategies relying on images have already been proposed, based on environment representation either by image databases [10] or by visual landmarks. Classically, the latter are detected by the robot, mapped into the environment representation and recognized during the execution of a navigation task. In general, the robot's position estimate is computed mainly from the integration of outputs of odometers, which tends to accumulate small displacement errors and produces drift. When recognized, visual landmarks allow to make this drift vanish, so they play a key role in making navigation systems efficient.

The work presented here aims to be part of a navigation strategy relying on "natural" visual landmarks, i.e., salient objects a mobile robot detects/recognizes and from which it can either simply localize itself (if the map of the environment is known) or incrementally build a metric map integrating perceptual data and position estimation, according to the simultaneous localization and mapping (SLAM) paradigm [17]. Our robot has several localization modalities, based either on laser segments learnt using a laser range finder and on visual landmarks detected from a single B&W camera: this paper is mainly devoted to the visual modality. The reader could refer to [1] for a description of these modalities.

Numerous techniques have been proposed to model landmarks for navigating in indoor environments. They all rely on two assumptions: (1) landmarks have to be easily detected in the image signal and (2) they can be locally characterized to distinguish them from others. In that scope, landmark-based navigation research has started by using remarkable characteristics of office-like environments (3D room corners, lights, etc.) [3,11,9], or collections of simple edge segments [16]. Point sets can also serve as landmarks when combined to define projective invariants [2].

Most recent work make use of points to define landmarks [4,15], taking advantage of new, powerful interest point detection and characterization algorithms such as SIFT, which makes landmark-point recognition much easier [5].

* Corresponding author. Tel.: +32 436 62627.
E-mail address: Jean-Bernard.Hayet@ensta.org (J.B. Hayet).

0262-8856/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2006.08.006

In our work, planar quadrangular objects (posters, doors, cupboards, etc.) are selected as landmarks, as they are one of the basic structures man-made environments are made of. Among research works similar to ours, we can quote [12], where the authors take advantage from genetic algorithms techniques to recognize 2D landmarks.

The paper is organized as follows. Section 2 details the landmarks detection process, with results on images acquired during robot navigation. Section 3 presents the landmarks recognition process; an evaluation of our recognition method, with respect to different acquisition criteria, proves its robustness. Navigation experiments are presented in Section 4. Finally, Section 5 sums up our approach and opens a discussion for our future work.

2. Landmarks detection

The landmark extraction is focused on planar, mostly quadrangular objects, e.g., doors, windows, posters, cupboards, etc. A natural way of extracting quadrilaterals relies upon perceptual grouping on edge segments.

2.1. Overview of the method

Let a set of n_L edge segments set be $\mathscr{L} = \{l_i\}$, $1 \leqslant i \leqslant n_L$. A naive approach to test all possible 4-uples inside $\mathscr{L}$ does not make sense, as illustrated in Fig. 2.

To reduce the problem complexity, we propose a two-step algorithm: first, mapping $\mathscr{L}$ to $\mathscr{L} \cup \{\emptyset\}$ so that each segment is matched with at most one segment; second, associating pairs of matched segments to form quadrangles. The whole process is described in Fig. 1.

2.1.1. Extracting edge segments

The output of a Canny-Deriche edge detector is first thinned and chained. The resulting edge chains are then recursively segmented to produce the set $\mathscr{L}$ of line segments as illustrated in Fig. 2. Before the matching process starts, small segments are filtered, altogether with segments that may correspond to repetitive patterns. Typically, segments corresponding to the floor tiling (as in the central image of Fig. 5) are found by an accumulator technique and are eliminated.

2.1.2. Generating segment matches

An initial set of matches is generated by looking for couples $(l_k, l_l)_{k\neq 1}$ for which a similarity measure s_{kl} is above a given level. Indices k and l are associated to individual segments. This measure combines several cues, as explained hereafter, so that segments corresponding to opposite sides of quadrangles have high values of s_{kl}.

Moreover, a set of geometric constraints on segment pairs denoted by Q^1_{klmn} is used in a first relaxation scheme to validate pairs belonging to quadrangles, i.e., to generate a set of coherent potential landmarks. Again, indices k, l, m, n represent individual segments.

2.1.3. Generating potential quadrangular landmarks

With constraints on pairs of detected quadrangles, a second relaxation process selects only the more consistent four-segment sets corresponding to landmarks; these constraints are denoted by Q^2_{klmn}. Three-segment sets are useful as they may correspond to occluded landmarks or doors, so a simple heuristic is used to combine two-segment sets rejected from the second relaxation process with single segments rejected from the first one by using constraints T^2_{klm}. All these constraints, specified in Section 2.3 are applied through a relaxation scheme depicted hereafter.

2.2. Relaxation scheme

Given two sets S_1 (n_1 elements) and S_2 (n_2 elements), the principle of relaxation is to iteratively make all the

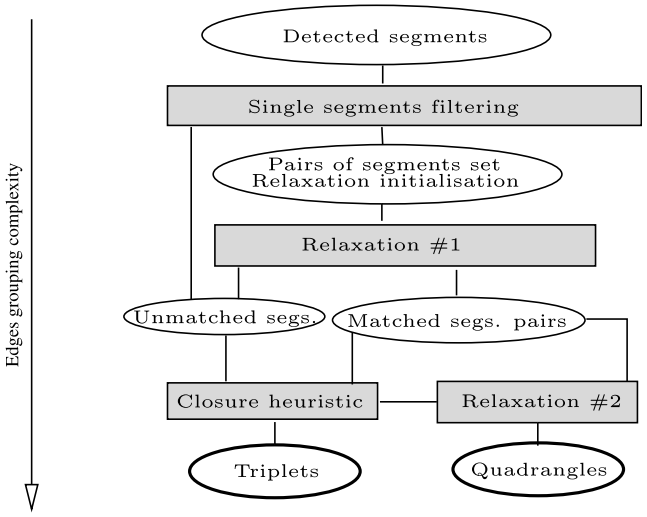


Fig. 1. The landmark detection scheme.

Fig. 2. Segments in a typical indoor scene.

probabilities p_{kl} of associations between items k (index for an element of S_1) and l (index for an element of S_2) evolve towards 1 or 0, i.e., towards unambiguous match or mismatch. Let A be the $n_1 \times n_2$ matrix such as $A_{kl} = p_{kl}$.

We define the variety $\mathscr{A} = \{A \in M_{n_1 \times n_2} \mid \forall (k,l) A_{kl} \geqslant 0 \text{ and } \forall k \sum_l A_{kl} = 1\}$.

The relaxation steps maximize iteratively a global consistency score using gradient ascent in $\mathscr{A}$. In our specific case, for each relaxation process $i \in \{1,2\}$, we maximize a score $G^i(A)$:

$$G^i(A) = \sum_{klmn} Q^i_{klmn} A_{kl} A_{mn}.$$

The terms Q^1_{klmn} (resp. Q^2_{klmn}) represent a compatibility degree between pairs of segment pairs (resp. quadrangles) (k,l) and (m,n). It is derived from constraints detailed in Section 2.3.

The gradient step $\alpha^{(p)}$ at iteration p is adaptive and defined by $\alpha^{(p)} = \text{argmin}_\alpha\, G^i(A^{(p)} - \alpha \nabla G^{i,(p)})$. Regarding initialization, a priori probabilities are computed from similarity measures s_{kl} only. If the measure s_{kl} is below a threshold $s_{\min}$, $p^{(0)}_{kl}$ is set to 0, otherwise it is estimated by:

$$p^{(0)}_{kl} = \frac{s_{kl}}{\sum_{s_{kn} > s_{\min}} s_{kn}}. \quad (1)$$

The next section describes the different criteria and constraints we use in the relaxation schemes.

2.3. Comparing sets of segments

In this section, we make the way we use sets of segments more explicit. We first describe the similarity measure s_{kl} between two segments l_k and l_l used to initialize probabilities p_{kl}. Then, we give details on the constraints Q^1_{klmn} between pairs of segments, and Q^2_{klmn} between quadrangles which are used in the two relaxation schemes.

2.3.1. Segment similarity

The measure s_{kl} is defined by a weighted sum of the following geometric and luminance cues:

- segments length ratio $\frac{1}{2}(\frac{|l_l|}{|l_k|} + \frac{|l_k|}{|l_l|})$ in Fig. 3,
- angular difference $|\theta_{l_k} - \theta_{l_l}|$ in Fig. 3,
- a shape criteria giving favour to square-like shapes $\frac{1}{2}(\frac{|l_l|+|l_k|}{h_{kl}+h_{lk}} + \frac{h_{kl}+h_{lk}}{|l_l|+|l_k|})$ where h_{kl} represents the distance defined in Fig. 3,
- the overlapping rate between l_k and l_l,

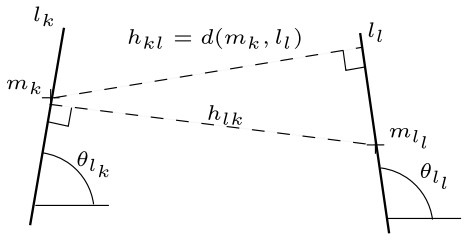


Fig. 3. Conventions for segment matching.

- presence of a third segment in the neighbourhood that forms a convex three-segments set with the given pair. Segments pairs (l_k, l_l) without at least a third segment l_m are discarded for the next.

As far as luminance criteria are concerned, an average grey-level profile is computed in the direction orthogonal to each segment, so that an association (l_k, l_l) is characterized by the zero normalized cross correlation (ZNCC) score between the two segments profiles. In fact, we assume here that the intensity in the background is uniform around a trustworthy quadrangle while its two opposite insides are supposed to include quite similar texture.

2.3.2. Second degree constraints

Here, uniqueness and convexity of potential matches among segments pairs are checked. Uniqueness constraint allows to reduce the relaxation algorithm complexity and enforces the assumption that landmarks are supposed to be locally unique. Convexity rule says that two segments pairs, correspond to opposite sides of two trustworthy quadrangles which must verify rules of full inclusion or no intersection as shown in the left part of Fig. 4.

From the constraints Q^1_{klmn} described above, the first relaxation outputs a set of segments pairs. The next step is to match two segment pairs delimiting trustworthy quadrangles. Indexes k, l, m and n refer now to segments *pairs*.

2.3.3. Third and fourth degree constraints

The fourth degree constraints Q^2_{klmn} ensue from accepted configurations for two quadrangles which are shown in the right part of Fig. 4 and are applied throughout the second relaxation scheme.

From the previous steps, it is possible to extract 3-segment sets that can be helpful in robot navigation. These sets involve an unmatched segment pair (k,l) coming from relaxation #2 and an unmatched segment m coming from relaxation #1. The selection of these potential landmarks is based on uniqueness, on the resulting shape convexity and on vicinity relationships (constraints T_{klm}).

2.4. Detection results

Experiments have been performed on a large database of about 300 images acquired from our robot navigating either in a corridor network or in cluttered open areas. The robot is a Nomadic XR4000, equipped with a SICK laser range finder and a CCD camera mounted on a pan-tilt platform.

Fig. 5 shows examples of landmarks detected in an open cluttered environment. We note that both quadrangles and three-segment sets are extracted.

During the environment exploration, the robot executes two operations: (1) a SICK laser map is built by a classical SLAM procedure and (2) visual landmarks are detected and combined with the laser segments. The resulted map is represented in Fig. 6, with all laser segments and all

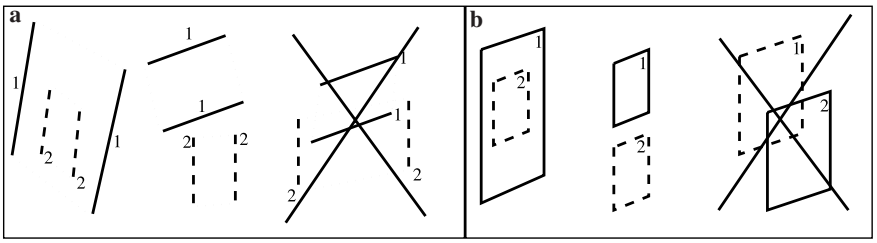


Fig. 4. Examples of accepted configurations for two segments pairs or two quadrangles.

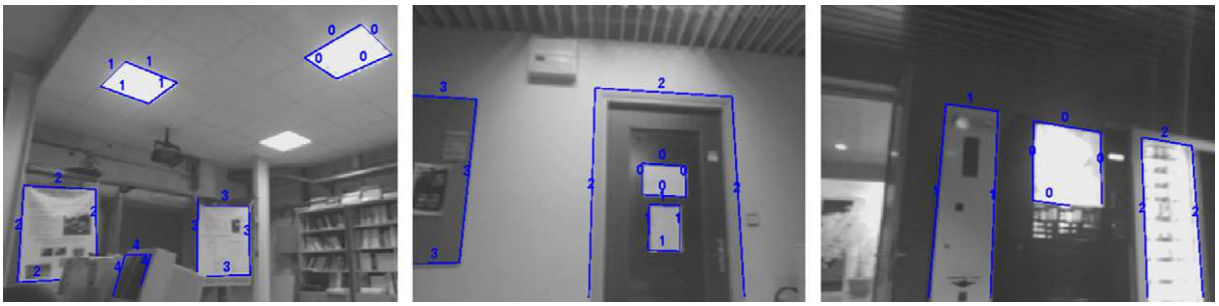


Fig. 5. Examples of landmarks detection: the numbers on the segments indicate the final tag associated to the detected landmark.

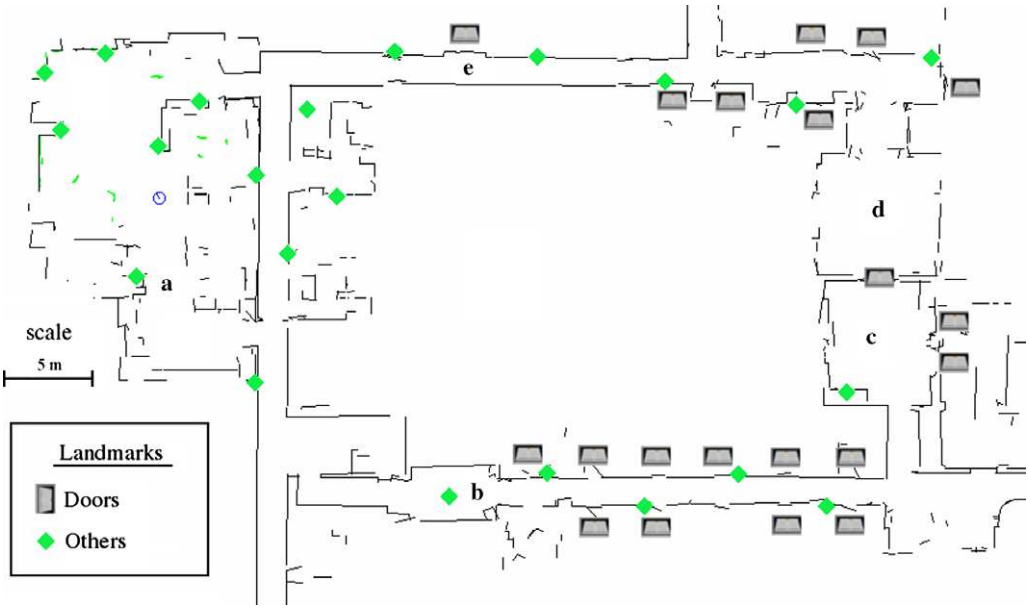


Fig. 6. Landmarks detection in office environment.

detected landmarks: windows or posters in green (lateral walls or ceiling), doors by a grey icon. Their associated locations on the walls are triangulated from their perspective views and the planes defined by the laser segments assuming the multisensory system is fully calibrated. For every detected landmark, a visibility map (not shown here) is statically computed according to the environment model by an analytical method.

Detection rates are computed over the database of images taken by the robot. In this database, all quadrangular objects have been identified by a human operator; the landmarks detection module extracts 88% of existing landmarks, without any false detection.

During this environment exploration step, the robot could stop to perform both detection and recognition processes, so that only the representation of new discovered landmarks is learnt. Only quadrangular objects which are successfully detected from different view points (Section 3.3) are considered as landmarks in the environment model. Later, when the robot navigates using the set of learnt landmarks, it must be able to achieve these tasks dynamically.

3. Landmarks recognition

Once a landmark has been detected, an appearance model is built so that it can be recognized from different viewpoints. In Section 3.1, we describe the landmark representation: boundaries of a detected landmark allow to rectify the observed pattern; and such a mapping provides an *invariant* representation under scale and perspective changes. We call it "icon".

In Section 3.2, we propose distances to compare icons and perform recognition. Section 3.3 thereafter describes the landmark model. Based on this model, a confidence factor on the recognition process is proposed in Section 3.4. In Section 3.5, a correlation-based method is compared with an approach based on interest points extracted from icons.

3.1. Landmarks iconification

Let us consider, (1) an extracted quadrangular landmark Q from an image I and (2) a fixed-size reference square S. The two shapes are related by a homography H_{SQ} that maps points from S to Q.

By using H_{SQ}, a new small-sized image I' is built from the image I by averaging pixels from I into pixels in I' (see Fig. 7). The computation of H_{SQ} is straightforward as four point correspondences are available [14].

Averaging is performed in order to avoid too much information compression in the low-scale front view I': the grey level value of a pixel (a,b) in image I' is determined by taking into account all pixels in image I belonging to a certain neighbourhood of $H_{SQ}(a,b,l)^{\mathrm{T}}$, i.e., its image in I. This neighbourhood is computed by approximating the image of a pixel square with simple heuristics (see Fig. 8).

The icon I' is processed by the Harris operator to get a set of n interest points $\{X_i\}_{1\leqslant i\leqslant n}$ and a local descriptor [13] in $\mathbb{R}^7$, based on Gaussian derivatives, is associated to every interest point.

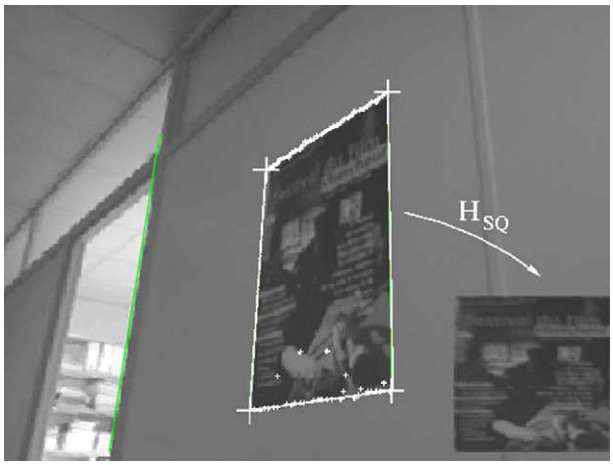


Fig. 7. Model construction: quadrilaterals are transformed into icons by the mean of H_{SQ}.

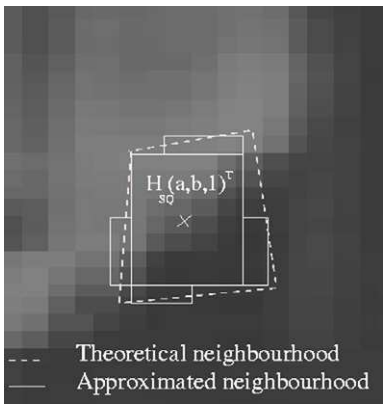


Fig. 8. Approximated averaging for iconification: zoom of the neighbourhood of the image of an icon pixel (a,b).

3.2. Metric on icons

To perform the recognition between a set of learnt landmarks noted $\{C_l\}_{1\leqslant l\leqslant N}$ and a detected landmark $\mathcal{Q}$, metrics on icons are defined.

3.2.1. A correlation-based distance

The centered and normalized correlation score $\mathscr{C}$ provides a distance which is theoretically invariant to overall light changes.

To be less sensitive to local variations or occlusions, the icons $\mathcal{Q}$ (from the new landmark) and C_l (from the reference landmark l) are divided into 5×5 buckets. Then, we define a robust correlation score $\mathscr{C}^r$ between two icons by using separated correlations $\mathscr{C}_{ij}(\mathcal{Q}, C_l)$ between buckets i and j, and by choosing the k^{th} greatest correlation score between buckets. The number k is expressed as a ratio r of admissible outliers among all the buckets. It allows to ignore the most important local differences between $\mathcal{Q}$ and C_l. From this new score, we derive the distance:

$$\mathscr{C}^r(\mathcal{Q}, C_l) = 1 - k^{\text{th}}_{1\leqslant i,j\leqslant 5}\mathscr{C}_{ij}(\mathcal{Q}, C_l).$$

3.2.2. A local features-based distance

Many popular appearance-based methods for object recognition are based on interest points matched thanks to their local descriptors [13,15]; these local features are remarkably stable under moderate rotation or light changes. We propose to use the partial Hausdorff distance [8] to compare sets of interest points $\{X_i\}_{1\leqslant i\leqslant n}$ extracted from icons.

Let be two sets of points $S_l = \{X_i^l\}_{1\leqslant i\leqslant n_l}$ associated with a known landmark C_l, and $S = \{X_i\}_{1\leqslant j\leqslant n}$, extracted from a new landmark $\mathcal{Q}$. To handle outliers, the Hausdorff distance between S_l and S is modified in the same way as $\mathscr{C}^r$, i.e., by considering only a fraction r of all the points, $k = r\,\min(n_l, n)$:

$$\begin{cases} d_h^r(S_l, S) = \max(h^r(S_l, S), h^r(S, S_l)), \\ h^r(S_l, S) = k^{\text{th}}_{1\leqslant i\leqslant n_l}\min_{1\leqslant j\leqslant n} d(X_i^l, X_j). \end{cases}$$

A threshold τ_l on the distance is set to recognize landmarks $\mathcal{Q}$ as instances of known landmarks C_l, as it will be described in Section 3.3.2. An interpretation of this distance is that an object is recognized provided that for at least k points of the second set, a similar point can be found in the first set, and reciprocally.

The partial Hausdorff distance between two sets of points depends on the *local distance* d between points. We could simply use the Euclidean distance, but we would lose explicit local photographetric information. In order to take into account both spatial and photographetric similarities between points, we define a local distance noted d_p:

$$d(a, b) = d_v(a, b)\|a - b\|,$$

where $d_v(a,b)$ is the Mahalanobis distance between the descriptor vectors at points a and b. The Hausdorff distance based on d is denoted by $\mathscr{H}^r$.

3.3. Building appearance models

For each landmark C_l, a model is built from a set of N_l representative images I_i at several viewpoints (typically $N_l = 50$), from which iconified views I'_i are extracted.

3.3.1. Reducing landmark representation

A principal component analysis is first performed on the set of raw icons. We keep only three icons, denoted respectively by $\mathcal{Q}_l^1$, $\mathcal{Q}_l^2$, $\mathcal{Q}_l^3$. The first one $\mathcal{Q}_l^1$ corresponds to the mean icon of I'_i, $1 \leqslant i \leqslant N_l$, whereas $\mathcal{Q}_l^2$ and $\mathcal{Q}_l^3$ correspond to the more significant modes on this icon set.

For distance $\mathscr{H}^r$, such a process is followed by the extraction of Harris points and their characteristics in the I'_i icons closest to the selected eigenvectors.

3.3.2. Determining recognition thresholds

During the recognition step, a detected landmark is compared to each known landmark C_l, using a recognition threshold τ_l specific to it. During the modelling step, an optimal threshold is computed for each landmark C_l by computing distances ($\mathscr{C}^r$ or $\mathscr{H}^r$) between extracted icons for this landmark, with either the C_l model or all the other models noted $\neg C_l$.

The distance distributions on representative sets of icons from C_l and $\neg C_l$ give us a good approximation of the probability densities on the distances, given the knowledge of C_l or $\neg C_l$. To specify an optimal threshold τ_l, we minimize:

$$S(\tau_l) = \lambda \underbrace{\int_0^{\tau_l} p(d|\neg C_l)\mathrm{d}d}_{\neg S_l(\tau_l)} + \mu \underbrace{\int_{\tau_l}^{+\infty} p(d|C_l)\mathrm{d}d}_{S_l(\tau_l)},$$

with λ and μ being two weights for respectively false positive and false negative, noted $\neg S_l(\tau_l)$ and $S_l(\tau_l)$. The choice $\mu = \frac{1}{6}\lambda$ allows to give more importance to false positives than to false negatives. The security in the robot navigation being critical, the recognition of a landmark in a bad position cannot be accepted, i.e., false positive are more important to be avoided.

3.3.3. Validation gates

For every landmark C_l, the modelling step ends with a verification of two criteria: (1) C_l must be salient enough, and (2) the N_l images from which the C_l appearance model has been generated, must give a good approximation of all possible viewpoints on C_l.

The *saliency* criterion is verified from the *covariance* of the icons I', and from the number of stable extracted interest points. The *visibility* criterion indicates how far from each other are the extreme positions at which the landmark has been detected during this learning step. For all couples $(i,j) \in [1, N_l]^2$, an inter-image homography H^{ij} maps corresponding vertices of the landmark in images I_i and I_j. Let us consider the normalized homography $\hat{\mathrm{H}}^{ij}$, such as $\hat{H}^{ij}_{33} = 1$, and where image coordinates have been centered and normalized. Then, we define a visibility confidence as: $v_c = \max_{ij}\|\hat{\mathrm{H}}^{ij} - I_{33}\|$.

I_{33} is the 3×3 identity matrix. The greater is v_c, the more extended is the area on which the landmark has been perceived during the learning step. The value v_c is clearly correlated to the planarity: planar landmarks are recognized in a larger area and under greater camera parameters changes than non-planar ones.

3.4. Confidence in the recognition result

The recognition task requires to index and compare detected landmarks. For a set of N modelled landmarks $\{C_l\}_{i \leqslant z \leqslant N}$ and a detected landmark $\mathcal{Q}$, let us note $\mathscr{D}_l = \mathscr{D}(\mathcal{Q}, C_l)$, the distance between $\mathcal{Q}$ and each class C_l ($\mathscr{D}$ being either $\mathscr{C}^r$ or $\mathscr{H}^r$). The probability $P(C_l|\mathcal{Q})$ of labeling $\mathcal{Q}$ to C_l, is defined by:

$$\begin{cases} P(C_\emptyset|\mathcal{Q}) = 1 \text{ and } \forall l\ P(C_l|\mathcal{Q}) = 0 \text{ when } \forall l\ \mathscr{D}_l > \tau_l \\ P(C_m|\mathcal{Q}) = 1 \text{ and } \forall l \neq m\ P(C_l|\mathcal{Q}) = 0 \text{ when } \exists! m \mathscr{D}_m < \tau_l \\ P(C_\emptyset|\mathcal{Q}) = 0 \text{ and } \forall l\ P(C_l|\mathcal{Q}) = \frac{h(\tau_l - \mathscr{D}_l)}{\sum_p h(\tau_l - \mathscr{D}_p)} \text{ otherwise} \end{cases}$$

where $C_\emptyset$ refers to the empty class and h the Heaviside function: $h(x) = 1$ if $x > 0$, 0 otherwise. This allows us to use the entropy-based measure:

$$m(\mathcal{Q}, \{C_l\}) = 1 + \frac{1}{N+1}\sum_j P(C_j|\mathcal{Q}) \log P(C_j|\mathcal{Q}).$$

3.5. Recognition evaluation

An important issue for our recognition process, is the way the algorithm behaves with light effects, scale/perspective changes and bad segmentation from the detection step. Other questions are related to the discriminating power of proposed distances. To investigate this robustness problem, a large test image database has been constituted both by:

(1) Two hundred and seventy real images of different landmarks acquired while the robot wandered around the lab (see Fig. 9) represented by the map of Fig. 6.
(2) synthetic images of 300 movie posters with different light, scale/perspective conditions and occlusions, these modalities remaining quite difficult to perform and quantify in real conditions (see Fig. 10).

3.5.1. Discriminating power

Let us consider probability densities computed from the distribution of distances between a given landmark and other ones from the database of real images. A poster found in this database has been selected and learnt as a landmark, and Fig. 11 now represents distributions of distance values obtained (a) for the objects corresponding which are instances of this landmark (class C_l) and (b) for objects that are not (class $\neg C_l$). This distribution can be approximated by a Gaussian function, which center and variance depends on the Hausdorff fraction and on sets cardinals.

The overlapping surface under the two curves are relatively small for the two distances that have been investigated. By following the process described in Section 3.3.3, we have rates of false positive around 1%, whereas false negative where about 30%, which reflects the high level of disturbances we put on synthetic data sets.

3.5.2. Behavior under viewpoint changes

The graphs in the left part of Fig. 12 represent the evolution of the ratio $\frac{\text{distances}}{\text{threshold}}$ for distances and $\mathscr{H}^r$ and $\mathscr{C}^r$

Fig. 9. Examples of real images with variable scales, occlusions, brightness variations or specular reflexions.

Fig. 10. Examples of synthetic images with variable scales, occlusions, brightness variations or specular reflexions. Movie posters were used as the basic texture.

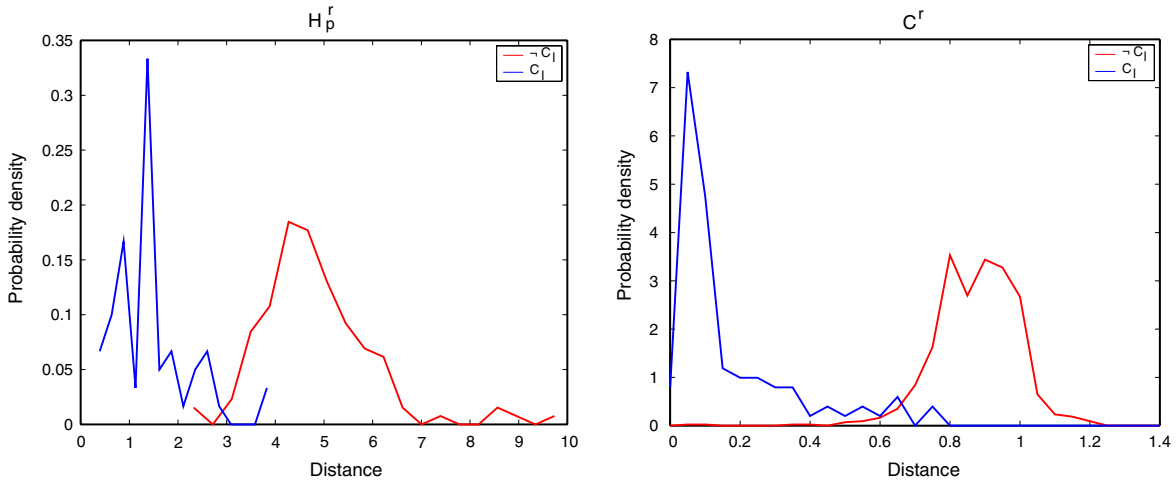


Fig. 11. Discriminating power: distribution of distances on classes C_l and $\neg C_l$ for $\mathscr{H}^r$ (left) and $\mathscr{C}^r$ (right).

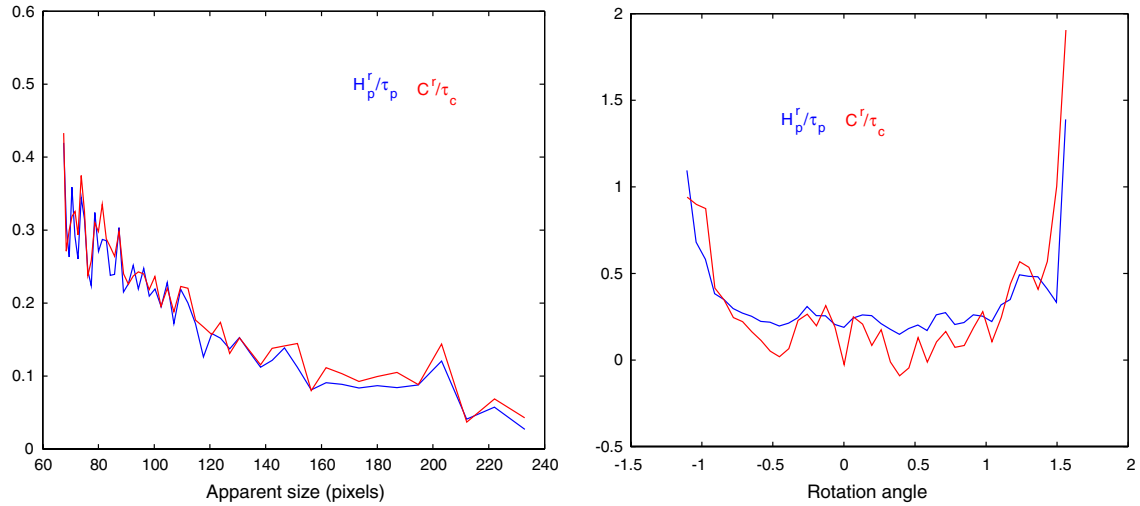


Fig. 12. Variations of the ratio $\frac{\text{distance}}{\text{thershold}}$ under scale changes (left) and rotations around the vertical axis (right) for distances $\mathscr{H}^r$ and $\mathscr{C}^r$.

under scale change. This ratio has to remain below 1 to ensure recognition. Even for a scale factor about three, values for both of the compared distances remain small w.r.t. their respective thresholds. However, as expected, results are degrading fast as soon as the apparent size of the extracted pattern is below the size of the square used for the iconic representation.

As far as perspective distortions are concerned, the evolution of the ratio $\frac{\text{distances}}{\text{threshold}}$ have been studied for distances $\mathscr{C}^r$ and $\mathscr{H}^r$ by performing a planar rotation in the horizontal plane of a landmark. Results on the right part of Fig. 12 show that the combination of invariants vectors and interest points is a powerful tool to achieve recognition of planar objects, as distances remain reliable up to $\pm 75°$ from the normal to the landmark plane, which is reasonable.

3.5.3. Behavior under light effects and occlusions

The left graph in Fig. 13 shows that it is possible to have good recognition results for the two distances until local or global light saturations appear in the image.

Moreover, as it can be seen on the right part of Fig. 13, the representation is also robust to partial occlusions, which occurs for partially detected landmarks, that compose the majority of detected landmarks in indoor environment. With the distances $\mathscr{H}^r$ and $\mathscr{C}^r$, occlusions of the landmark up to 46% and 56% of its area do not prevent the landmark from being recognized.

3.6. Discussion: comparing the two metrics

We have compared two different representations and associated metrics by applying tests w.r.t the main sources of image noise and variations. Both of the metrics have quite satisfactory results on ambient brightness variations, scale or perspective changes, which makes our concept of quadrangles-landmarks a powerful tool for modelling environments. The $\mathscr{C}^r$ metric gives slightly better recognition results on all these tests, but it is limited by the size of data that have to be stored, i.e., all the icons have to be stored.

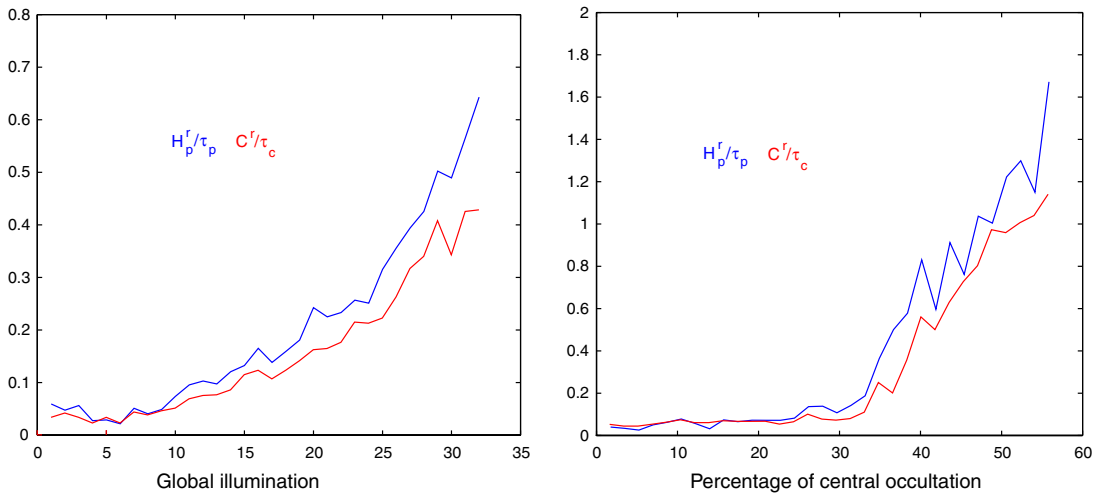


Fig. 13. Variations of the ratio $\frac{\text{distance}}{\text{thershold}}$ under light changes (left) and central occlusions (right) for distances $\mathscr{H}^r$ and $\mathscr{C}^r$.

That is why in practice $\mathscr{H}^r$ is preferred for our experimental work: this distance is compact and gives fairly good recognition results.

4. Application to robot navigation

Our landmark detection and recognition scheme has been integrated as a visual localization module in our Diligent Nomadic XR4000 robot, shown in Fig. 14. Section 4.1 describes the landmark localization with calibrated vision, and experiments showing our robot navigating in indoor environments are presented. Then, we introduce an extension we developed to handle unknown camera parameters.

4.1. Localization with a calibrated camera

Let us assume that our vision system is fully calibrated and that a 3D model of the quadrangle $\mathscr{Q}$ has been determined i.e. its four corners noted $\{P_i^n\}_{i=1,\ldots,4}$ in the poster frame are *a priori* known. The landmark localization in the camera frame, i.e., the displacement $[R^{cn}, T^{cn}]$, is based on the decomposition of the homography H^m relating four matches of image points p_i and model points P_i^n. This matrix H^m can be interpreted in terms of a displacement between the poster frame and the camera frame [14]:

$$[r_2^{cn}, r_3^{cn}, T^{cn}] = \lambda K^{-1}[h_1, h_2, h_3] \quad (2)$$

$[r_1^{cn}, r_2^{cn}, r_3^{cn}]$ (resp. $[h_1, h_2, h_3]$) are columns of R^{cn} (resp. H^m), $[t_x^{cn}, t_y^{cn}, t_z^{cn}]$ are the components of T^{cn}, K the intrinsic parameters matrix and λ the scale factor.

Let us recall the robot is considered as a complete system, equipped not only by a camera, but also by a laser range finder and by odometry. A localization module is associated to every sensor: all computed positions are logically fused by a dedicated position manager module [1]. The localization strategy is based on a loose coupling of these modalities. During an off line statistical analysis, the robot learns the better localization modality it must execute to locate itself in every area in the environment, according to their intrinsic performances and to local configurations of learnt landmarks or features. For example, the robot learns by itself, that: (1) in open space, it is relevant to fuse localizations computed by all modalities, even if they are computed at various frequencies, (2) in a given place, due to an uneven area on the ground, the odometry modality gives an important bias, (3) in a long corridor, vision modality is better than laser modality.

Figs. 15 and 16 illustrates navigation experiments in a 25 m long corridor (annoted (b) in Fig. 6) where laser localization is known to be inefficient as there is no identifiable beacon in the direction orthogonal to the corridor. In each left sub-figure, the blue trace corresponds to current odometry positions; without another modality, the robot would clearly bump against the left wall of the corridor. The red trace gives the current corrected position from the vision method, executed on four previously learnt posters annotated #0 to #3 (red color) on the laser map. Sub-figures show the robot, respectively at corridor entry, at two positions close to posters and finally at corridor exit. Each upper right image shows the current robot perception while the bottom right image shows the robot in its environment. The robot perception is ensured by the camera mounted on

Fig. 14. The XR4000 "DILIGENT" robot we used in the experiments is equipped with a laser and BW cameras.

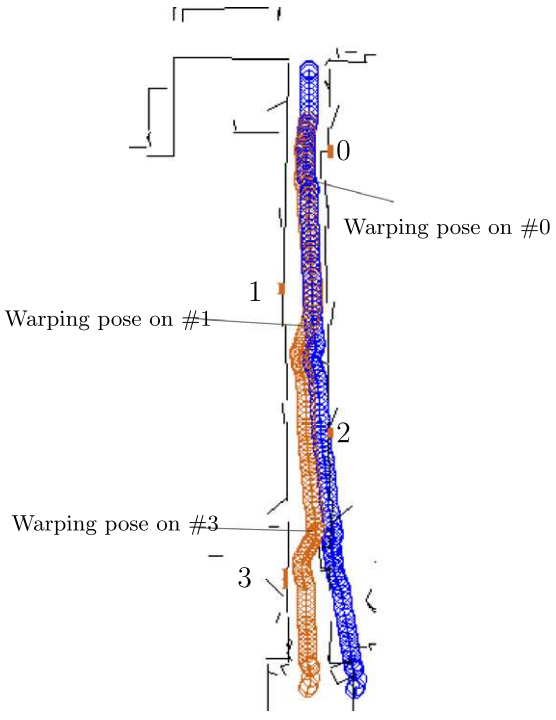


Fig. 15. Robot localization: recognizing known landmarks (marked $0, 1, 2, 3$) allow to correct the robot's position.

Fig. 16. Robot localization: external view (bottom) and robot view (up).

a pan and tilt platform; in every place, the camera is pointed towards the best landmark, selected with respect to its visibility area and saliency coefficients estimated during the learning step. The number of positions corrections performed since the robot enters the corridor is displayed in the superimposed box on each sub-figure. The robot's position is corrected 13 times during its navigation. In such a corridor, the robot can be localized in the corridor direction, with an error lower than 20 cm.

4.2. Auto-calibration with quadrangles

An extension of our work deals with active vision, which implies to re-estimate camera intrinsic parameters. We propose to do it online, from several views of a planar quadrangle. It is assumed here that these parameters are constant on these views. Using Eq. (2), we evaluate the image of the absolute conic $\omega = K^{-\mathrm{T}}K^{-1}$, under the simplified form:

$$\omega = \begin{pmatrix} \omega_1 & 0 & \omega_2 \\ 0 & \omega_1 & \omega_3 \\ \omega_2 & \omega_3 & \omega_4 \end{pmatrix}.$$

Let $\Omega = (\omega_1, \omega_2, \omega_3, \omega_4)^{\mathrm{t}}$ be the vector to estimate. Such a parametrization allows to write linear constraints on intrinsic parameters[14]. First, constraints on planar homography deduced from Eq. (2) lead to:

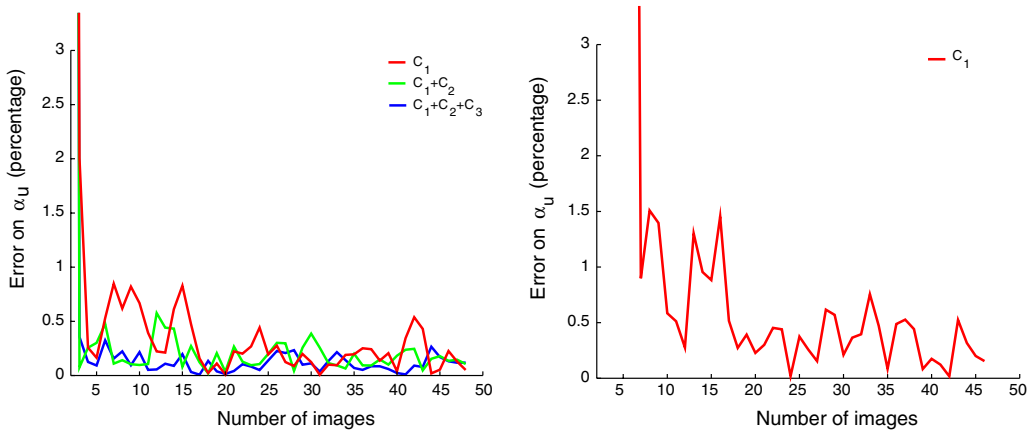


Fig. 17. On-line calibration: errors on α_u constraints for synthetic (left) or real (right) views vs. number of views.

$$\begin{cases} h_1^{\mathrm{T}}\omega h_1 = h_2^{\mathrm{T}}\omega h_2, \\ h_1^{\mathrm{T}}\omega h_2 = 0. \end{cases} \quad (C_1)$$

Second, assuming the roll angle of the camera platform to be neglected, makes the skyline and vertical vanishing point be known. This entails also:

$$(0 \quad 1 \quad 0).\,\omega h_2 = 0. \quad (C_2)$$

In the same way, given Eq. (2), the constraint of planar robot motion can be written as follows:

$$-t_z^{nc}(h_1^{\mathrm{T}}\omega h_1) = h_2^{\mathrm{T}}\omega h_3. \quad (C_3)$$

Combining these three constraints (C_1), (C_2) and (C_3) allows to solve intrinsic parameters K.

Fig. 17 shows calibration results for different constraint combinations. Synthetic experiments (see Fig. 17, left) show that the first constraint (C_1) seems to be sufficient. On the right part of Fig. 17, calibration results for real images are presented. The relative error to ground truth is inferior to 1% which is suitable for active vision purposes. From five to ten views are required to recover intrinsic parameters with a good precision.

5. Conclusion and future works

We present an original framework to use quadrangular visual landmarks for robot navigation in indoor environment. A first contribution concerns a method for extracting quadrangles in open cluttered and corridor-like spaces. These quadrangles can correspond to planar objects (posters, doors, cupboards, etc.). A new representation and associated recognition method for such landmarks is presented. It has been verified that this method remains efficient despite ambient brightness variations or viewing changes.

Navigation experiments have been performed; the extraction of visual landmarks is very efficient, as well as the landmark recognition method. During the environment exploration, about 90% of pertinent landmarks are extracted; then, when the robot goes along a path planned in the environment model, landmarks are actively searched and exploited for the robot localization. When only posters are considered as landmarks, the recognition rate is greater than 97%. Failures are due to unforeseen occlusions or specific ambient brightness variations. Our method proposed to select the thresholds, allows to avoid false positive errors.

Two directions are currently studied regarding our visual landmarks based navigation system. First, visual functions described here are exploited for topological navigation and qualitative localization purpose [7]. Considering ambiguous landmarks (doors, etc.), a Markovian localization [6] will be implemented to handle multi-hypothesis on the robot position. Second, a more tied coupling strategy is studied to improve the explicit robot localization; the land-mark model, will be learnt together with the laser map, using a SLAM approach to build a heterogeneous stochastic map.

References

[1] B.Morisset, M.Ghallab, Synthesis of supervision policies for robust sensory-motor behaviors, in: Proceedings of the International Conference on Intelligent Autonomous Systems (IAS'02), 2002.

[2] A. Branca, E. Stella, A. Distante, Landmark-based navigation using projective invariants, in: Proceedings of the International Symposium on Robotics and Automation (ISRA'00), 2000, pp. 569–574.

[3] H. Choset, K. Nagatani, A. Rizzi, Sensor based planning : using a honing strategy and local map method to implement the generalized Voronoi graph, in: Graph. SPIE Mobile Robotics, 1997.

[4] C.I. Colios, P.E. Trahanias, Landmark identification based on projective and permutation invariant vectors, in: Proceedings of the International Conference on Pattern Recognition (ICPR'00), 2000, pp. 128–131.

[5] P. Elinas, R. Sim, J.J. Little, σrslam: stereo vision slam using the rao-blackwellised particle filter and a novel mixture proposal distribution, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2006.

[6] D. Fox, W. Burgard, S. Thrun, Markov localization for mobile robots in dynamics environments, Journal of Artificial Intelligence Research 11 (1999) 1265–1278.

[7] J.B. Hayet, F. Lerasle, M. Devy, Environment modeling for topological navigation using visual landmarks and range data, in: Proceedings of the International Conference on Robotics and Automation (ICRA'03), 2003.

[8] D.P. Huttenlocher, A. Klanderman, J. Rucklidge, Comparing images using the Hausdorff distance, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 15 (9) (1993).

[9] G. Jang, S. Kim, W. Lee, I. Kweon, Color landmark-based self-localization for indoor mobile robots, in: Proceedings of the International Conference on Robotics and Automation (ICRA'02), 2002, pp. 1037–1042.

[10] J. Santos-Victor, R.Vassallo, H.J. Schneebeli, Topological maps for visual navigation, in: Proceedings of the International Conference on Computer Vision Systems (ICVS'99), 1999, pp. 1799–1803.

[11] F. Launay, A. Ohya, S. Yuta, A corridors lights based navigation system including path definition using a topologically corrected map for indoor mobile robots, in: Proceedings of the International Conference on Robotics and Automation (ICRA'02), 2002, pp. 3918–3923.

[12] M. Mata, J.M. Armingol, A. de la Escalera, M.A. Salichs, A visual landmark recognition system for topological navigation of mobile robots, in: Proceedings of the International Conference on Robotics and Automation (ICRA'01), 2001.

[13] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, in: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR'03), 2003.

[14] R. Hartley, A. Zimmerman, Multiple View Geometry in Computer Vision, Cambridge University Press, Cambridge, 2000.

[15] S. Se, D.G. Lowe, J. Little, Global localization using distinctive visual features, in: Proceedings of the International Conference on Intelligent Robots and Systems (IROS'02), 2002, pp. 226–231.

[16] R. Sim, G. Dudek, Learning visual landmark for pose estimation, in: Proceedings of the International Conference on Robotics and Automation (ICRA'99), 1999, pp. 1972–1978.

[17] S. Thrun, Robotic mapping: a survey, in: G. Lakemeyer, N. Nebel (Eds.), Exploring Artificial Intelligence in the New Millenium, Morgan Kaufman, Los Altos, CA, 2002.

Qualitative Modeling of Indoor Environments from Visual Landmarks and Range Data

Jean-Bernard Hayet Claudia Esteves* Michel Devy Frédéric Lerasle

Laboratoire d'Analyse et d'Architecture des Systèmes - LAAS-CNRS
7, avenue du Colonel Roche, 31077 Toulouse Cedex 04, France
e-mail: {jbhayet, cesteves, michel, lerasle}@laas.fr

Abstract

This article describes the integration in a complete navigation system of an environment modeling method based on a Generalized Voronoi Graph (GVG), relying on laser data, on the one hand, and of a localization method based on monocular vision landmark learning and recognition framework, on the other hand. Such a system is intended to work in structured environments. It is shown that the two corresponding modules — laser GVG construction and visual landmarks learning and recognition — can cooperate to complete each other, as image processing can be enhanced by some structural knowledge about the scene, whereas the GVG is annotated, even as far as its edges are concerned, by qualitative visual information.

1 Introduction

Mobile robot navigation can be considered as the art to overcome the inaccuracy of internal sensors and to take advantage of exteroceptive sensors like cameras, sonars or laser range finders to allow the robot move and act in its environment. Many strategies have already been proposed, some based on explicit localization of the robot with respect to the environment, others only relying on relative localization with respect to some interesting objects, landmarks, perceived by the robot. The topology of this set of landmarks is generally embedded in a graph. The work presented in this paper has been done in the latter framework and designed for a service robot moving in an office environment composed of a network of corridors and open spaces.

We have already presented a preliminary work in [8] using the ultrasonic-based Generalized Voronoi Graph (GVG) representation proposed by H.Choset [2]. This representation is a topological graph describing the paths on which the robot must navigate; in this approach, nodes are associated to "distinctive places", where "distinctiveness" is determined according to the US sensors. In this case, it corresponds to the discontinuities of the GVG edges, i.e. "meet points", associated to intersections between corridors or to crossings (doors) towards open spaces (rooms, hallways...). The graph edges correspond to paths in corridors or in open spaces. To overcome the classical self-localization problem resulting from US data ambiguity, we annotated each node with visual landmarks, planar, quadrangular objects (e.g. doors, windows, posters) that were automatically discovered, learned *around the meet points only*. Landmark intrinsic representations *independent from the viewpoint* were used and were shown in [5] to be stable with respect to illumination, scale changes and small occlusions.

In order to better validate an hypothesis about a node identification and to make the incremental construction more robust, we worked in two directions : (1) change the range sensor from the ultrasonic sensors belt to a *laser range finder (horizontal scanning)* and (2) annotate not only the GVG nodes *but also its edges* to maintain a qualitative position along an edge. Some authors proposed methods to deal with the incremental construction of a GVG representation using laser data. In [10], the GVG was explicitly built, coping with a lot of geometrical situations that made the method slow and unreliable. In [9], an implicit modeling strategy was proposed, using the sensor servoing, namely the task function formalism, to keep on the GVG or to detect meet points. The authors used only laser data, so that this very efficient method could have some problems in very ambiguous situations, like a regular network of corridors.

Qualitative spatial reasoning implies to work on some

*The stay of Claudia Esteves at LAAS-CNRS in France, is funded by the French-Mexican lab in Computer Science.

space notions without using any representation or reasoning method requiring numeric or quantitative descriptions. Qualitative information covers *topology* — connectivity, topological relationships — *orientation*, and *order*. In [3], the available approaches to model *topological* relationships were reviewed, and some new ones were proposed. *Orientation* and *order* relationships are also widely used : in [4] the notion of intrinsic orientation is underlined, i.e. the orientation relative to the robot current position on its trajectory.

Nevertheless, in order to achieve robust navigation, it is desirable to forget the pure "qualitative" notions so that the robot could benefit from all the possible data it could use and combine metric and topological levels of information. As an example, Kuipers [7] introduced the notion of *Spatial Semantic Hierarchy* that included, among the others, two levels for topological and metric information.

The sections 2 and 3 present the modules devoted respectively to the GVG construction from laser data and to the landmarks detection from visual input. In section 4, we introduce the compound environment representation and finally the section 5 sums up this work and opens a discussion for our future works.

2 Building a GVG from laser data

The Generalized Voronoi Graph representation associates the set of points equidistant from at least two obstacles to its edges and *meetpoints* — points equidistant to at least three obstacles — to its nodes. The latter ones are salient features in the environment, distinctive places, such as corridor intersections, crossings to open spaces (rooms or hallways) and corridor ends. The incremental construction of the GVG does not only provide a natural way to capture the topology of the environment free space but also greatly reduces the error accumulation due to odometry by observing the change of local coordinates. The GVG construction consists in going over every possible path in a corridor-based environment, memorizing the path connections in the GVG and learning visual landmarks at the nodes and along the edges, as presented in section 3. Note that the two traditional tasks, exploration and navigation, have no clear boundary here, as they are performed at the same time.

The robot can be controlled to navigate along a GVG edge by keeping equidistant to the two closest obstacles which are mainly the two walls on the corridor. The inputs of this control law are the distance and orientation to the GVG edge computed from the segment information provided from the laser range finder. A prediction and correction steps are merged to obtain a smooth path. Detected laser segments provide a representation (figure 1) that is required either to keep on the GVG ((a) and (f)), to detect a meet point ((b) and (c)) or to detect an obstacle or a dead-end ((d) and (e)).

Because the laser range finder sensor we are using has a limited angle, a meetpoint detection approach by watching for an abrupt change in the direction of the gradients to the two closest obstacles as proposed in [2] becomes unsuitable. In order to detect and move to a meetpoint, our approach relies on the observation of filtered and segmented range data. Two main *events* on the tracked corridor can be identified. One, when two segments belonging to the same wall on the corridor are disconnected by a length superior to a given threshold (*discontinuity*) and the other at the end of a wall (*end*). Such events are closely related to the nature of the other obstacles found on the same scan as they can be produced due to occlusions, open doors or new paths.

A model of the approaching meetpoints mp can be determined according to the configuration of these events and the nature of the found obstacles before the robot actually gets there, typically at about 5 meters from the meetpoint. The underlying hypothesis generation-verification scheme relies on the a priori knowledge of models presented on the figure 1. The following lines sum up our strategy :

0: Search for two major line segments within the segmented data → *wall1* and *wall2*.
1: Access to GVG by gradient ascent. $mp = \emptyset$
2: Main loop.
  while ($remaining\_paths \neq \emptyset$) **do**
    if (*closest*(mp) not reached) **then**
      track *wall1* and *wall2*
      check for *events*
    else
      confirm *closest*(mp)
      update graph and follow exploration
    end if
    update mp
  end while

As an illustration, figure 2 shows the robot getting onto the GVG (1), following it while generating hypothesis (2) and reaching a meet point (3). Note that obstacles inside the corridor are easily detected, so the robot performs avoiding strategy (figure 1,(f)) or adds a dead-end (figure 1,(e)), according to the obstacles relative size and orientation.

3 Visual landmarks detection

With corridor-like environments, extracting vanishing points and skyline is relatively easy. We show

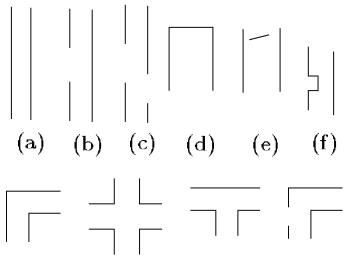


Figure 1: *Corridor and meetpoints configurations*

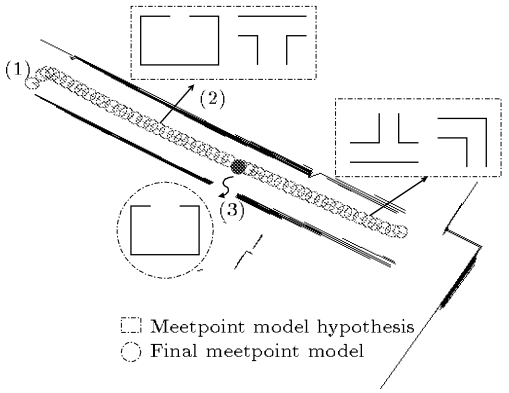


Figure 2: *GVG incremental construction*

how we can take this into account to improve our landmark detection strategy.

3.1 Vision in corridor-based environments

With some simple assumptions about the environment, well-adapted visual functions can be proposed to help the robot navigate. We focused our previous work on the use of rectangular visual landmarks. We suppose for the moment that the camera intrinsic parameters matrix K is known :

$$K = \begin{pmatrix} \alpha_i & 0 & i_0 & 0 \\ 0 & \alpha_j & j_0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

The use of the *vanishing points* information inside the image processing steps seems inevitable in this case. Let i and j be the image coordinates. Let ϕ be the platform tilt angle and θ the horizontal angle between the camera optical axis and the corridor direction. As we illustrate it on figure 3 the robot planar motion constraint and the camera platform movements restricted to pan and tilt motions make the skyline $i = i_s$ and *vertical* vanishing point $p_v = (i_v, j_v, t_v)$ (in homogeneous coordinates) be known. The platform and camera internal parameters K are read to have :

$$\begin{cases} i_s = i_0 - \alpha_i \tan(\phi) \\ i_v = i_0 \tan(\phi) + \alpha_i \\ j_v = j_0 \tan(\phi) \\ t_v = \tan(\phi) \end{cases}$$

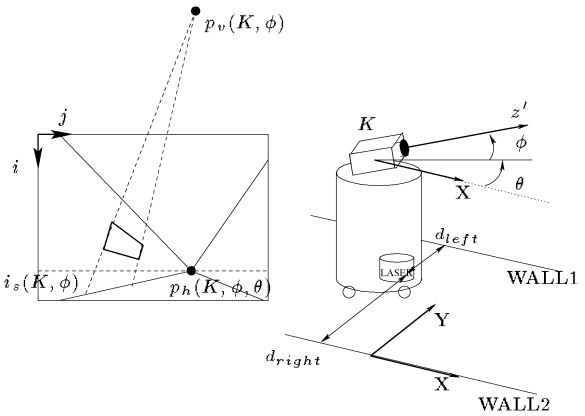


Figure 3: *Spatial configuration in a corridor*

Moreover, the knowledge of the skyline is very useful to perform, on a second step, a quick search of the *horizontal vanishing point* p_h, reduced to one dimensional problem *along the skyline*. This approach is a classical one, but requires (1) an image segmentation into edge segments and (2) a Hough-like transform. It is not very convenient for on-line processing but may be useful, as it will be explained hereafter.

3.2 Laser/camera transformation

A key problem to use laser data in our image processing functions is to have a good estimate of the transformation T_{sc} between the two sensors. Let $(\alpha_{sc}, \beta_{sc}, \gamma_{sc}, tx_{sc}, ty_{sc}, tz_{sc})^T$ be the transformation parameters and T_{sc} the corresponding 4×4 matrix. Physical measuring allows to have a first approximation of T_{sc}. A reasonable hypothesis is that γ_{sc}, resulting from the two roll angles, is close to zero. The other parameters have to be found in a preliminary *calibration phase*.

An interesting method to calibrate the T_{sc} transform consists in *decoupling the angles and translations* parameters thanks to the infinite points. Indeed, let be some corridor images, from both laser and camera. From visual segments, we can apply the Hough transform-based search we mentioned above to get a visually detected horizontal vanishing point $p_h^v = (i_h^v, j_h^v, 1)$ corresponding to the corridor. We can also re-project the infinite point from laser data into a point $p_h^r = (i_h^r, j_h^r)$ depending on T_{sc}. Defining :

$$\begin{cases} f(i) = \arctan(\frac{i_0 - i}{\alpha_i}) \\ g(i,j) = \arctan(\cos(f(i))\frac{j - j_0}{\alpha_j}) \end{cases}$$

α_{sc} and β_{sc} are computed by minimizing :

$$\begin{cases} \alpha_{sc} = \arg\min_\alpha(i_h^v - (i_0 - \alpha_i \tan(f(i_h^r) + \alpha))) \\ \beta_{sc} = \arg\min_\beta(j_h^v - (j_0 + \frac{\alpha_j \tan(g(i_h^v, j_h^r) + \beta)}{\cos(f(i_h^v))})) \end{cases}$$

Finally, the angle errors we get are about 0.02 radians, as seen on figure 4. The histogram represents the angular error on θ corresponding to the shift between p_h^v and p_h^r, for the computed α_{sc}, β_{sc}.

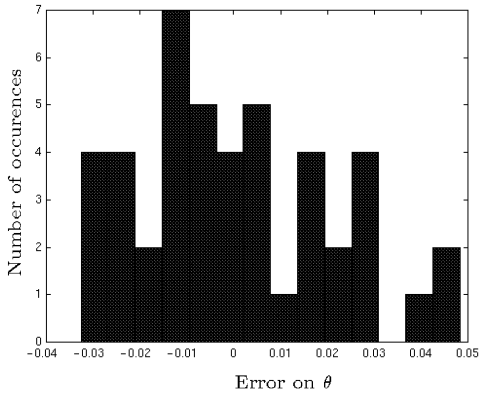


Figure 4: *Error on θ over a test sequence*

At this point, let's define the **corridor lines** by the four parallel straight lines d_k, $k = 1..4$ defining the corridor 3D model. Then, matching the corridor lines projections **detected** onto the image with the model-based **re-projected** ones allows to find the translation parameters. However, the needed precision on these parameters is much less important than the one on the angles, as they will only have influence on the corridor lines projections, not on the directions of detected primitives.

3.3 Projection of laser data

Laser data from the GVG module may provide some useful information to enhance our visual functions. Indeed, we have proposed in [1] a simple method to detect planar, quadrangular landmarks lying on a vertical wall, posters for example. One of the key features of this system was that salient zones detection and segmentation was partially done in a *1D* image resulting from an *averaging procedure over the whole image* along *vertical* direction only. No a priori information about the scene or the robot was used.

However, when the robot enters a corridor-like part of the environment, laser data can provide a reliable estimate of the robot direction along the corridor, so we can get the θ angle from figure 3 and the *horizontal* vanishing point $(i_h, v_h, 1)$:

$$p_h = (i_h, j_h, 1)^T = (i_s, j_0 - \alpha_j \frac{\tan(\theta)}{\cos(\phi)}, 1)^T$$

When the GVG module computes the distances d_{left} and d_{right} to the wall (see figure 3), we can get the four projected corridor lines d_k. We know that they go through point p_h so that for each d_k we only need one more point projection. We can take the intersections of the camera horizontal axis with the wall. For the right wall bottom line d_1, for instance :

$$d_1 = (p_h, KT_\phi T_{sc} \begin{pmatrix} -d_{right}\sin(\theta) \\ d_{right}\cos(\theta) \\ -H_s \\ 1 \end{pmatrix})$$

H is a height arbitrarily chosen, H_s the laser height from the floor.

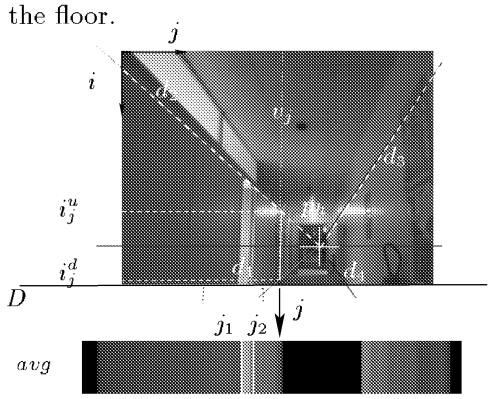


Figure 5: *How laser enhances image processing*

3.4 Detection of salient quadrangles

From the four d_k half straight lines, we can perform an efficient search : let us consider the j values along the horizontal line D, at the bottom of the image, as illustrated in figure 5, with j varying from $j = j_{min}$ to $j = j_{max}$. The equation of D is $i = i_{max}$.

In each j we can then define a direction v_j passing through p_v and two bounds i_j^u and i_j^d on this direction corresponding to the projections of the areas of the lateral walls. The averaging procedure can be done *on this segment only* to get a 1D image *avg* as in figure 5. *avg* is processed to detect salient transitions on D points j_k. These transitions are treated separately to isolate corresponding vertical segments in the image.

From all the points we detailed before, we can now present the detection algorithm as follows :

0: Gets *attitude_data* (ϕ)
1: Computes p_v from *attitude_data*
2:
if (ReadCorridorInfoFromGvg()=OK) **then**
gets θ and computes p_h and d_k, $k = 1..4$
set vertical averaging bounds from d_k
else
set vertical averaging bounds to default
end if
3: Averages $\rightarrow$ image avg along lines v_j
4: Detects in avg transitions points j_k
5:
for all k **do**
Detects transitions along v_{j_k}.
Segmentation and RANSAC estimation.
end for
6: Matches vertical segments together by relaxation.
7: Closes all matched pairs.

The closure procedure, already described in [1], is based on a RANSAC function. We adapted it to take the vanishing point into account. The closure may be *total*, when both lower and upper vertices have been isolated, as the left quadrangle in figure 6; it may be *partial* when, as the right example from the same figure, only one horizontal edge has been found.

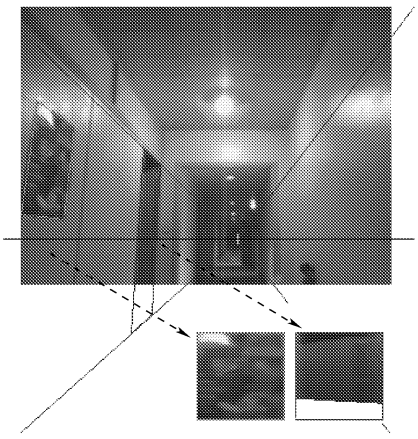


Figure 6: *Results of landmark detection in a typical corridor environment*

The landmark representation is computed from the "iconification" of detected quadrangular landmarks, by applying an homography H on the original image to a 75×75 square SQ. These icons are shown on figure 6. More details can be found in [5].

3.5 Case of partially detected landmarks

The *partially* detected landmarks compose the majority of detected landmarks in indoor environments : doors, cupboards are very frequent in office environments. As we have only three available lines and we need four lines to perform the representation construction, we propose to use the corridor lines to complete them, and we call "door-like" this kind of landmark.

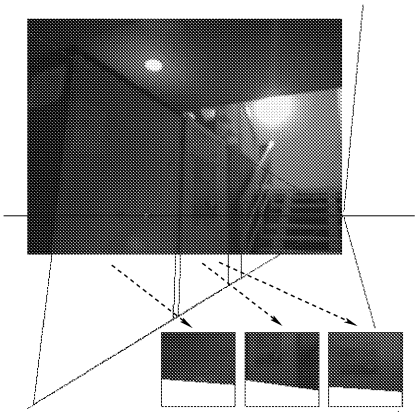


Figure 7: *Partially detected "door-like" landmarks*

Figure 7 illustrates this process on a set of cupboards, where landmarks models are defined from H between the three detected lines and d_1, on the one hand, and the previously defined square SQ, on the other hand.

4 Integrating visual landmarks into the environment representation

Once landmarks have been detected, we have to integrate them into the graph-based representation of the environment. The GVG approach can also embed this kind of information. We saw in [8] that nodes could be annotated with visual landmarks. In this work, we also try to enrich the graph edges with visual information.

There are different levels of information we have to process from visual landmarks :

- *intrinsic data* embedded in the landmark.
- *orientation* relatively to the edge
- *topological and order relationships* with other landmarks

The intrinsic data are extracted from the iconified views, as described in [5]. This appearance representation is robust to illumination, viewpoint and scale changes, so that we are able to recognize the same landmark at different points in the corridor.

Orientation gives the position of the landmark in one of the corridor sides : left/right. Last, topological and order information represent the relative

relationships between landmarks, if available : relative positioning along the wall, relationships of inclusion/intersection/disjunction. As in figure 7, landmarks do not necessarily correspond to physically distinct objects.

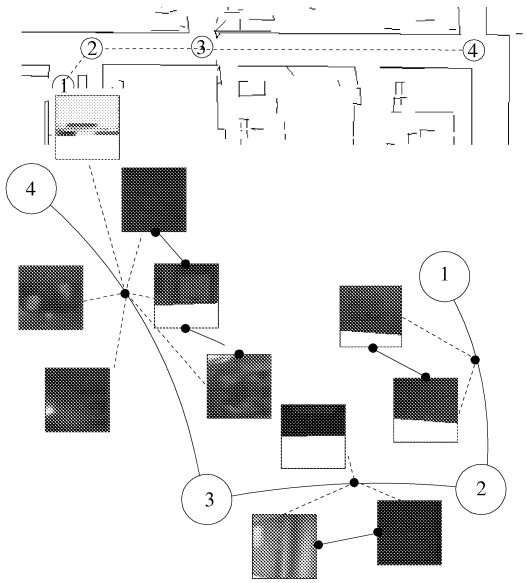


Figure 8: *A graph for visual and laser information*

Figure 8 shows an example of such a graph for the portion of the metric space represented on the upper part. Dashed lines represent links from landmarks to edges, solid lines the existing inter-landmarks topological relationships. Note that not all the connections between landmarks from the same walls have been defined.

Maintaining qualitative knowledge is still part of our current work. We base our approach on [6], where orientation and topology are taken into account simultaneously.

5 Conclusion and future works

This paper has presented the integration of two topological based representations required for the navigation of a mobile robot in an office environment. We take advantage both from the GVG model, suitable to represent a network of corridors, and from a landmark-based topological map to provide a qualitative localization of the robot with respect to the nodes and the edges of the GVG.

The GVG construction relies on laser data, more accurate and less noisy than the ultrasonic data we used in a preliminary work [8]. The landmark learning and recognition method has been improved to become more robust and more reactive. In the corridors, this module takes profit of the laser data to focus the landmark detection procedure only on the lateral walls.

In our lab, the environment is more complex than a simple network of orthogonal corridors, so that the GVG approach is very useful, but vision is mandatory to guarantee a good recognition of the nodes. We are currently trying to get more significant experimental results, including the crossing of open spaces like large hallways, in which the topology cannot be reliably described by a GVG. For such places, we intend to associate edges to visually-controlled actions (for example, *goto landmark* or *follow the line on the wall*, ...). We also intend to consider more types of visual landmarks: vertical edges, lines on the ground or on the walls, quadrangular and planar objects located on the ceiling ... so that the robot could always locate itself.

References

[1] V. Ayala, J.B. Hayet, M. Devy and F. Lerasle. Visual Localization of a Mobile Robot in Indoor Environments using Planar Landmarks IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, Takamatsu, Japan.

[2] H. Choset, K. Nagatani and A. Rizzi. Sensor Based Planning: Using a Honing Strategy and Local Map Method to Implement the Generalized Voronoi Graph. SPIE Mobile Robotics, Pittsburgh, PA, 1997.

[3] E. Clementini and P. DiFelice. A Comparison of Methods for Representing Topological Relationships. Information Sciences,3,149-178,1995.

[4] C. Freksa and K. Zimmermann. On the Utilization of Spatial Structures for Cognitively Plausible and Efficient Reasoning. IEEE Int. Conf. on Systems, Man and Cybernetics, October 1992.

[5] J.B. Hayet, F. Lerasle and M. Devy. A Visual Landmark Framework for Indoor Mobile Robot Navigation. IEEE Int. Conf. on Robotics and Automation, May 2002, pp. 3942-3947, Washington, USA.

[6] D. Hernández. Maintaining Qualitative Spatial Knowledge. European Conf. on Spatial Information Theory, Elba, 1993.

[7] B. Kuipers. The Spatial Semantic Hierarchy. Artificial Intelligence. 2000.

[8] P. Ranganathan, J.B. Hayet, M. Devy, S. Hutchinson and F. Lerasle. A Visual Landmark Framework for Indoor Mobile Robot Navigation. Int. Symp. on Intelligent Robotics Systems, Toulouse, France, 2001.

[9] A.C. Vitorino, P. Rives and J.J. Borrelly. Mobile Robot Navigation Using a Sensor-Based Control Strategy, IEEE Int. Conf. on Robotics and Automation, Seoul, Korea, May 2001.

[10] D. Van Zwynsvoorde, T. Simeon and R. Alami. Incremental Topological Modeling using Local Voronoi-like Graphs, IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, Takamatsu, Japan.

Machine Vision and Applications
DOI 10.1007/s00138-008-0174-7

ORIGINAL PAPER

Particle filtering strategies for data fusion dedicated to visual tracking from a mobile robot

Ludovic Brèthes · Frédéric Lerasle · Patrick Danès · Mathias Fontmarty

Received: 7 December 2006 / Accepted: 15 September 2008

Abstract This paper introduces data fusion strategies within particle filtering in order to track people from a single camera mounted on a mobile robot in a human environment. Various visual cues are described, relying on color, shape or motion, together with several filtering strategies taking into account all or parts of these measurements in their importance and/or measurement functions. A preliminary evaluation enables the selection of the most meaningful visual cues associations in terms of discriminative power, robustness to artifacts and time consumption. The depicted filtering strategies are then evaluated in order to check which people trackers, regarding visual cues and algorithms associations, best fulfill the requirements of the considered scenarios. The performances are compared through some quantitative and qualitative evaluations. Some associations of filtering strategies and visual cues show a significant increase in the tracking robustness and precision. Future works are finally discussed.

Keywords Monocular vision · Person tracking · Particle filtering · Data fusion · Mobile robotics

L. Brèthes · F. Lerasle (✉) · P. Danès · M. Fontmarty
LAAS-CNRS, 7 avenue du Colonel Roche,
31077 Toulouse, France
e-mail: Frederic.Lerasle@laas.fr

L. Brèthes
e-mail: Ludovic.Brethes@laas.fr

P. Danès
e-mail: Patrick.Danes@laas.fr

M. Fontmarty
e-mail: Mathias.Fontmarty@laas.fr

F. Lerasle · P. Danès
Univ. Paul Sabatier, 118 rte de Narbonne,
31062 Toulouse, France

1 Introduction

Tracking people in dynamically changing environments is a critical task over a wide range of applications, e.g. human–computer interface [8,22], teleconferencing [41], surveillance [18,19], motion capture [29], video compression [35], and driver assistance [4]. This paper focuses on mobile robotic applications, where visual tracking of people is one of the ubiquitous elementary functions. Tracking from a mobile platform is a very challenging task, which imposes several requirements. First, the sensors being embedded on the robot, they are usually moving instead of static, and have a restricted perception of the environment. Moreover, the robot is expected to evolve in a wide variety of environmental conditions. Consequently, several hypotheses must be handled simultaneously and a robust integration of multiple visual cues is required in order to achieve some robustness to artifacts. Finally, on-board computational power is limited so that only a small percentage of these resources can be allocated to tracking, the remaining part being required to enable the concurrent execution of other functions as well as decisional routines within the robot's architecture. Thus, care must be taken to design efficient algorithms.

Many 2D people tracking paradigms with a single camera have been proposed in the literature which we shall not attempt to review here. The reader is referred to [15,46] for details. One can mention Kalman filtering [35], the mean-shift technique [12] or its variant [11], tree-based filtering [36] among many others. Beside these approaches, one of the most successful paradigms, focused in this paper, undoubtedly concerns sequential Monte Carlo simulation methods, also known as particle filters [14]. The popularity of these strategies stems from their simplicity, ease of implementation, and modeling flexibility over a wide variety of applications. They seem well-suited to visual tracking as they make

no assumption on the probability distributions entailed in the characterization of the problem and enable an easy combination/fusion of diverse kind of measurements.

Particle filters represent the posterior distribution by a set of samples, or particles, with associated importance weights. This weighted particles set is first drawn from the state vector initial probability distribution, and is then updated over time taking into account the measurements and a prior knowledge on the system dynamics and observation models.

In the Computer Vision community, the formalism has been pioneered in the seminal paper [20] by Isard and Blake, which coins the term CONDENSATION. In this scheme, the particles are drawn from the dynamics and weighted by their likelihood w.r.t. the measurement. CONDENSATION is shown to outperform Kalman filter in the presence of background clutter.

Following the CONDENSATION algorithm, various improvements and extensions have been proposed for visual tracking. Isard et al. in [22] introduce a mixed-state CONDENSATION tracker in order to perform multiple model tracking. The same authors propose in [21] another extension, named ICONDENSATION, which has introduced for the first time importance sampling in visual tracking. It constitutes a mathematically principled way of directing search, combining the dynamics and measurements. So, the tracker can take advantage of the distinct qualities of the information sources and re-initialize automatically when temporary failures occur. Particle filtering with history sampling is proposed as a variant in [37]. Rui and Chen in [34] introduce the Unscented Particle Filter (UPF) into audio and visual tracking. The UPF uses the Unscented Kalman filter to generate proposal distributions that seamlessly integrate the current observation. Partitioned sampling, introduced by MacCormick and Isard in [27], is another way of applying particle filters to tracking problems with high-dimensional configuration spaces. This algorithm is shown to be well-suited to track articulated objects [28]. The hierarchical strategy [33] constitutes a generalization. Last, though outside the scope of this paper, particle filters have also become a popular tool to perform simultaneous tracking of multiple persons [27,42].

As mentioned before, the literature proposes numerous particle filtering algorithms, yet a few studies comparing the efficiency of these filtering strategies have been carried out. When doing so, the associated results are mainly compared against those of the original CONDENSATION approach [26,34,37].

Another observation concerns data fusion. It can be argued that data fusion using particle filtering schemes has been fairly seldom exploited within this visual tracking context. The numerous visual trackers referred to in the literature consider a single cue, i.e. contours [20,28,34] or color [30,32]. The multiple cues association has often been confined to contours and color [8,21,37,47], or color and motion [6,10,33,43].

This data fusion problem has been extensively tackled by Pérez et al. in [33]. The authors propose a hierarchical particle filtering algorithm, which successively takes account of the measurements so as to efficiently draw the particles. To our belief, using multiple cues simultaneously, both in the importance and measurement functions, not only allows to use complementary and redundant information but also enables a more robust failures detection and recovery. More globally, other existing particle filtering strategies should also be evaluated in order to check which ones best fulfill the requirements for the envisaged application.

From these considerations, a first contribution of this paper relates to visual data fusion in robotics scenarios covering a wide variety of environmental conditions. A large spectrum of plausible multi-cues association for such a context is depicted. Evaluations are then performed in order to exhibit the most meaningful visual cues associations in terms of discriminative power, robustness to artifacts and time consumption, be these cues involved in the particle filter importance or measurement functions. A second contribution concerns a thorough comparison of the various particle filtering strategies for data fusion dedicated to the applications envisaged here. Some experiments are presented, in which the designed trackers efficiency is evaluated with respect to temporary target occlusion, presence of significant clutter, as well as large variations in the target appearance and in the illumination of the environment. These trackers have been integrated on a tour-guide robot named Rackham whose role is to help people attending an exhibition.

The paper is organized as follows. Section 2 describes Rackham and outlines the embedded visual trackers. Section 3 sums up the well-known particle filtering formalism, and reviews some variants which enable data fusion for tracking. Then, Sect. 4 specifies some visual measurements which rely on the shape, color or image motion of the observed target. A study comparing the efficiency of various particle filtering strategies is carried out in Sect. 5. Section 6 reports on the implementation of these modalities on Rackham. Last, Sect. 7 summarizes our contribution and puts forward some future extensions.

2 Rackham and the tour-guide scenario progress

Rackham is an iRobot B21r mobile platform whose standard equipment has been extended with one pan-tilt camera EVI-D70 dedicated to H/R interaction, one digital camera for robot localization, one ELO touch-screen, a pair of loudspeakers, an optical fiber gyroscope and wireless Ethernet (Fig. 1).

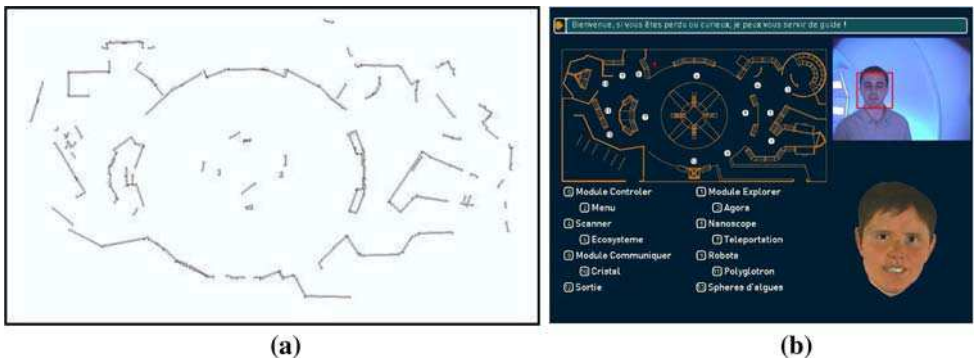
Fig. 1 Rackham

Rackham has been endowed with functions enabling it to act as a tour-guide robot. So, it embeds robust and efficient basic navigation abilities in human-crowded environments. For instance, Fig. 2a shows the laser map of an exhibition, which the robot first builds automatically during an exploration phase with no visitor and then uses for localization. Besides, our efforts have concerned the design of visual functions in order to track, recognize and interact with visitors attending an exhibition. Figure 2b reports the robot's interface display gathering the outputs from such visual functions (top right) together with other interaction facilities: selection of exhibition areas (top left, down left), human-like clone (down right), etc.

When Rackham is left alone with no mission, it tries to find out people whom he could interact with, a behavior hereafter called "search for interaction". As soon as a visitor comes into its neighborhood, it introduces itself, tries to identify his/her face and explains how to use its services thanks to the touch-screen. More precisely, if the interlocutor is unknown, the robot opens a face learning session, then asks him/her to define the mission. On the contrary, when interacting with a formerly identified user, the robot can suggest missions/services which are complementary to the ones executed in the past. Once the robot and its tutor have agreed on an area to visit, Rackham plans and displays its trajectory, prior to inviting its user to follow. While navigating, the robot keeps on giving information about the progress of the ongoing path and verifies the user presence. Whenever the guided visitor leaves during the execution of this "guidance mission", the robot detects this and stops. If, after a few seconds, this user is not re-identified, the robot restarts a "search for interaction" session. Otherwise, when a known user is re-identified, the robot proposes him/her to continue the ongoing "guidance mission".

The design of visual modalities has been undertaken within this demonstration scenario. Three basic tracking modalities, focused in this paper, have been outlined which the robot must basically deal with:

1. *The search for interaction*, where the robot, static and left alone, visually tracks visitors thanks to the camera mounted on its helmet, in order to heckle them when they enter the exhibition (Fig. 3a). This modality involves the whole human body tracking at long H/R distances (>3 m);
2. *The proximal interaction*, where a user can interact through the ELO touch-screen, to select the area he/she wants to visit (Fig. 3b); during this interaction, the robot remains static and must keep, thanks to the camera materializing its eye, the visual contact with its tutor's face at short H/R distances (<1 m);
3. *The guidance mission,* where the robot drives the visitor to the selected area; during its mission, the robot must also maintain the interaction with the guided visitor (Fig. 3c). This modality involves the upper human body tracking at medium H/R distances.



(a) (b)

Fig. 2 Interface display (a), SICK laser map of the exhibition (b)

Fig. 3 The robot visual modalities: **a** search for interaction, **b** proximal interaction, **c** guidance mission

These trackers involves the camera EVI-D70 whose characteristics are: image resolution 320×240 pixels, retina dimension $1/2''$, and focal length 4.1mm.

3 Particle filtering algorithms for data fusion

3.1 A generic algorithm

Particle filters are sequential Monte Carlo simulation methods to the state vector estimation of any Markovian dynamic system subject to possibly non-Gaussian random inputs [3,13,14]. Their aim is to recursively approximate the posterior probability density function (pdf) $p(x_k|z_{1:k})$ of the state vector x_k at time k conditioned on the set of measurements $z_{1:k} = z_1, \ldots, z_k$. A linear point-mass combination

$$p(x_k|z_{1:k}) \approx \sum_{i=1}^{N} w_k^{(i)} \delta\left(x_k - x_k^{(i)}\right), \quad \sum_{i=1}^{N} w_k^{(i)} = 1, \qquad (1)$$

is determined—with $\delta(\cdot)$ the Dirac distribution—which expresses the selection of a value—or "particle"—$x_k^{(i)}$ with probability—or "weight"—$w_k^{(i)}$, $i = 1, \ldots, N$. An approximation of the conditional expectation of any function of x_k, such as the minimum mean square error (MMSE) estimate $\mathrm{E}_{p(x_k|z_{1:k})}[x_k]$, then follows.

Let the system be fully described by the prior $p(x_0)$, the dynamics pdf $p(x_k|x_{k-1})$ and the observation pdf $p(z_k|x_k)$. The generic particle filtering algorithm—or "Sampling Importance Resampling" (SIR)—is shown in Table 1. Its initialization consists in an independent identically distributed (i.i.d.) sequence drawn from $p(x_0)$. At each further time k, the particles keep evolving stochastically, being sampled from an *importance function* $q(x_k|x_{k-1}^{(i)}, z_k)$ which aims at adaptively exploring "relevant" areas of the state space. They are then suitably weighted so as to guarantee the consistency of the approximation (1). To this end, step 5 affects each particle $x_k^{(i)}$ a weight $w_k^{(i)}$ involving its *likelihood* $p(z_k|x_k^{(i)})$ w.r.t. the measurement z_k as well as the values at $x_k^{(i)}$ of the importance function and dynamics pdf.

In order to limit the degeneracy phenomenon, which says that whatever the sequential Monte Carlo simulation method, after few instants all but one particle weights tend to zero, step 8 inserts a resampling stage, e.g. the so-called "systematic resampling" defined in [25]. There, the particles associated with high weights are duplicated while the others collapse, so that the resulting sequence $x_k^{(s^{(1)})}, \ldots, x_k^{(s^{(N)})}$ is i.i.d. according to $\sum_{i=1}^{N} w_k^{(i)} \delta(x_k - x_k^{(i)})$. Note that this resampling stage should rather be fired only when the filter efficiency—related to the number of "useful" particles—goes beneath a predefined threshold [14].

3.2 Importance sampling from either dynamics or measurements: basic strategies

The CONDENSATION—for "Conditional Density Propagation" [20]—can be viewed as the instance of the SIR algorithm in which the particles are drawn according to the system dynamics, viz. when $q(x_k|x_{k-1}^{(i)}, z_k) = p(x_k|x_{k-1}^{(i)})$. This endows CONDENSATION with a prediction-update structure, in that $\sum_{i=1}^{N} w_{k-1}^{(i)} \delta(x_k - x_k^{(i)})$ approximates the prior $p(x_k|z_{1:k-1})$. The weighting stage becomes $w_k^{(i)} \propto w_{k-1}^{(i)} p(z_k|x_k^{(i)})$. In the visual tracking context, the original algorithm [20] defines the particles likelihoods from contour primitives, yet other visual cues have also been exploited [33].

Resampling by itself cannot efficiently limit the degeneracy phenomenon. In addition, it may lead to a loss of diversity in the state space exploration. The importance function must thus be defined with special care.

In visual tracking, the modes of the likelihoods $p(z_k|x_k)$, though multiple, are generally pronounced. As CONDENSATION draws the particles $x_k^{(i)}$ from the system dynamics but "blindly" w.r.t. the measurement z_k, many of these may well be assigned a low likelihood $p(z_k|x_k^{(i)})$ and thus a low weight in step 5, significantly worsening the overall filter performance. An alternative, henceforth labeled "Measurement-based SIR" (MSIR), merely consists in sampling the particles at time k—or just some of their entries—according to an importance function $\pi(x_k|z_k)$ defined from the current image. The first MSIR strategy was ICONDENSATION [21], which guides the state space exploration by a color blobs detector. Other visual detection functionalities can be used as well, e.g. face detector (Sect. 4), or any other intermittent primitive which, despite its sporadicity, is very discriminant when present [33]: motion, sound, etc.

In an MSIR scheme, a particle $x_k^{(i)}$ whose entries are drawn from the current image may be inconsistent with its predecessor $x_{k-1}^{(i)}$ from the point of view of the state dynamics. As expected, the smaller is the value $p(x_k^{(i)}|x_{k-1}^{(i)})$, the lesser

Table 1 Generic particle filtering algorithm (SIR)

$[\{x_k^{(i)}, w_k^{(i)}\}]_{i=1}^N = \mathrm{SIR}([\{x_{k-1}^{(i)}, w_{k-1}^{(i)},\}]_{i=1}^N, z_k)$

1: **IF** $k = 0$, **THEN** Draw $x_0^{(1)}, \ldots, x_0^{(i)}, \ldots, x_0^{(N)}$ i.i.d. according to $p(x_0)$, and set $w_0^{(i)} = \frac{1}{N}$ **END IF**
2: **IF** $k \geq 1$ **THEN** $\{$—$[\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^N$ being a particle description of $p(x_{k-1}|z_{1:k-1})$—$\}$
3: **FOR** $i = 1, \ldots, N$, **DO**
4: "Propagate" the particle $x_{k-1}^{(i)}$ by independently sampling $x_k^{(i)} \sim q(x_k|x_{k-1}^{(i)}, z_k)$
5: Update the weight $w_k^{(i)}$ associated with $x_k^{(i)}$ according to $w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(z_k|x_k^{(i)}) p(x_k^{(i)}|x_{k-1}^{(i)})}{q(x_k^{(i)}|x_{k-1}^{(i)}, z_k)}$, prior to a normalization step s.t. $\sum_i w_k^{(i)} = 1$
6: **END FOR**
7: Compute the conditional mean of any function of x_k, *e.g.* the MMSE estimate $\mathrm{E}_{p(x_k|z_{1:k})}[x_k]$, from the approximation $\sum_{i=1}^N w_k^{(i)} \delta(x_k - x_k^{(i)})$ of the posterior $p(x_k|z_{1:k})$
8: At any time or depending on an "efficiency" criterion, resample the description $[\{x_k^{(i)}, w_k^{(i)}\}]_{i=1}^N$ of $p(x_k|z_{1:k})$ into the equivalent evenly weighted particles set $[\{x_k^{(s^{(i)})}, \frac{1}{N}\}]_{i=1}^N$, by sampling in $\{1, \ldots, N\}$ the indexes $s^{(1)}, \ldots, s^{(N)}$ according to $P(s^{(i)} = j) = w_k^{(j)}$; set $x_k^{(i)}$ and $w_k^{(i)}$ with $x_k^{(s^{(i)})}$ and $\frac{1}{N}$
9: **END IF**

is the weight $w_k^{(i)}$. One solution to this problem, as proposed in the genuine ICONDENSATION algorithm, consists in sampling some of the particles from the dynamics and some w.r.t. the prior, so that the importance function reads as, with $\alpha, \beta \in [0; 1]$

$$q\left(x_k|x_{k-1}^{(i)}, z_k\right) = \alpha\pi(x_k|z_k) + \beta p\left(x_k|x_{k-1}^{(i)}\right) + (1 - \alpha - \beta) p_0(x_k). \quad (2)$$

This combination enables the tracker to benefit from the distinct qualities of the information sources and to re-initialize automatically when temporary failures occur.

3.3 Towards the "optimal" case: the auxiliary particle filter

It can be shown [14] that the "optimal" recursive scheme, i.e. which best limits the degeneracy phenomenon, must define $q^*(x_k|x_{k-1}^{(i)}, z_k) \triangleq p(x_k|x_{k-1}^{(i)}, z_k)$ and thus $w_k^{*(i)} \propto w_{k-1}^{*(i)} p(z_k|x_{k-1}^{(i)})$ in the SIR algorithm (Table 1). Each weight $w_k^{*(i)}$ can then be computed before drawing $x_k^{(i)}$. So, the overall efficiency can be enhanced by resampling the weighted particle set $[\{x_{k-1}^{(i)}, w_k^{*(i)}\}]_{i=1}^N$, which in fact represents the smoother pdf $p(x_{k-1}|z_{1:k})$, just before its "propagation" through the optimal importance function $q^*(x_k|x_{k-1}^{(i)}, z_k)$.

Despite such an algorithm can be seldom implemented exactly, it can be mimicked by the "Auxiliary Particle Filter" (AUXILIARY_PF) along the lines of [31], see Table 2. Let the importance function $\pi(x_k|x_{k-1}^{(i)}, z_k) = p(x_k|x_{k-1}^{(i)})$ be defined in place of $q^*(x_k|x_{k-1}^{(i)}, z_k)$, and $\hat{p}(z_k|x_{k-1}^{(i)})$ be an approximation of the *predictive likelihood* $p(z_k|x_{k-1}^{(i)})$ (steps 3–5); for instance, one can set $\hat{p}(z_k|x_{k-1}^{(i)}) = p(z_k|\mu_k^{(i)})$, where $\mu_k^{(i)}$ characterizes the distribution of x_k conditioned on $x_{k-1}^{(i)}$, e.g. $\mu_k^{(i)} = \mathrm{E}_{p(x_k|x_{k-1}^{(i)})}[x_k]$ or $\mu_k^{(i)} \sim p(x_k|x_{k-1}^{(i)})$. First, an *auxiliary weight* $\lambda_k^{(i)} \propto w_{k-1}^{(i)} \hat{p}(z_k|x_{k-1}^{(i)})$ is associated with each particle $x_{k-1}^{(i)}$. The approximation $[\{x_{k-1}^{(i)}, \lambda_k^{(i)}\}]_{i=1}^N$ of $p(x_{k-1}|z_{1:k})$ is then resampled into $[\{x_{k-1}^{(s^{(i)})}, \frac{1}{N}\}]_{i=1}^N$ (step 6), prior to its propagation until time k through $\pi(x_k|x_{k-1}^{(s^{(i)})}, z_k)$ (step 8). Finally, the weights of the resulting particles $x_k^{(i)}$ must be corrected (step 9) in order to take account of the "distance" between $\lambda_k^{(i)}$ and $w_k^{*(i)}$, as well as of the dissimilarity between the selected and optimal importance functions $\pi(x_k^{(i)}|x_{k-1}^{(s^{(i)})}, z_k)$ and $p(x_k^{(i)}|x_{k-1}^{(s^{(i)})}, z_k)$.

The particles cloud can thus be steered towards relevant areas of the state space. In the visual tracking context, the approximate predictive likelihood can rely on distinct visual cues from these involved in the computation of the "final-stage" likelihoods $p(z_k|x_k^{(i)})$. The main limitation of the AUXILIARY_PF algorithm is its bad performance when the dynamics is uninformative compared to the state final-stage likelihood, e.g. when the dynamics is very noisy or when the observation density has sharp modes. Therefore, $\mu_k^{(i)}$ being a bad characterization of $p(x_k|x_{k-1}^{(i)})$, the pdf $p(x_{k-1}|z_{1:k})$ of the smoother is poorly approximated by $[\{x_{k-1}^{(i)}, \lambda_k^{(i)}\}]_{i=1}^N$. The resampling stage in step 6 of Table 2 may well eliminate some particles which, once propagated through the dynamics, would be very likely w.r.t. z_k. At the same time, other particles may well be duplicated which, after the prediction step, come to lie in the final-stage likelihood tails.

In the framework of auxiliary particle filters, the Unscented Transform [24] can constitute a way to define a better approximation $\hat{p}(z_k|x_{k-1})$ of the predictive likelihood $p(z_k|x_{k-1})$, which is the basis of the auxiliary resampling stage, see steps 3–6 of Table 2. As is the case in the Unscented Particle Filter [39], this transform can also be entailed in the association to each particle of the Gaussian near-optimal importance function from which it is sampled. Andrieu et al. propose such a strategy in [2]. Nevertheless, despite

Table 2 Auxiliary particle filter (AUXILIARY_PF)

$[\{x_k^{(i)}, w_k^{(i)}\}]_{i=1}^N = \text{AUXILIARY\_PF}([\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^N, z_k)$

1: **IF** $k = 0$, **THEN** Draw $x_0^{(1)}, \ldots, x_0^{(i)}, \ldots, x_0^{(N)}$ i.i.d. according to $p(x_0)$, and set $w_0^{(i)} = \frac{1}{N}$ **END IF**
2: **IF** $k \geq 1$ **THEN** $\{$—$[\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^N$ being a particle description of $p(x_{k-1}|z_{1:k-1})$—$\}$
3: **FOR** $i = 1, \ldots, N$, **DO**
4: From the approximation $\hat{p}(z_k|x_{k-1}^{(i)}) = p(z_k|\mu_k^{(i)})$ –*e.g.* with $\mu_k^{(i)} \sim p(x_k|x_{k-1}^{(i)})$ or $\mu_k^{(i)} = \mathrm{E}_{p(x_k|x_{k-1}^{(i)})}[x_k]$–, compute the auxiliary weights $\lambda_k^{(i)} \propto w_{k-1}^{(i)} \hat{p}(z_k|x_{k-1}^{(i)})$, prior to a normalization step s.t. $\sum_i \lambda_k^{(i)} = 1$
5: **END FOR**
6: Resample $[\{x_{k-1}^{(i)}, \lambda_k^{(i)}\}]_{i=1}^N$ –or, equivalently, sample in $\{1, \ldots, N\}$ the indexes $s^{(1)}, \ldots, s^{(N)}$ of the particles at time $k-1$ according to $P(s^{(i)} = j) = \lambda_k^{(j)}$ – in order to get $[\{x_{k-1}^{(s^{(i)})}, \frac{1}{N}\}]_{i=1}^N$; both $\sum_{i=1}^N \lambda_k^{(i)} \delta(x_{k-1} - x_{k-1}^{(i)})$ and $\frac{1}{N}\sum_{i=1}^N \delta(x_{k-1} - x_{k-1}^{(s^{(i)})})$ mimic $p(x_{k-1}|z_{1:k})$
7: **FOR** $i = 1, \ldots, N$, **DO**
8: "Propagate" the particles by independently drawing $x_k^{(i)} \sim p(x_k|x_{k-1}^{(s^{(i)})})$
9: Update the weights, prior to their normalization, by setting $w_k^{(i)} \propto \frac{p(z_k|x_k^{(i)}) p(x_k^{(i)}|x_{k-1}^{(s^{(i)})})}{\hat{p}(z_k|x_{k-1}^{(s^{(i)})}) \pi(x_k^{(i)}|x_{k-1}^{(s^{(i)})}, z_k)} = \frac{p(z_k|x_k^{(i)})}{\hat{p}(z_k|x_{k-1}^{(s^{(i)})})} = \frac{p(z_k|x_k^{(i)})}{p(z_k|\mu_k^{(s^{(i)})})}$
10: Compute $\mathrm{E}_{p(x_k|z_{1:k})}[x_k]$ from the approximation $\sum_{i=1}^N w_k^{(i)} \delta(x_k - x_k^{(i)})$ of the posterior $p(x_k|z_{1:k})$
11: **END FOR**
12: **END IF**

its attractiveness and its ability to mimic the optimal case, this is more difficult to implement and shows a higher computational cost.

3.4 Other strategies suited to visual tracking

3.4.1 History sampling

Several interesting particle filtering alternatives to visual tracking are proposed in [37]. One of them considers dynamic models of order greater than or equal to 2, in which the state vector has the form $x_k = (u_k', v_k', h_k')'$, with $[.]'$ the transpose operator. The subvector $(u_k', v_k')'$ or "innovation part"—of x_k obeys a stochastic state equation on x_{k-1}, while h_k—called "history part"—is a deterministic function $f(x_{k-1})$. It is assumed that the innovations $(u_k^{(i)\prime}, v_k^{(i)\prime})'$ are sampled from an importance function such as $q_I(u_k, v_k|x_{k-1}^{(i)}, z_k) = \pi(u_k|z_k) p(v_k|u_k^{(i)}, x_{k-1}^{(i)})$, i.e. the subparticles $u_k^{(i)}$ are positioned from the measurement only while the $v_k^{(i)}$'s are drawn by fusing the state dynamics with the knowledge of $u_k^{(i)}$—and that the pdf of the measurement conditioned on the state satisfies $p(z_k|x_k) = p(z_k|u_k, v_k)$. This context is particularly well-suited to visual tracking, for state-space representations of linear AR models entail the above decomposition of the state vector, and because the output equation does not involve its "history part".

The authors define procedures enabling the avoidance of any contradiction between $(u_k^{(i)\prime}, v_k^{(i)\prime})'$ and its past $x_{k-1}^{(i)}$. Their "Rao-Blackwellized Subspace SIR with History Sampling" (RBSSHSSIR) is summarized in Table 3. Its step 5 consists, for each subparticle $u_k^{(i)}$ drawn from $\pi(u_k|z_k)$, in the resampling of a predecessor particle—and thus of the "history part" of $x_k^{(i)}$—which is at the same time likely w.r.t. $u_k^{(i)}$ from the dynamics point of view and assigned with a significant weight. The RBSSHSSIR algorithm noticeably differs from ICONDENSATION precisely because of this stage, yet necessary lest the weighted particles $[\{x_k^{(i)}, w_k^{(i)}\}]_{i=1}^N$ is not a consistent description of the posterior $p(x_k|z_{1:k})$.

An original proof of the RBSSHSSIR algorithm is sketched in [9], using arguments similar to these underlying the AUXILIARY_PF. It is shown that the algorithm applies even when the state process is of the first order, by just suppressing the entry $f(x_{k-1})$ from x_k.

3.4.2 Partitioned and hierarchical sampling

Partitioned and Hierarchical sampling can significantly enhance the efficiency of a particle filter in cases when the system dynamics comes as the successive application of elementary dynamics, provided that intermediate likelihoods can be defined on the state vector after applying each partial evolution model. The classical single-stage sampling of the full state space is then replaced by a layered sampling approach: thanks to a succession of sampling operations followed by resamplings based on the intermediate likelihoods, the search can be guided so that each sampling stage refines the output from the previous stage.

To outline the technical aspects of each strategy, let $\xi_0 = x_{k-1}, \xi_1, \ldots, \xi_{M-1}, \xi_M = x_k$ be $M+1$ "auxiliary vectors" such that the dynamics $p(x_k|x_{k-1})$ reads as the convolution

$$p(x_k|x_{k-1}) = \int \tilde{d}_M(\xi_M|\xi_{M-1}) \ldots \tilde{d}_1(\xi_1|\xi_0)\, \mathrm{d}\xi_1 \ldots \mathrm{d}\xi_{M-1}, \qquad (3)$$

Table 3 Rao-Blackwellized subspace particle filter with history sampling (RBSSHSSIR)

$[\{x_k^{(i)}, w_k^{(i)}\}]_{i=1}^N = \text{RBSSHSSIR}([\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^N, z_k)$

1: **IF** $k = 0$, **THEN** Draw $x_0^{(1)}, \ldots, x_0^{(i)}, \ldots, x_0^{(N)}$ i.i.d. according to $p(x_0)$, and set $w_0^{(i)} = \frac{1}{N}$ **END IF**
2: **IF** $k \geq 1$ **THEN** $\{$—$[\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^N$ being a particle description of $p(x_{k-1}|z_{1:k-1})$—$\}$
3: **FOR** $i = 1, \ldots, N$, **DO**
4: Draw $u_k^{(i)} \sim \pi(u_k|z_k)$
5: Sample in $\{1, \ldots, N\}$ the index $I_k^{(i)}$ of the predecessor particle of $u_k^{(i)}$ according to the weights $(w_{k-1}^{(1)} p(u_k^{(i)}|x_{k-1}^{(1)}), \ldots, w_{k-1}^{(.)} p(u_k^{(i)}|x_{k-1}^{(.)}), \ldots, w_{k-1}^{(N)} p(u_k^{(i)}|x_{k-1}^{(N)}))$, *i.e.* according to $P(I_k^{(i)} = j) = \frac{w_{k-1}^{(j)} p(u_k^{(i)}|x_{k-1}^{(j)})}{\sum_{l=1}^N w_{k-1}^{(l)} p(u_k^{(i)}|x_{k-1}^{(l)})}$
6: Draw $v_k^{(i)} \sim p(v_k|u_k^{(i)}, x_{k-1}^{(I_k^{(i)})})$
7: Set $x_k^{(i)} = \left(u_k^{(i)\prime}, v_k^{(i)\prime}, f(x_{k-1}^{(I_k^{(i)})})' \right)'$
8: Update the weights, prior to their normalization, by setting $w_k^{(i)} \propto \frac{p(z_k|u_k^{(i)}) \sum_{l=1}^N w_{k-1}^{(l)} p(u_k^{(i)}|x_{k-1}^{(l)})}{\pi(u_k^{(i)}|z_k)}$
9: Compute the conditional mean of any function of x_k, *e.g.* the MMSE estimate $\text{E}_{p(x_k|z_{1:k})}[x_k]$, from the approximation $\sum_{i=1}^N w_k^{(i)} \delta(x_k - x_k^{(i)})$ of the posterior $p(x_k|z_{1:k})$
10: **END FOR**
11: **END IF**

i.e. the successive application of $\tilde{d}_1(\xi_1|\xi_0), \ldots, \tilde{d}_m(\xi_m|\xi_{m-1}), \ldots, \tilde{d}_M(\xi_M|\xi_{M-1})$. The measurement pdf $p(z_k|x_k)$ is supposed to factorize as $p(z_k|x_k) = \prod_{m=1}^M p_m(z_k|x_k)$.

The second partitioned particle filtering algorithm proposed in [28] assumes that the state vector dynamics are component-wise independent, i.e. if all the vectors ξ_m, $m = 1, \ldots, M$, are analogously partitioned into M subvectors $\xi_m^1, \ldots, \xi_m^M$, then $\tilde{d}_m(\xi_m|\xi_{m-1}) = p(\xi_m^m|\xi_{m-1}^m) \prod_{r \neq m} \delta(\xi_m^r - \xi_{m-1}^r)$ holds for all $m = 1, \ldots, M$ so that $p(x_k|x_{k-1}) = \prod_{m=1}^M p(x_k^m|x_{k-1}^m)$. In addition, the intermediate likelihoods $p_m(z_k|x_k)$ are supposed to concern a subset of the state vector all the more important as $m \longrightarrow M$, i.e. to have the form $p_m(z_k|x_k) = l_m(z_k|x_k^1, \ldots, x_k^m)$. Under these hypotheses, the partitioned particle filter follows the algorithm outlined in Table 4, with $\tilde{q}_m(\xi_m|\xi_{m-1}, z_k) = \tilde{d}_m(\xi_m|\xi_{m-1})$, $p_m(z_k|x_k) = l_m(z_k|x_k^1, \ldots, x_k^m)$.

Partitioned sampling has been successfully applied to the visual tracking of an open kinematic chain in [28], by organizing the state vector so that its first entries depict the top elements of the chain—to be accurately positioned sooner for a higher efficiency—while its last components are related to the extremities. A branched algorithm has also proved to be able to track multiple persons in [27].

The hierarchical particle filter developed in [33] can be viewed as a generalization of the partitioned scheme outlined above. No restriction is imposed on the functions $\tilde{d}_1(\cdot|\cdot), \tilde{d}_M(\cdot|\cdot)$. The measurement z_k is supposed made up with M sensory information $z_k^1, \ldots, z_k^M$ conditionally independent given x_k, so that the intermediate likelihoods come as $p_m(z_k|x_k) = p_m(z_k^m|x_k)$. Importantly, the particles relative to the auxiliary vectors $\xi_1, \ldots, \xi_M$ are not sampled from $\tilde{d}_1(\cdot|\cdot), \ldots, \tilde{d}_M(\cdot|\cdot)$ but instead from distributions $\tilde{q}_1(\cdot|\cdot), \ldots, \tilde{q}_M(\cdot|\cdot)$ related to the importance function $q(x_k|x_{k-1}, z_k)$ by $q(x_k|x_{k-1}, z_k) = \int \tilde{q}_M(x_k|\xi_{M-1}, z_k^M) \ldots \tilde{q}_1(\xi_1|x_{k-1}, z_k^1)\, d\xi_1 \ldots d\xi_{M-1}$.

Incorporating each likelihood $p_m(z_k|\cdot)$ after applying the intermediate dynamics leads to the algorithm depicted in Table 4, with $\tilde{q}_m(\xi_m|\xi_{m-1}, z_k) = \tilde{q}_m(\xi_m|\xi_{m-1}, z_k^m)$ and $p_m(z_k|x_k) = p_m(z_k^m|x_k)$.

4 Importance and measurement functions

Importance sampling offers a mathematically principled way of directing search according to visual cues which are discriminant though possibly intermittent, e.g. motion. Such cues are logical candidates for detection modules and efficient proposal distributions. Besides, each sample weight is updated taking into account its likelihood w.r.t. the current image. This likelihood is computed by means of measurement functions, according to visual cues (e.g. color, shape) which must be persistent but may however be proner to ambiguity in cluttered scenes. In both importance sampling and weight update steps, combining or fusing multiple cues enables the tracker to better benefit from distinct information sources, and can decrease its sensitivity to temporary failures in some of the measurement processes. Measurement and importance functions are depicted in the next subsections.

4.1 Measurement functions

4.1.1 Shape cue

The use of shape cues requires that silhouette templates of human limbs have been learnt beforehand (Fig. 4). Each

Table 4 Partitioned (PARTITIONED_PF) and hierarchical (HIERARCHICAL_PF) particle filtering: $\tilde{q}_m(\xi_m|\xi_{m-1}, z_k)$ and $p_m(z_k|x_k)$ are defined either as: (PARTITIONED_PF) $\tilde{q}_m(\xi_m|\xi_{m-1}, z_k) = \tilde{d}_m(\xi_m|\xi_{m-1})$, $p_m(z_k|x_k) = l_m(z_k|x_k^1, \ldots, x_k^m)$, or: (HIERARCHI- CAL_PF) $\tilde{q}_m(\xi_m|\xi_{m-1}, z_k) = \tilde{q}_m(\xi_m|\xi_{m-1}, z_k^m)$, $p_m(z_k|x_k) = p_m(z_k^m|x_k)$

$[\{x_k^{(i)}, w_k^{(i)}\}]_{i=1}^N$= PARTITIONED_OR_HIERARCHICAL_PF $\left([\{x_{k-1}^{(i)}, w_{k-1}^{(i)},\}]_{i=1}^N, z_k\right)$

1: **IF** $k = 0$, **THEN** Draw $x_0^{(1)}, \ldots, x_0^{(i)}, \ldots, x_0^{(N)}$ i.i.d. according to $p(x_0)$, and set $w_0^{(i)} = \frac{1}{N}$ **END IF**
2: **IF** $k \geq 1$ **THEN** —$[\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^N$ being a particle description of $p(x_{k-1}|z_{1:k-1})$—
3: Set $\{\xi_0^{(i)}, \tau_0^{(i)}\} = \{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}$
4: **FOR** $m = 1, \ldots, M$, **DO**
5: **FOR** $i = 1, \ldots, N$, **DO** Independently sample $\xi_m^{(i)} \sim \tilde{q}_m(\xi_m|\xi_{m-1}^{(i)}, z_k)$ and associate $\xi_m^{(i)}$ the weight $\tau_m^{(i)} \propto \tau_{m-1}^{(i)} \frac{p_m(z_k|\xi_m^{(i)})\tilde{d}_m(\xi_m^{(i)}|\xi_{m-1}^{(i)})}{\tilde{q}_m(\xi_m^{(i)}|\xi_{m-1}^{(i)}, z_k)}$ **END FOR**
6: Resample $[\{\xi_m^{(i)}, \tau_m^{(i)}\}]_{i=1}^N$ into the evenly weighted particles set $[\{\xi_m^{(s^{(i)})}, \frac{1}{N}\}]_{i=1}^N$; rename $[\{\xi_m^{(s^{(i)})}, \frac{1}{N}\}]_{i=1}^N$ into $[\{\xi_m^{(i)}, \tau_m^{(i)}\}]_{i=1}^N$
7: **END FOR**
8: Set $\{x_k^{(i)}, w_k^{(i)}\} = \{\xi_M^{(i)}, \tau_M^{(i)}\}$, which is a consistent description of $p(x_k|z_{1:k})$
⋮
(…): **END IF**

particle x is classically given an edge-based likelihood $p(z^S|x)$ that depends on the sum of the squared distances between N_p points uniformly distributed along the template corresponding to x and their nearest image edges [20], i.e.

$$p\left(z^S|x\right) \propto \exp\left(-\frac{D^2}{2\sigma_s^2}\right), \quad D = \sum_{j=1}^{N_p} |x(j) - z(j)|, \qquad (4)$$

where the similarity measure D involves each jth template point $x(j)$ and associated closest edge $z(j)$ in the image, the standard deviation σ_s being determined a priori.

A variant [17] consists in converting the edge image into a Distance Transform image. Interestingly, the DT is a smoother function of the model parameters. In addition, the DT image reduces the involved computations as it needs to be generated only once whatever the number of particles involved in the filter. The similarity distance D in (4) is replaced by

$$D = \sum_{j=1}^{N_p} I_{\mathrm{DT}}(j), \qquad (5)$$

where $I_{\mathrm{DT}}(j)$ terms the DT image value at the j-th template point. Figure 5 plots this shape-based likelihood for an example where the target is a 2D elliptical template corresponding coarsely to the subject on the right of the input image.

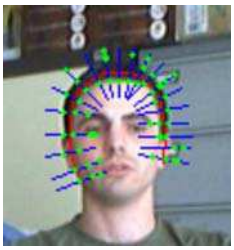


Fig. 4 Shape cue

In case of cluttered background, using only shape cues for the model-to-image fitting is not sufficiently discriminant, as multiple peaks are present.

4.1.2 Color cue

Reference color models can be associated with the targeted ROIs. These models are defined either a priori, or on-line using some automatic detection modules. Let h_{ref}^c and h_x^c two N_{bi}-bin normalized histograms in channel $c \in \{R, G, B\}$, respectively corresponding to the model and to a region B_x parametrized by the state x. The color likelihood $p(z^C|x)$ must favor candidate color histograms h_x^c close to the reference histogram h_{ref}^c. The likelihood has a form similar to (4), provided that D terms the Bhattacharyya distance [30] between the two histograms h_x^c and h_{ref}^c, i.e. for a channel c,

$$D(h_x^c, h_{\mathrm{ref}}^c) = \left(1 - \sum_{j=1}^{N_{bi}} \sqrt{h_{j,x}^c \cdot h_{j,\mathrm{ref}}^c}\right)^{1/2}. \qquad (6)$$

A single histogram does not capture any information on the spatial arrangement of colors and so can lead to noticeable drift. This drift can be avoided by splitting the tracked region into sub-regions with individual reference color models. Let

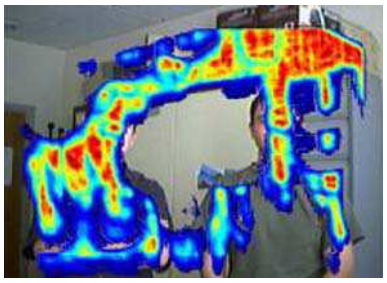


Fig. 5 Shape DT-based likelihood

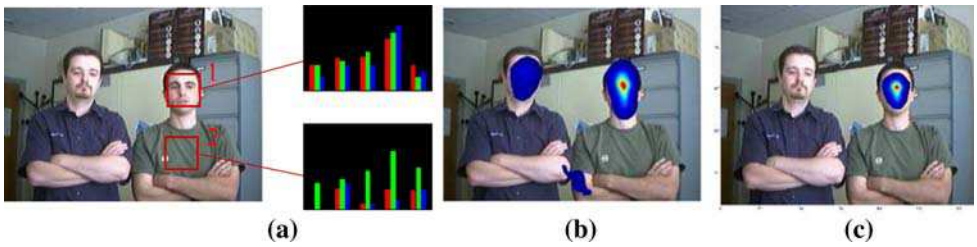


Fig. 6 **a** Color-based regions of interest (ROIs) and corresponding RGB histograms. **b**, **c** Likelihoods regarding single-part and multiple-part color models, respectively

the union $B_x = \bigcup_{p=1}^{N_R} B_{x,p}$ be associated with the set of N_R reference histograms $\{h^c_{\text{ref},p} : c \in \{\text{R, G, B}\}, \text{p} = 1, \ldots, \text{N}_\text{R}\}$. By assuming conditional independence of the color measurements, the likelihood $p(z^C|x)$ becomes

$$p(z^C|x) \propto \exp\left(-\sum_c \sum_{p=1}^{N_R} \frac{D^2(h^c_{x,p}, h^c_{\text{ref},p})}{2\sigma_c^2}\right). \quad (7)$$

Figure 6b and c plots single and multi-patch likelihoods for the above example. The ROIs corresponding to the face and clothes of the person on the right, are compared to their reference model shown in Fig. 6a.

4.1.3 Motion cue

For a static camera, a basic method consists in computing the luminance absolute difference image from successive frames. To capture motion activity, we propose to embed the frame difference information into a likelihood model similar to the one developed for the color measurements.

Pérez et al. in [33] define a reference histogram model for motion cues. For motionless regions, the measurements fall in the lower histograms bins while moving regions fall a priori in all the histograms bins. From these considerations, the reference motion histogram h^M_{ref} is given by $h^M_{j,\text{ref}} = \frac{1}{N'_{bi}}$, $j = 1, \ldots, N'_{bi}$. The motion likelihood is set to

$$p(z^M|x) \propto \exp\left(-\frac{D^2(h^M_x, h^M_{\text{ref}})}{2\sigma_m^2}\right), \quad (8)$$

and is illustrated on Fig. 7a, b, and c.

4.1.4 Multi-cues fusion

Fusing multiple cues enables the tracker to better benefit from M distinct measurements $(z^1, \ldots, z^M)$. Assuming that these are mutually independent conditioned on the state, the unified measurement function thus factorizes as

$$p(z^1, \ldots, z^M|x) = \prod_{m=1}^{M} p(z^m|x). \quad (9)$$

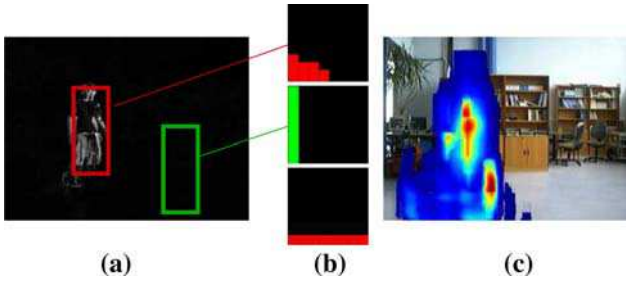


Fig. 7 **a** Absolute luminance frame difference. **b** Motion histograms of two ROIs (*top*, *middle*) and of a reference ROI (*bottom*). **c** Consequent associated likelihood

Yet, to avoid the evaluation of the likelihood for each cue, we hereafter propose some variants so as to combine multiple cues into a single likelihood model.

4.1.5 Shape and motion cues combination

Considering a static camera, it is highly possible that the targeted subject be moving, at least intermittently. To cope with background clutter, we thus favor the moving edges (if any) by combining motion and shape cues into the definition of the likelihood $p(z^S, z^M|x)$ of each particle x. Given $\overrightarrow{f}(z(j))$ the optical flow vector at pixel $z(j)$, the similarity distance D in (4) is then replaced by

$$D = \sum_{j=1}^{N_p} |x(j) - z(j)| + \rho \cdot \gamma(z(j)), \quad (10)$$

where $\gamma(z(j)) = 0$ (resp. 1) if $\overrightarrow{f}(z(j)) \neq 0$ (resp. if $\overrightarrow{f}(z(j)) = 0$) and $\rho > 0$ terms a penalty. Figure 8 plots this more discriminant likelihood function for the example seen above. The target is still the subject on the right, but is assumed to be moving.

4.1.6 Shape and color cues combination

We propose in [7] a likelihood model $p(z^S, z^C|x)$ which combines both shape and color cues through a skin-colored regions segmentation. The use of color features makes the tracker more robust to situations where there is poor grey-level contrast between the human limbs and the background.

Fig. 8 Likelihood combining shape and motion cues

Numerous techniques for skin blobs segmentation are based on a skin pixel classification (see a review in [44]) as human skin colors have specific color distributions. Training images from the Compaq database [23] enable to construct a reference color histogram model [35] in a selected color space. The originality of the segmentation method [7] lies in the sequential application of two watershed algorithms, the first one being based on chromatic information and the last one relying on the intensity of the selected skin-color pixels. This second phase is useful to segment regions with similar colors but different luminance values (like hand and sleeve in Fig. 9).

A new DT image I'_{DT} is defined from the set of Canny edges points I_{DT} and from the contours of the segmented skin blobs. The latters enable to define a mask applied onto the original DT image I_{DT}. Canny edges points which are outside the mask are thus given a penalty in the DT image I'_{DT} as illustrated in Fig. 9d.

The similarity distance D in (4) is then replaced by

$$D = \sum_{j=1}^{N_p} I'_{\mathrm{DT}}(j) + \rho \cdot \gamma(j), \tag{11}$$

where $\gamma(j) = 0$ (resp. 1) if $z^{\mathrm{mask}}(j) = 1$ (resp. if $z^{\mathrm{mask}}(j) = 0$) and $\rho > 0$ terms a penalty. This strategy makes the model $p(z^S, z^C|x)$ relevant even if skin colored regions are not fully extracted or are not detected at all. Typically, overexposure (close to a bay window) or underexposure (in a corridor) make more uncertain the separation of the skin regions from background. In these situations, all the edges have the same strength in the DT image. More details on the segmentation process can be found in [7].

4.2 Importance functions

4.2.1 Shape cue

We use the face detector introduced by Viola et al. [45] which covers a range of ±45° out-of-plane rotation. It is based on a boosted cascade of Haar-like features to measure relative darkness between eyes and nose/cheek or nose bridge. Let B be the number of detected faces and $\mathbf{p}_i = (u_i, v_i)$, $i = 1, \ldots, B$, the centroid coordinate of each such region. An importance function $\pi(.)$ at location $\mathbf{x} = (u, v)$ follows, as the Gaussian mixture proposal

$$\pi(\mathbf{x}|z^S) = \sum_{i=1}^{B} \frac{1}{B} \mathcal{N}\left(\mathbf{x};\ \mathbf{p}_i, \mathrm{diag}(\sigma_{u_i}^2, \sigma_{v_i}^2)\right), \tag{12}$$

where $\mathcal{N}(\cdot;\ \mu, \Sigma)$ denotes the Gaussian distribution with mean μ and covariance Σ.

4.2.2 Color cue

Skin-colored blobs detection is performed by subsampling the input image prior to grouping the classified skin-like pixels. Then, the importance function $\pi(\mathbf{x}|z^C)$ is defined from the resulting blobs by a Gaussian mixture similar to (12).

4.2.3 Motion cue

The Bhattacharyya distance $D(h_x^M, h_{\mathrm{ref}}^M)$ to the reference motion histogram h_{ref}^M is evaluated on a subset of locations obtained by subsampling the input image and keeping the scale factor fixed. These locations are taken as the nodes of a regular grid. Nodes that satisfy $D^2(h_x^M, h_{\mathrm{ref}}^M) > \tau$ are selected. The importance function $\pi(\mathbf{x}|z^M)$ is a Gaussian mixture (12) centered on the detected locations of high motion activity. Figure 10 reports an importance function derived from the motion cues developed in Fig. 7a and b.

4.2.4 Multi-cues mixture

The importance function $\pi(\cdot)$ can be extended to consider the outputs from any of the M detectors, i.e.

$$\pi(\mathbf{x}|z^1, \ldots, z^M) = \frac{1}{M} \sum_{j=1}^{M} \pi(\mathbf{x}|z^j). \tag{13}$$

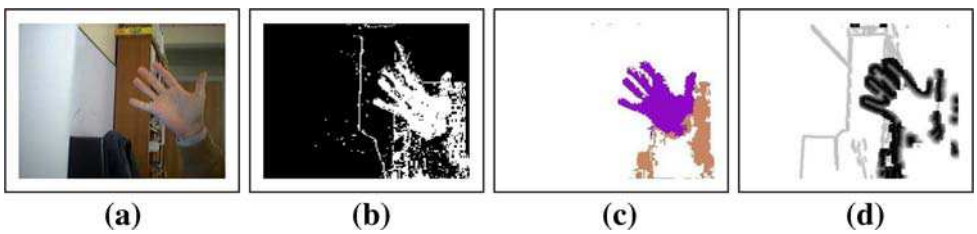


Fig. 9 **a** Input image. **b** Map of the skin color. **c** Skin blobs segmentation. **d** DT image after masking

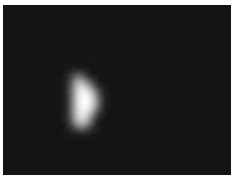


Fig. 10 Motion-based importance function

Figure 11b shows an importance function based on two detectors.

5 People tracking modalities

For our three visual modalities, the aim is to fit the *template* relative to the tracked visitor all along the video stream, through the estimation of its image coordinates (u, v), its scale factor s, as well as, if the template is shape-based, its orientation θ. All these parameters are accounted for in the state vector x_k related to the k-th frame. With regard to the dynamics model $p(x_k|x_{k-1})$, the image motions of observed people are difficult to characterize over time. This weak knowledge is thus formalized by defining the state vector as $\mathbf{x_k} = (u_k, v_k, s_k, \theta_k)^{'}$ and assuming that its entries evolve according to mutually independent random walk models, viz. $p(\mathbf{x_k}|\mathbf{x}_{k-1}) = \mathcal{N}(\mathbf{x_k}\,;\,\mathbf{x}_{k-1}, \Sigma)$, where $\mathcal{N}(\cdot\,;\,\mu, \Sigma)$ terms the Gaussian distribution with mean μ and covariance $\Sigma = \mathrm{diag}(\sigma_u^2, \sigma_v^2, \sigma_s^2, \sigma_\theta^2)$.

5.1 Visual cues evaluation

A preliminary evaluation enables the selection of the most meaningful visual cues associations in terms of discriminative power, precision, time consumption and robustness to artifacts (e.g. clutter or illumination changes), be these cues involved in the importance or measurement functions. Results are computed from a database of over than 400 images acquired from the robot in a wide range of typical conditions. For each database image, a "ground truth" is worked out beforehand regarding the presence/absence and possibly the location of a targeted moving head. The discriminative power of a measurement (resp. importance) function is then computed by comparing the likelihood peaks (resp. the detections) locations with the "true" target location. A peak (resp. detection) in a region of interest around the target is counted as a true positive while outside peaks (resp. detections) are considered as false positives. At last, a false negative occurs when no peaks (resp. no detections) are found inside the region of interest.

Fig. 11 **a** Skin blobs (*blue*) and face (*red*) detectors. **b** Importance function mixing their outputs

5.1.1 Measurement functions

Figure 12 illustrates the average discriminative power of the measurement functions depicted in Sect. 4.1, and Fig. 13 assesses their precisions. Function $S1$ (resp. $S2$) terms the shape-based likelihood $p(z^S|\mathbf{x}_k)$ built upon the similarity distance (4) (resp. (5)). Function $C1$ is relative to the color-based likelihood $p(z_k^C|\mathbf{x}_k)$ relying on the similarity distance (6). The measurement functions $S1C1$ and $S2C1$ fuse shape and color cues by multiplying their likelihoods according to (9). Function $S2C2$ combines both shape and color cues through the skin-colored regions segmentation detailed in [7]. $S1M$ combines shape and motion cues according to (10). Finally, the functions $S1MC1$ and $MC1$ enable the fusion in (9) of all or parts of the three aforementioned cues, respectively shape, motion and color. As Fig. 13 shows, shape cues provide the best accuracy, thus it is important to privilege shape-based measurement functions.

In terms of discriminative power (Fig. 12), using only one cue in the definition of a likelihood is a bad choice. For example, the shape-based likelihoods $S1$, $S2$ are very sensitive to clutter and thus generate a high false positives rate in spite of their good true positives rates. To explain the good results of the color-based likelihood $C1$, it is important to notice that for each image of the database, a color model is computed from the true target location so that no

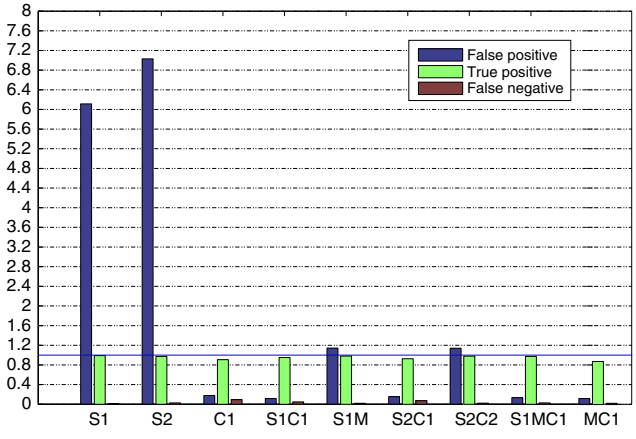


Fig. 12 Average number of detections relative to false positives, true positives and false negatives, for various likelihood functions. The *horizontal red line* depicts the mean target presence rate

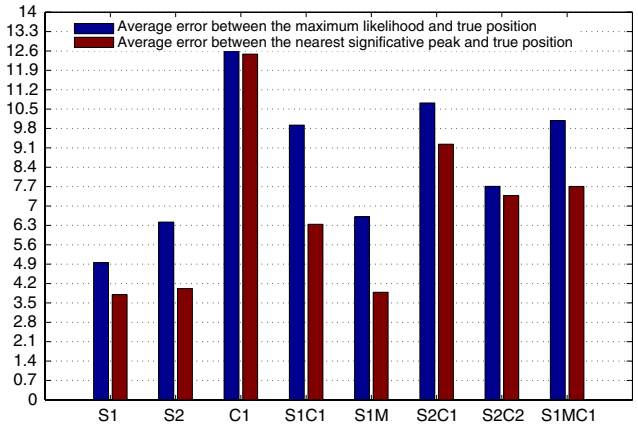


Fig. 13 Average distance between the true target position and (*1*) the maximum likelihood peak, (*2*) the nearest significative peak

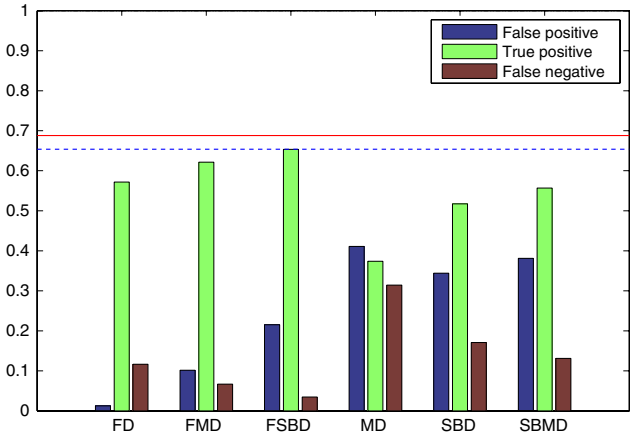


Fig. 15 Average detection rate relative to false positives, true positives and false negatives for various importance functions. The *red* and *blue* lines depict the real true positives rate and the frontal face recognition rate in the database, respectively

color model drift is taken into account in the evaluation. As expected, the more cues are involved in the definition of a likelihood, the higher is its discriminative power. For instance, the motion attribute can penalize motionless contours due to the static background. Fusing color-based likelihoods with shape-based likelihoods eliminates the influence of background contours and makes conveniently colored regions become more likely. Though the fusion strategy $S1MC1$ slightly increases the discriminative power, it is not selected because of its important time consumption. Its running time and these of the other measurement functions are illustrated in Fig. 14.

Similar arguments lead to the rejection of $S2C2$. In fact, the associations of either shape and color cues ($S1C1$, $S2C1$), shape and motion ($S1M$) or color and motion ($MC1$) show the best tradeoff between computational cost and discriminative power. These which enjoy the least time consumption, namely $S2C1$, $S1M$ and $MC1$, have been kept for future evaluations.

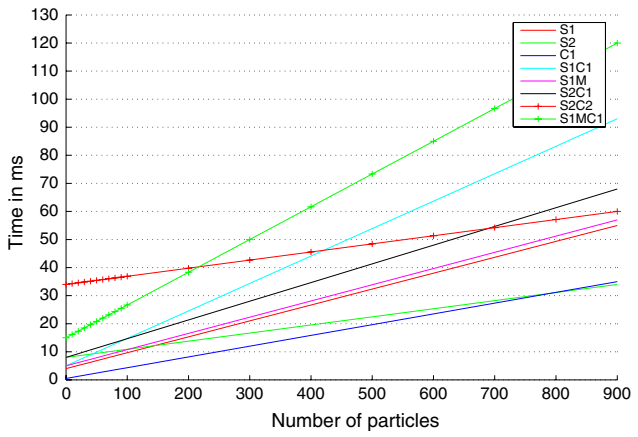


Fig. 14 Average running time per image for various likelihood functions

5.1.2 *Importance functions*

Recall that the importance functions in Sect. 4.2 are relative to face detection (yielding $\pi(\mathbf{x}_k|z_k^S)$ and denoted FD), motion detection (yielding $\pi(\mathbf{x}_k|z_k^M)$ and denoted MD) or skin blob detection (yielding $\pi(\mathbf{x}_k|z_k^C)$ and denoted SBD). The importance functions FMD, $FSBD$, $SBMD$ mix all or parts of the three aforementioned detectors thanks to (13). Figure 15 illustrates the average discriminative power for importance functions associated with a single detector or merging the outputs from several detectors.

Though the FD importance function enjoys a high detection rate and a low false positives rate, it is unfortunately restricted to frontal faces located in the H/R distances interval [0.5 m; 3 m] to permit Haar-like feature extraction. For such short and medium H/R distances[1] ($< 3m$), the multi-cues importance function $FSBD$, which associates FD and SBD into (13), clearly enlarges the spectrum of detected faces as its true positives rate is higher. Moreover, as reported in Fig. 16, the time consumption induced by SBD is negligible compared to the one of FD. The performances are significantly worse for MD, yet this detector is well-suited for long-range H/R distances[2] (>3 m) where shape and skin-color are not sufficiently reliable to use FD-based or SBD-based detectors.

In the selection of an importance function, strategies enjoying a higher true positives rate are preferred. This

[1] i.e. modalities #2 and #3 in Sect. 2.

[2] i.e. modality #1 in Sect. 2.

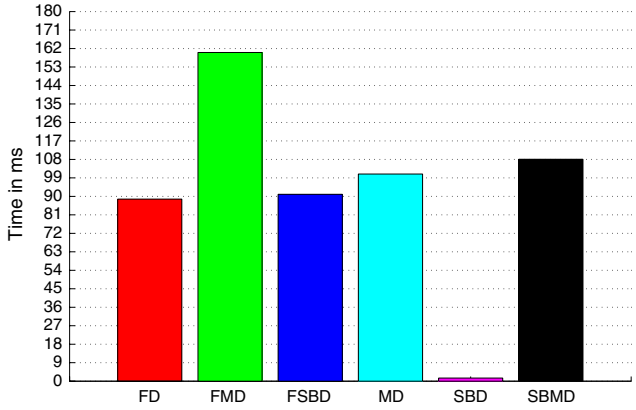


Fig. 16 Average computation time of one image for each importance function. Note that this time is independent of the number of particles involved in the tracking

ensures a sufficient number of particles to be sampled on the target, despite some may be positioned on false detections. Consequently, the importance functions $FSBD$ and MD are considered as candidate elements of the aforementioned visual tracking modalities, to be characterized and evaluated below.

5.2 Particle filtering strategies evaluations

The filtering strategies depicted in Sect. 3 must be examined in order to check which ones best fulfill the requirements of the considered H/R interaction modalities. For the sake of comparison, importance functions rely on dynamics or measurements alone (and are respectively noted DIF for "Dynamics-based Importance Function" and MIF for "Measurement-based Importance Function"), or combine both (and are termed DMIF for "Dynamics and Measurement-based Importance Function"). Further, each modality is evaluated on a database of sequences acquired from the robot in a wide range of typical conditions: cluttered environments, appearance changes or sporadic disappearance of the targeted subject, jumps in his/her dynamics, etc. For each sequence, the mean estimation error with respect to "ground truth", together with the mean failure rate (% of target loss), are computed from several filter runs and particles numbers. The error is computed as the distance (in pixels) between the estimated position and the true position of the object to be tracked. It is important to keep in mind that a failure is registered when this error exceeds a threshold (related to the region of interest), and is followed by a re-initialization of the tracker. Due to space reasons, only a subset of the associated figure plots is shown here. This analysis motivates our choices depicted hereafter for the three visual tracking modalities. The presented results have been obtained on a 3 GHz Pentium IV personal computer.

5.2.1 *Face tracker (proximal interaction)*

This modality involves the state vector $\mathbf{x}_k = (u_k, v_k, s_k, \theta_k)'$. As the robot remains static, both shape and motion cues are combined into the $S1M$ measurement function. The tracker is evaluated in nominal conditions, viz. under no disturbance, as well as against cluttered environments and illumination changes. A typical run is shown in Fig. 17. Figures 18 and 19 plot the tracking failure rate as well as tracking errors averaged over scenarios involving cluttered backgrounds. Dynamics-based Importance Functions lead to a better precision (about 10 pixels) together with a low failure rate, so that detection modules are not necessary in this "easiest" context.

The AUXILIARY_PF strategy shows an higher time consumption than the CONDENSATION algorithm though with no improvement of the approximation of the posterior. The increased computational cost is of course due to the auxiliary sampling step. The fair performance comes from the poorly informative dynamics model, see the end of Sect. 3.3. This is why we opt for a CONDENSATION algorithm, which can run at ≈ 40 Hz for $N = 200$ particles (Fig. 20).

The parameters values of our face tracker are listed in Table 5. The standard deviations of the Gaussian noises entailed in the random walk dynamics are set using classical arguments relating them to the admissible evolution of the template between two consecutive images. The standard deviations of the importance functions come from an offline prior study of the underlying detectors, as was done in [21]. The parameter σ_s involved in the shape-based likelihood (and, in Tables 6, 7, the parameters σ_c, σ_m involved in the color-based and motion-based likelihoods), are defined as follows: first, for each element of an image database, a "ground truth" is determined by manually adjusting the template position, orientation and scale; then the distances involved in the various likelihoods are computed for several perturbations of the template parameters around their "true" values; assuming that these distances are samples of a Gaussian centered on 0 enables the estimation of $\sigma_s, \sigma_c, \sigma_m$. The remaining coefficients, including the number of particles, are selected by trial-and-error, so as to ensure overall good performance of the tracker while limiting its computational cost.

5.2.2 *Upper human body tracker (guidance mission)*

This modality involves the state vector $\mathbf{x}_k = (u_k, v_k, s_k)'$—the orientation θ_k being set to a known constant—as well as two color models $h^c_{\text{ref},1}$,$h^c_{\text{ref},2}$, respectively corresponding to the head and the torso of the guided person, in the measurement function (7). To overcome appearance changes of these ROIs in the video stream, their associated color models are

Fig. 17 Tracking scenario with CONDENSATION over a cluttered background. The *red* template depicts the MMSE estimate

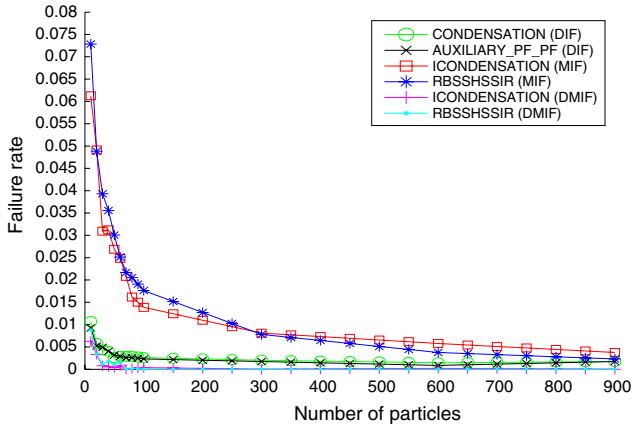


Fig. 18 Average failure rate versus number of particles on sequences involving cluttered background

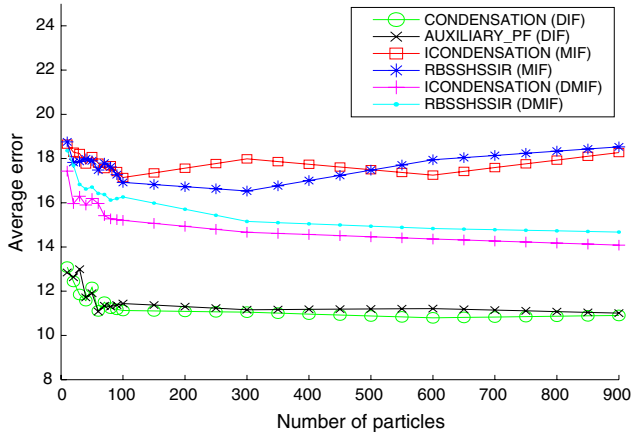


Fig. 19 Tracking errors versus number of particles on sequences involving cluttered background

updated online through a first-order discrete-time filtering process entailing the state estimate i.e.

$$h^c_{\text{ref},k} = (1-\kappa)\cdot h^c_{\text{ref},k-1} + \kappa \cdot h^c_{E[\mathbf{x}_k]}, \tag{14}$$

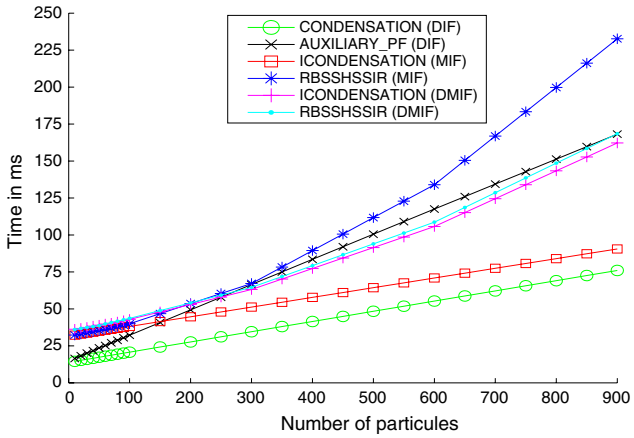


Fig. 20 Average time consumption versus number of particles on all sequences for several strategies

where κ weights the contribution of the mean state histogram $h^c_{E[\mathbf{x}_k]}$ to the target model $h^c_{\text{ref},k-1}$ and index p has been omitted for compactness reasons. Drift and possible subsequent target loss are experienced in any tracker which involves models updating. To avoid this, the particles weighting step considers the likelihood $S2C1$ which fuses, thanks to (9), color distributions cue but also shape cue relatively to the head silhouette (Fig. 4).

Given the H/R interaction distance and the evaluations results in Sect. 4.2, the common importance function of ICONDENSATION and RBSSHSSIR strategies is based on color blobs and face detectors, namely $FSBD$. These proposals permit automatic initialization when persons appear or re-appear in the scene and improve the recovery of deadlocks induced by target loss.

In nominal conditions, all the particle filtering strategies lead to a similar precision and failure rate. Experiments on sequences including appearance or illumination changes, such as the two runs reported in Fig. 21, also show similar results. Indeed, fusing shape and color cues in the

Table 5 Parameter values used in our face tracker

Symbol	Meaning	Value
$(\sigma_u, \sigma_v, \sigma_s, \sigma_\theta)$	Standard deviation of the random walk dynamics noise on the state vector $\mathbf{x_k} = (u_k, v_k, s_k, \theta_k)'$	$(15, 6, 0.01, 0.3)$
σ_s	Standard deviation in likelihood $S1M$ combining shape and motion cues	36
ρ	Penalty in Eq. (10)	0.12
N	Number of particles	150

Table 6 Parameter values used in our upper human body tracker

Symbol	Meaning	Value
(α, β)	Mixture coefficients in the importance function $q(\mathbf{x}_k \vert \mathbf{x}_{k-1}, z_k)$ along Eq. (2)	(0.3, 0.6)
$(\sigma_u, \sigma_v, \sigma_s)$	Standard deviation of the random walk dynamics noise on the state vector $\mathbf{x_k} = (u_k, v_k, s_k)'$	$(11, 6, \sqrt{0.1})$
$(\sigma_{u_i}, \sigma_{v_i})$	Standard deviation in importance function $\pi(\mathbf{x}_k \vert z^S)$ for FD-based detector	(6, 6)
$(\sigma_{u_i}, \sigma_{v_i})$	Standard deviation in importance function $\pi(\mathbf{x}_k \vert z^C)$ for SBD detector	(6, 6)
σ_s	Standard deviation in shape-based likelihood $p(z_k^S \vert \mathbf{x}_k)$	25
N_R	Number of patches in $p(z_k^C \vert \mathbf{x}_k)$	2
σ_c	Standard deviation in color-based likelihood $p(z_k^C \vert \mathbf{x}_k)$	0.03
N_{bi}	Number of color bins per channel involved in $p(z_k^C \vert \mathbf{x}_k)$	32
κ	Coefficients for reference histograms $h^c_{\text{ref},1}$, $h^c_{\text{ref},2}$ update in Eq. (14)	(0.1, 0.05)
N	Number of particles	150

Table 7 Parameter values used in our whole human body tracker

Symbol	Meaning	Value
(α, β)	Mixture coefficients in the importance function $q(\mathbf{x}_k \vert \mathbf{x}_{k-1}, z_k)$ along Eq. (2)	(0.3, 0.6)
$(\sigma_u, \sigma_v, \sigma_s)$	Standard deviation of the random walk dynamics noise on the state vector $\mathbf{x_k} = (u_k, v_k, s_k)'$	$(7, 5, \sqrt{0.1})$
v	Threshold for importance function $\pi(\mathbf{x}_k \vert z_k^M)$	10
$(\sigma_{u_i}, \sigma_{v_i})$	Standard deviation in importance function $\pi(\mathbf{x}_k \vert z^M)$ for MD-based detector	(8, 8)
σ_m	Standard deviation in motion-based likelihood $p(z_k^M \vert \mathbf{x}_k)$	0.2
N'_{bi}	Number of motion bins involved in $p(z_k^M \vert \mathbf{x}_k)$	32
σ_c	Standard deviation in color-based likelihood $p(z_k^C \vert \mathbf{x}_k)$	0.03
N_{bi}	Number of color bins per channel involved in $p(z_k^C \vert \mathbf{x}_k)$	32
N_R	Number of patches in $p(z_k^C \vert \mathbf{x}_k)$	1
κ	Coefficient for reference histogram h^c_{ref} update in Eq. (14)	0.1
N	Number of particles	150

measurement function improves the discriminative power so that in these contexts, a robust tracking can be performed whatever the used particle filtering strategy.

Experiments on sequences including additional sporadic disappearances (due to the limits of the camera field of view) or jumps in the target dynamics highlight the efficiency of ICONDENSATION/RBSSHSSIR strategies in terms of failure rate (Fig. 22). In fact, these two strategies, by drawing some particles according to the output from detection modules, permit automatic initialization and aid recovery from transient tracking failures. In addition, the RBSSHSSIR filter leads to a slightly better precision than ICONDENSATION. This is a consequence of the more efficient association of subparticles sampled from the proposal with plausible predecessors thanks to the intermediate resampling stage (step 6 in Table 3).

Experiments on sequences including spurious detections due to the presence of another non-occluding person in the

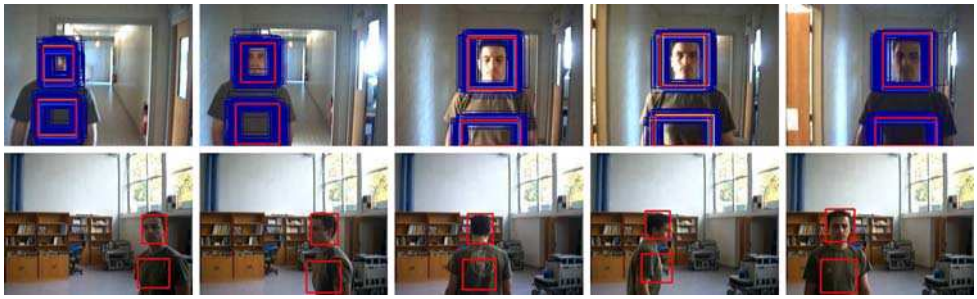


Fig. 21 Tracking scenario involving illumination (*top*) or appearance (*bottom*) changes with DMIF-ICONDENSATION. The *blue* (resp. *red*) *rectangles* depict the particles (resp. the MMSE estimate)

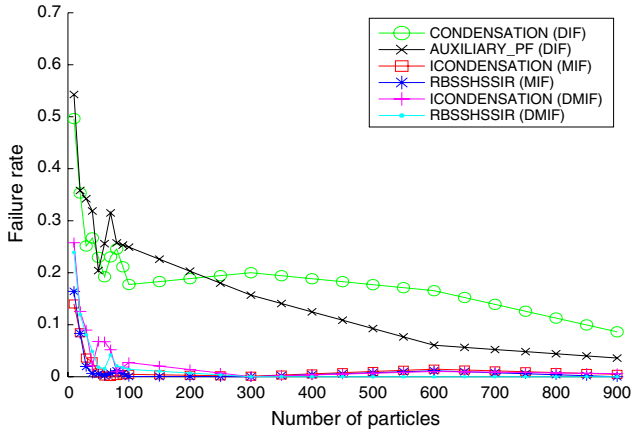


Fig. 22 Average failure rate versus number of particles on sequences including jumps in the target dynamics

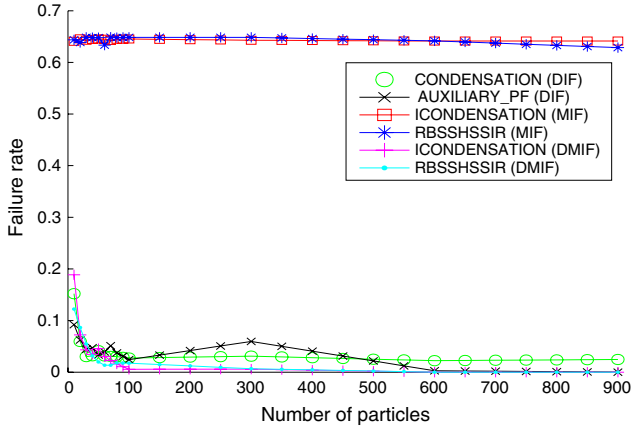


Fig. 23 Average failure rate versus number of particles on sequences including a spurious detection without occlusion

camera field of view, bring out that Measurement-based Importance Functions lead to a worse failure rate (Fig. 23). Conversely, the DMIF strategies ensure a proper tracking thanks to the sampling of some particles from the dynamics.

To illustrate these observations, Fig. 24 shows a tracking run including two persons for MIF-ICONDENSATION (Fig. 24, top) and DMIF-ICONDENSATION (Fig. 24, bottom). In the MIF-ICONDENSATION case, only the non-targeted person is detected so that the importance function

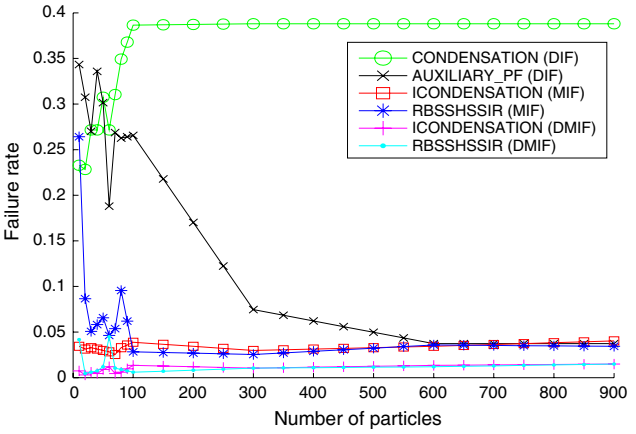


Fig. 25 Average failure rate versus number of particles on sequences including two people occluding each other

draws all the particles on wrong regions, leading to a failure on and after the third frame.

Experiments on sequences which involve two people occluding each other highlight the efficiency of the ICONDENSATION/RBSSHSSIR strategies in terms of failure rate (Fig. 25).

DIF strategies lead to track the person on the foreground (Fig. 26, top), whereas ICONDENSATION/RBSSHSSIR strategies keep locking on the right target throughout the sequence (Fig. 26, bottom) thanks to the sampling of some particles according to the visual detectors outputs, and to the discriminative power of the measurement function.

The above experiments emphasize the necessity of taking into account both the dynamics and the measurements so that the tracking can be robust and efficient enough in all considered scenarios related to the guidance modality. Therefore, the DIF and MIF particle filtering strategies are excluded in this context. Even if DMIF-ICONDENSATION and DMIF-RBSSHSSIR are well suited and have similar time consumption (Fig. 27) for the used number N of particles (between 100 and 200), we finally adopt DMIF-RBSSHSSIR for this guidance modality because of its slightly better performances compared to DMIF-ICONDENSATION. The parameters reported in Table 6 are used in the likelihoods, proposal and state dynamics involved in our upper human body tracker.

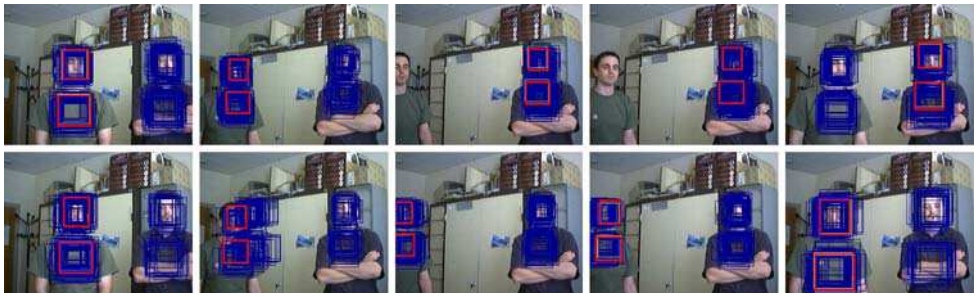


Fig. 24 Tracking scenario involving two people with MIF-ICONDENSATION (*top*) and DMIF-ICONDENSATION (*bottom*). On the *third top frame* the targeted person is not detected while the undesired person remains detected

Fig. 26 Tracking scenario involving occlusions with CONDENSATION (*top*) and DMIF-RBSSHSSIR (*bottom*)

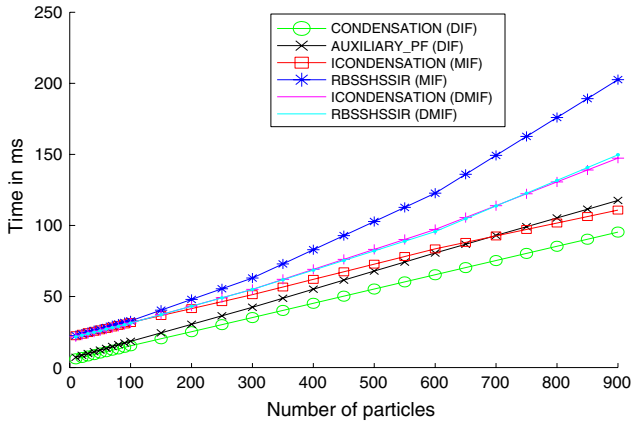


Fig. 27 Average time consumption versus number of particles on all sequences for several strategies

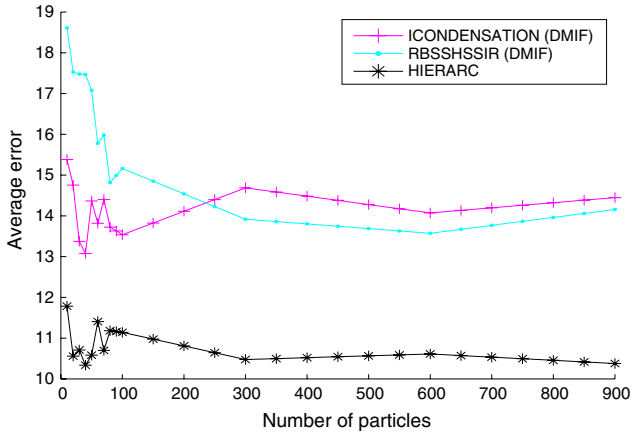


Fig. 28 Tracking errors versus number of particles in nominal conditions

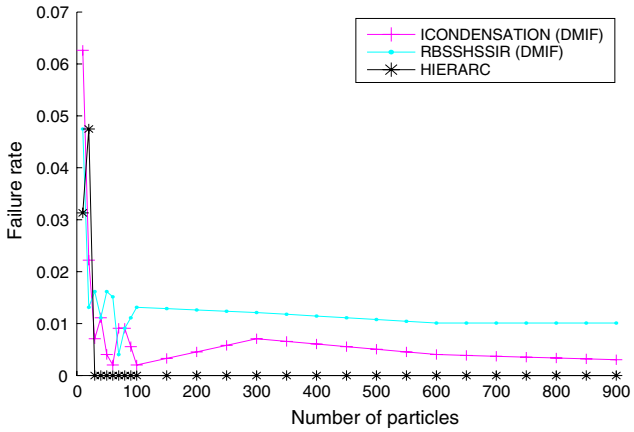


Fig. 29 Average failure rate versus number of particles in nominal conditions

5.2.3 Person tracker (search for interaction)

As was the case in the above section, the state vector has the form $\mathbf{x}_k = (u_k, v_k, s_k)'$. Color and motion cues are fused into the global measurement function (9) as the robot is supposed motionless. The selected importance function MD is defined from the motion detector output. Relying on the conclusions concerning the guidance modality, only DMIF-ICONDENSATION and DMIF-RBSSHSSIR are evaluated. As the Hierarchical Particle Filter (HIERARCHICAL_PF) defined in [33] constitutes an alternative to these strategies, it is also assessed. Thanks to its intermediate sampling (step 6 in Table 4), which enables the particles cloud to remain more focused on the target, it results in a significant decrease of the tracking error under nominal conditions, as illustrated in Fig. 28. In this helpful case, a slightly better failure rate is also observed as shown in Fig. 29.

Experiments on sequences including a full occlusion of the moving target by a static object (Fig. 32) highlight the efficiency of DMIF-ICONDENSATION/DMIF-RBSSHSSIR strategies in terms of of failure rate compared to the HIERARCHICAL_PF strategy (Fig. 30). Though some particles are sampled on the targeted person by the motion-based importance function (MD) as soon as he/she reappears after an occlusion, the HIERARCHICAL_PF strategy fails (Figs. 31, 32). Indeed, as these particles lie in the tails of the dynamics pdf, they are affected small weights and thus get eliminated during the first resampling step of the algorithm (step 6 in Table 4). Meanwhile, the other filtering strategies which rely on Dynamics and Measurement based importance functions can perform the tracking.

Similar conclusions hold when the target is motionless and subject to occlusion (Figs. 33, 34). This can again be explained by the action of its first resampling step which concentrates particles on high motion activity regions. Figures 35

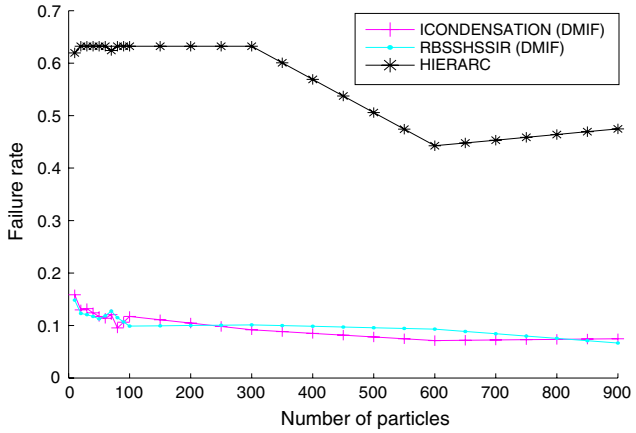


Fig. 30 Average failure rate versus number of particles on sequences including an occlusion by a static object

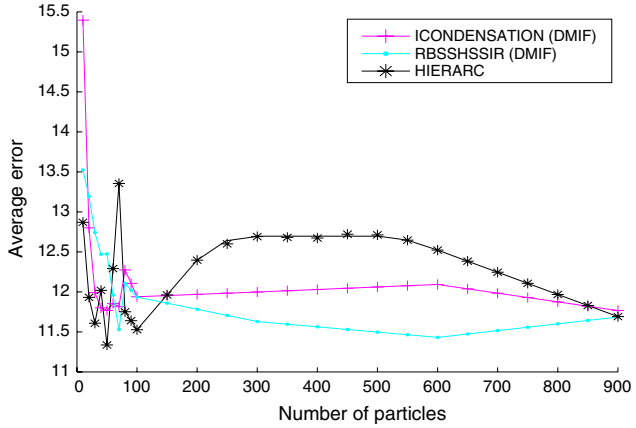


Fig. 31 Tracking errors versus number of particles on sequences including an occlusion by a static object

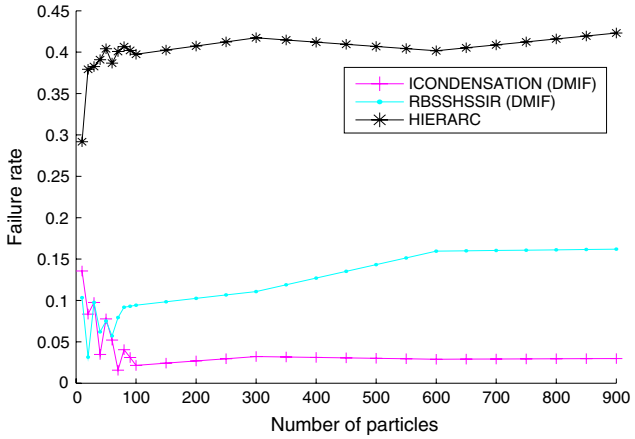


Fig. 33 Average failure rate versus number of particles on sequences including target occlusions

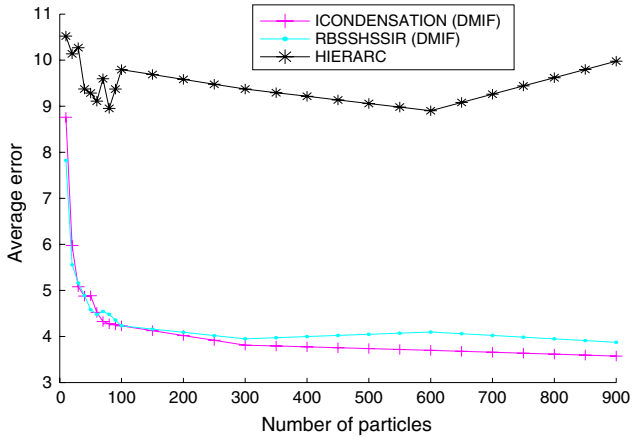


Fig. 34 Tracking errors versus number of particles on sequences including target occlusions

and 36 present two tracking scenarios involving several persons with mutual occlusions. In the first scenario, the HIERARCHICAL_PF tracker locks onto the undesired person, because only regions of high motion activity which are in the modes of the system dynamics pdf are explored by its first step. Regions corresponding to the target, even if they do comply with the color model, are thus discarded during the resampling procedure. In contrast, the RBSSHSSIR tracker which doesn't dissociate motion and color cues, keeps locking on the targeted person. The second scenario leads to the same observations and confirms the RBSSHSSIR efficiency. The two filters DMIF-ICONDENSATION/DMIF-RBSSHSSIR are well-suited to this modality. As robustness is preferred to precision for our application, we finally opt for the DMIF-RBSSHSSIR algorithm. The fixed parameters involved in the likelihoods, proposal and state dynamics of our human body tracker are given in Table 7.

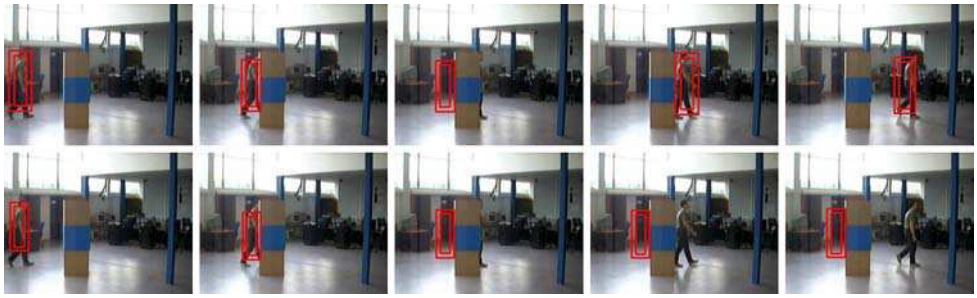


Fig. 32 Tracking scenario involving full occlusion by a static object with DMIF-RBSSHSSIR (*top*) and HIERARCHICAL_PF (*bottom*)

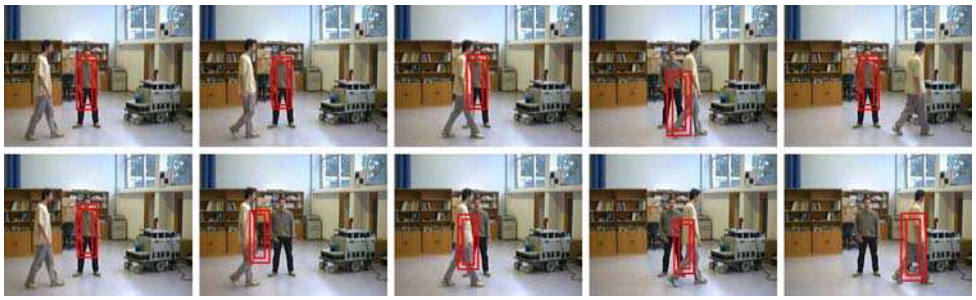


Fig. 35 Tracking scenario involving occlusion of the motionless target by another person crossing the field of view with DMIF-RBSSHSSIR (*top*) and HIERARCHICAL_PF (*bottom*)

Fig. 36 A scenario involving persistent occlusions due to persons. Tracker based on a DMIF into the RBSSHSSIR algorithm

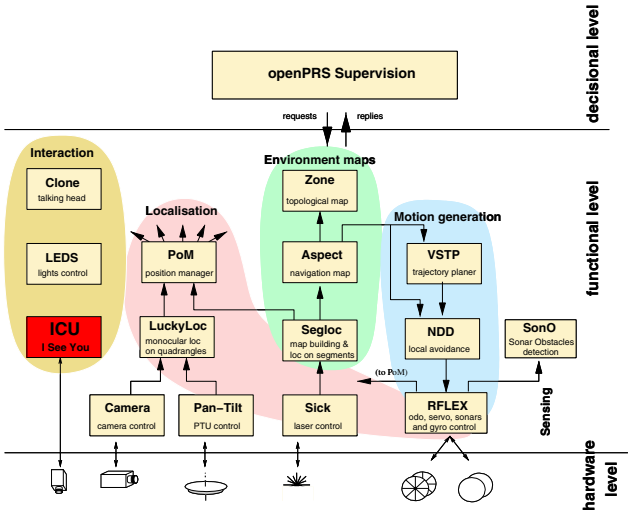


Fig. 37 Rackham's software architecture

6 Integration on Rackham robot

6.1 Outline of the overall software architecture

The above visual functions were embedded on the Rackham robot. To this aim, Rackham is fitted with the "LAAS" software architecture introduced in Fig. 37 and thoroughly presented in [1].

On the top of the hardware (sensors and effectors), the *functional level* encapsulates all the robot's action and perception capabilities into controllable communicating modules, operating at very strong temporal constraints. The *executive level* activates these modules, controls the embedded functions, and coordinates the services depending on the task high-level requirements. Finally, the upper *decision level* copes with task planning and supervision, while remaining reactive to events from the execution control level.

In addition to functional modules dedicated to exteroceptive sensors handling, e.g. cameras, laser and ultrasonic telemeters,…, low-level servo algorithms, elementary navigation functions, etc.[5], a module named ICU—for "I see you"—has been designed which encapsulates all the aforementioned tracking modalities. It is depicted below.

6.2 Considerations about the ICU software architecture

The C++ implementation of the module ICU is integrated in the "LAAS" architecture using a C/C++ interfacing scheme. It enjoys a high modularity thanks to C++ abstract classes and template implementations. This way, virtually any tracker can be implemented by selecting its components from predefined libraries related to particle filtering strategies, state evolution models, and measurement/importance functions. For more flexibility, specific components can be defined and integrated directly.

ICU sequentially invokes the tracking components through its processing pipe, as illustrated in Figure 38. So, the functions shared by several trackers running in parallel are processed only once.

Section 6.3 enumerates all the visual functions provided by the module ICU, which not limited to tracking. Section 6.4 details the way how they are entailed in the tour-guide scenario, and discusses the automatic switching between trackers.

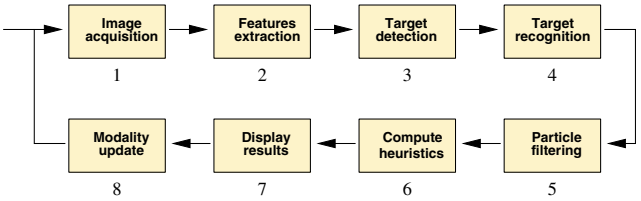


Fig. 38 Sequencing the module ICU

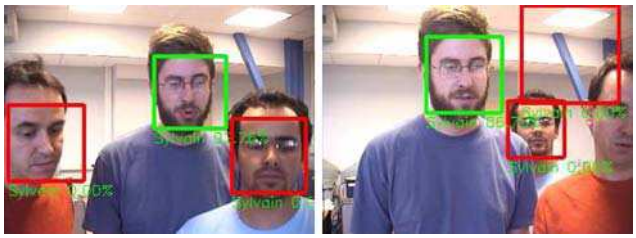


Fig. 39 Snapshots of detected (*red*)/recognized (*green*) faces with associated probabilities. The target is named Sylvain in this example

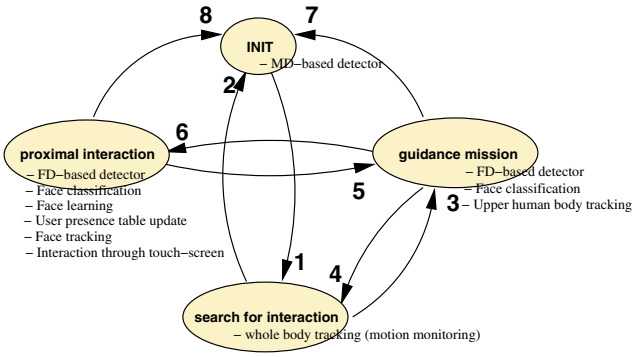


Fig. 40 Transitions between tracking modalities

6.3 Visual functions provided by the module ICU

These can be organized into three broad categories:

1. Functions related to human body/limbs detection: Independently from the tracking loop, FD-based or MD-based detectors (see Sect. 4.2) can be invoked depending on the current H/R distance and the scenario status.
2. Functions related to user face recognition: The face recognition process underlies the following functions
   - a *face learning function* based on the FD-based detector in order to train the classifier.
   - a *face classification function* based on these training samples and eigenfaces representation [38]. The face recognition probability associated with each detected face can be integrated both in the face and upper human body trackers. Some recognition snapshots are reported in Fig. 39. Further details can be found in [16].
   - a *user presence function* updates a presence table of the 30 previously learned robot users. The table update is similar to a FIFO stack, i.e. the oldest user added in the table is handled next.
3. Functions related to user tracking: These are
   - the *three tracking functions* characterized and evaluated in Sect. 5. Recall that they have been designed so as to best suit to the interaction modalities of Sect. 2.
   - an *estimator of the H/R distance* of the targeted person from the scale of the updated template during the tracking.

   The robot activates these functions depending on the current H/R distance, user identification and scenario status. The next subsection details the way how they are scheduled.

6.4 Heuristic-based switching between trackers

A finite-state automaton can be defined from the tour-guide scenario outlined in Sect. 2, as illustrated in Fig. 40. Its four states are respectively associated to the INIT mode and to the three aforementioned interaction modalities. Two heuristics relying on the current H/R distance and the face recognition status allow to characterize most of the transitions in the graph. Practically, outcomes from the face classifier and H/R distance functions are filtered on a sliding temporal window W_t of about 20 samples. The robot in INIT mode invokes the motion-based detector (MD), so that any visitor entering the exhibition initializes the whole body tracking (arrow 1). The robot assumes that the visitor is willing to interact when he/she has come closer and has got detected by the frontal face FD-detector over W_t. If so, the upper human body tracking mode is launched (arrow 3). If the user H/R distance keeps decreasing to less than $1m$ and his/her face remains detected/recognized, a "proximal interaction" begins, entailing the face tracker (arrow 6). The face learning function and the human presence table update function are possibly invoked if the user is unknown. When starting the "guidance mission", the robot switches to the upper human body tracker (arrow 5). Temporary target loss are notified when the face classifier fails for more than 70% of the 20 images composing W_t. Re-identification of the guided visitor in the next W_t is required in order to resume the ongoing mission. Finally, the robot returns in INIT mode when: (1) no moving blobs are detected, (2) the current user hasn't been recognized over W_t, (3) the end mission is signified by the robot (arrows 2, 7 and 8).

Thanks to its efficient modular implementation, all the ICU functions can be executed in real time on our robot. Experiments show their complementary and efficiency in cluttered scenes.

7 Conclusion

This paper has introduced mechanisms for visual data fusion/combination within particle filtering to develop people trackers from a single color camera mounted on a mobile robot. Most particle filtering techniques to single-target tracking

have been surveyed and tested. A first contribution concerns visual data fusion for the considered robotics scenarios, a context which has fairly seldom exploited particle filtering based solutions. The most persistent cues are used in the particles weighting stage. The others, logically intermittent, permit automatic initialization and aid recovery from transient tracking failures. Mixing these cues both into the importance and measurement functions of the underlying estimation scheme, can help trackers work under a wide range of conditions encountered by our Rackham robot during its displacements. A second contribution relates to the evaluation of dedicated particle filtering strategies in order to check which people trackers, regarding visual cues and algorithms associations, best fulfill the requirements of the considered scenarios. Let us point out that few studies comparing the efficiency of so many filtering strategies had been carried out in the literature before. A third contribution concerns the integration of all these trackers on a mobile platform, whose deployment in public areas has highlighted the relevance and the complementarity of our visual modalities. To our knowledge, quite few mature robotic systems enjoy such advanced capabilities of human perception.

Several directions are currently investigated. First, we study how to fuse other information such as laser or sound cues. The sound cue would not just contribute to the localization in the image plane, but will also endow the tracker with the ability to switch its focus between speakers. A next issue will concern the incorporation of appropriate degrees of adaptivity into our multiple cues based likelihood models depending on the target properties changes or the current viewing conditions [40]. In addition, our tracking modalities will be made much more active. Zooming will be used to actively adapt the focal length with respect to the H/R distance and to the current active visual modalities. Finally, the tracker will be enhanced so as to track multiple persons simultaneously [27,42].

Acknowledgments The work described in this paper was partially conducted within the EU Integrated Project COGNIRON (The Cognitive Companion) and the EU STREP Project CommRob (Advanced Behavior and High-Level Multimodal Communication with and among Robots), funded by the European Commission Division FP6-IST Future and Emerging Technologies under Contracts FP6-IST-002020 and FP6-IST-045441.

References

1. Alami, R., Chatila, R., Fleury, S., Ingrand, F.: An architecture for autonomy. Int. J. Robot. Res. **17**(4), 315–337 (1998)
2. Andrieu, C., Davy, M., Doucet, A.: Improved auxiliary particle filtering: Application to timevarying spectral analysis. In: IEEE Wk. on Statistical Signal Processing, pp. 309–312. Singapore (2001)
3. Arulampalam, S., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. IEEE Trans. Signal Process. **50**(2), 174–188 (2002)
4. Avidan, S.: Support vector tracking. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR'01), pp. 184–191. Kauai, HI (2001)
5. Bailly, G., Brèthes, L., Chatila, R., Clodic, A., Crowley, J., Danès, P., Elisei, F., Fleury, S., Herrb, M., Lerasle, F., Menezes, P., Alami, R.: HR+ : towards an interactive autonomous robot. In: Journées ROBEA, pp. 39–45. Montpellier, France (2005)
6. Bichot, E., Mascarilla, L., Courtellemont, P.: Particle filtering based on motion and color information. IEEE Trans. Information Sci. Appl. **2**, 220–227 (2005)
7. Brèthes, L., Menezes, P., Lerasle, F., Briot, M.: Face tracking and hand gesture recognition for human-robot interaction. In: IEEE Int. Conf. on Robotics and Automation (ICRA'04), pp. 1901–1906. New Orleans, LA (2004)
8. Bretzner, L., Laptev, I., Lindeberg, T.: Hand gesture using multi-scale colour features, hierarchical models and particle filtering. In: IEEE Int. Conf. on Automatic Face and Gesture Recognition (FGR'02), pp. 405–410. Washington D.C. (2002)
9. Brèthes, L.: Suivi visuel par filtrage particulaire. Application à l'interaction homme-robot. Ph.D. thesis, Université Paul Sabatier, LAAS-CNRS, Toulouse (2006)
10. Bullock, D., Zelek, J.: Real-time tracking for visual interface applications in cluttered and occluding situations. J. Vis. Image Comput. **22**, 1083–1091 (2004)
11. Chen, H., Liu, T.: Trust-region methods for real-time tracking. In: IEEE Int. Conf. on Computer Vision (ICCV'01), vol. 2, pp. 717–722. Vancouver, Canada (2001)
12. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. In: IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 25, pp. 564–575 (2003)
13. Doucet, A., De Freitas, N., Gordon, N.J.: Sequential Monte Carlo Methods in Practice. Series Statistics For Engineering and Information Science. Springer, New York (2001)
14. Doucet, A., Godsill, S.J., Andrieu, C.: On sequential monte carlo sampling methods for bayesian filtering. Stat. Comput. **10**(3), 197–208 (2000)
15. Gavrila, D.M.: The visual analysis of human movement : a survey. Comput. Vis. Image Underst. **1**, 82–98 (1999)
16. Germa, T., Brèthes, L., Lerasle, F., Simon, T.: Data fusion and eigenface based tracking dedicated to a tour-guide robot. In: Int. Conf. on Vision Systems (ICVS'07). Bielefeld, Germany (2007)
17. Giebel, J., Gavrila, D.M., Schnorr, C.: A bayesian framework for multi-cue 3D object. In: Eur. Conf. on Computer Vision (ECCV'04). Prague, Czech Republic (2004)
18. Haritaoglu, I., Harwood, D., Davis, L.: W4 : Real-time surveillance of people and their activities. IEEE Trans. Pattern Anal. Mach. Intell. **8**(22), 809–830 (2000)
19. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEE Trans. Syst. Man Cybernet. **34**(3), 334–352 (2004)
20. Isard, M., Blake, A.: Condensation—conditional density propagation for visual tracking. Int. J. Comput. Vis. **29**(1), 5–28 (1998)
21. Isard, M., Blake, A.: Icondensation: Unifying low-level and high-level tracking in a stochastic framework. In: Eur. Conf. On Computer Vision (ECCV'98), pp. 893–908. London, UK (1998)
22. Isard, M.A., Blake, A.: A mixed-state condensation tracker with automatic model-switching. In: IEEE Int. Conf. on Computer Vision (ICCV'98), pp. 107–112. Bombay, India (1998)
23. Jones, M., Rehg, J.: Color detection. Tech. rep., Compaq Cambridge Research Lab (1998)
24. Julier, S., Uhlmann, J.: A general method for approximating nonlinear transformations of probability distributions. Tech. rep., RRG, Dept. of Engineering Science, University of Oxford (1994)
25. Kitagawa, G.: Monte carlo filter and smoother for non-gaussian nonlinear state space models. J. Comput. Graph. Stat. **5**(1), 1–25 (1996)

26. Li, P., Zhang, T.: Visual contour based on sequential importance sampling/resampling algorithm. In: IEEE Int. Conf. on Pattern Recognition (ICPR'02), pp. 564–568. Quebec, Canada (2002)
27. MacCormick, J., Blake, A.: A probabilistic exclusion principle for tracking multiple objects. Int. J. Comput. Vis. **39**(1), 57–71 (2000)
28. MacCormick, J., Isard, M.: Partitioned sampling, articulated objects, and interface-quality hand tracking. In: Eur. Conf. on Computer Vision (ECCV'00), pp. 3–19. Springer, London, UK (2000)
29. Moeslund, T., Granum, E.: A survey on computer vision-based human motion capture. Comput. Vis. Image Underst. **81**, 231–268 (2001)
30. Nummiaro, K., Koller-Meier, E., Gool, L.V.: An adaptative color-based particle filter. J. Image Vis. Comput. **21**, 90–110 (2003)
31. Pitt, M., Shephard, N.: Filtering via simulation: auxiliary particle filters. J. Am. Stat. Assoc. **94**(446), 590–599 (1999)
32. Pérez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-based probabilistic tracking. In: Eur. Conf. on Computer Vision (ECCV'02), pp. 661–675. Berlin, Germany (2002)
33. Pérez, P., Vermaak, J., Blake, A.: Data fusion for visual tracking with particles. Proc. IEEE **92**(3), 495–513 (2004)
34. Rui, Y., Chen, Y.: Better proposal distributions: Object tracking using unscented particle filter. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR'01), pp. 786–793. Kauai, HI (2001)
35. Schwerdt, K., Crowley, J.L.: Robust face tracking using color. In: Int. Conf. on Face and Gesture Recognition (FGR'00), pp. 90–95. Grenoble, France (2000)
36. Thayananthan, A., Stenger, B., Torr, P., Cipolla, R.: Learning a kinematic prior for tree-based filtering. In: British Machine Vision Conf. (BMVC'03), vol. 2, pp. 589–598. Norwich, UK (2003)
37. Torma, P., Szepesvári, C.: Sequential importance sampling for visual tracking reconsidered. In: AI and Statistics, pp. 198–205 (2003)
38. Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR'91), pp. 586–591. Maui, HI (1991)
39. Van Der Merwe, R., De Freitas, N., Doucet, A., Wan, E.: The unscented particle filter. In: Advances in Neural Information Processing Systems, vol. 13 (2001)
40. Vermaak, J., Andrieu, C., Doucet, A., Godsill, S.: Particle methods for bayesian modeling and enhancement of speech signals. IEEE Trans. Speech Audio Process. **10**(3), 173–185 (2002)
41. Vermaak, J., Blake, A.: A nonlinear filtering for speaker tracking in noisy and reverberant environments. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'00). Istanbul, Turkey (2000)
42. Vermaak, J., Doucet, A., Pérez, P.: Maintaining multi-modality through mixture tracking. In: IEEE Int. Conf. on Computer Vision (ICCV'03). Nice, France (2003)
43. Vermaak, J., Pérez, P., Gangnet, M., Blake, A.: Towards improved observation models for visual tracking: Selective adaptation. In: Eur. Conf. on Computer Vision (ECCV'02), pp. 645–660. Berlin, Germany (2002)
44. Vezhnevets, V., Sazonov, V., Andreeva, A.: A survey on pixel-based skin color detection techniques. In: Proc. Graphicon-2003 pp. 85–92 (2003)
45. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR'01). Kauai, HI (2001)
46. Wachter, S., Nagel, H.: Tracking persons in monocular image sequences. Comput. Vis. Image Underst. **74**(3), 174–192 (1999)
47. Wu, Y., Huang, T.: A co-inference approach to robust visual tracking. In: IEEE Int. Conf. on Computer Vision (ICCV'01), vol. 2, pp. 26–33. Vancouver, Canada (2001)

2008 IEEE/RSJ International Conference on Intelligent Robots and Systems
Acropolis Convention Center
Nice, France, Sept, 22-26, 2008

Mutual assistance between speech and vision for human-robot interaction

Brice Burger[†‡¶], Frédéric Lerasle[†¶], Isabelle Ferrané[‡¶], Aurélie Clodic[†]
[†] CNRS; LAAS; 7 avenue du Colonel Roche, 31077 Toulouse Cedex, France
[‡] IRIT; 118 route de Narbonne, 31062 Toulouse Cedex, France
[¶] Université de Toulouse; UPS; LAAS-CNRS : F-31077 Toulouse, France
E-mail: {bburger, lerasle, aclodic}@laas.fr, ferrane@irit.fr

Abstract— **Among the cognitive abilities a robot companion must be endowed with, human perception and speech understanding are both fundamental in the context of multimodal human-robot interaction. First, we propose a multiple object visual tracker which is interactively distributed and dedicated to two-handed gestures and head location in 3D. An on-board speech understanding system is also developed in order to process deictic and anaphoric utterances. Characteristics and performances for each of the two components are presented. Finally, integration and experiments on a robot companion highlight the relevance and complementarity of our multimodal interface. Outlook to future work is finally discussed.**

keywords: multiple object tracking, speech understanding, multimodal interaction, assistance robotic.

I. Introduction and framework

As the number of senior citizens increases, more research efforts have been made to develop socially interactive household robots. This field of robotics is a deep challenge because robots moving out of laboratories have to gain more social skills to improve natural peer-to-peer interaction with a novice user in his/her daily life. As speech is the most prominent communication channel for humans, a considerable number of robot assistants embed advanced speech recognition system [4]. This is not enough to realize a user-friendly interface as we, humans, omit, abbreviate and underspecify things in our utterances, that are supposed to be obtained by vision. Only few research work addresses the development of such appropriate multimodal interfaces [10]. On one hand, the mutual assistance between the speech and vision capabilities of the robot, permits to specify parameters related to person/object IDs or location references in verbal statements. On the other hand, fusing auditive and visual features are supposed to be more robust to noisy/cluttered environments than using one single feature.

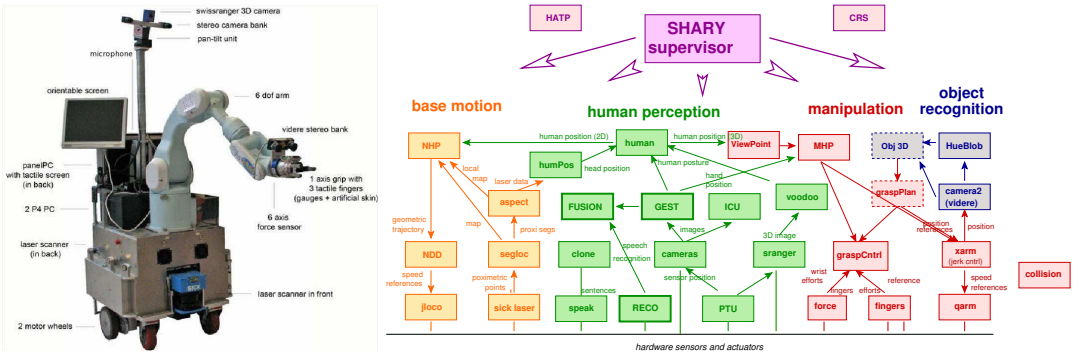
To complete/verify the message conveyed by the verbal communication channel, these interfaces consider vision techniques in order to: (i) characterize the robot surroundings *i.e.* places [6] or objects [13], (ii) perceive the human user [6], [9], [11], [13] *i.e.* his/her gestures and nonverbal reactive body motions. Besides image-based approaches [6], [9], [13], 3D positions of the user's head and hands are particularly useful, in combination with speech recognition, to specify parameters of location in verbal statements *e.g. "look here"* or *"give this object to me"*. Following [7], a first issue concerns the design of body and gesture tracker suited for gesture interpretation. This tracker has been extended in three ways. First, we propose an interactively distributed multiple object visual tracker dedicated to two-handed gestures and head location in 3D. Secondly, the tracker has been endowed with visual data fusion and automatic re-initialization. All this makes our tracker work under a wide range of viewing conditions and aid recovery from transient tracking failures due to the robot's motion or temporarily loss of observability when performing gestures. Finally, their combination with deictic and anaphoric utterances have been tested in household robotics operation with promising results. Here, gesture is used as an essential complementary information. Gesture detection could also help to reinforce communication in case of speech recognition errors.

The paper is organized as follows. Section II describes our robot companion Jido, outlines its embedded multimodal interfaces, and the target scenario we address. Section III presents the binocular tracking of the user's head and two-handed gestures in order to interpret symbolic and deictic gestures thanks to Hidden Markov Models. Section IV depicts the system dedicated to verbal communication between our robot companion and humans. Section V presents robotic experiments involving these two components. Last, section VI summarizes our contributions and discusses future extensions.

II. Jido and its target scenario

Our multimodal interface is embedded on a robot companion called Jido which is equipped with a 6-DOF arm, a pan-tilt stereo system at the top of a mast, two laser scanners (Figure 1(a)) while the human wears a wireless headset microphone. All these devices enable Jido to act as a robot assistant as it is endowed with basic functions enabling to: (i) navigate in its environment, (ii) recognize and grasp objects, (iii) detect, localize humans in its vicinity, (iv) interpret speech utterances and some gestures. All the embedded functions are managed thanks to the "LAAS" layered software architecture (Figure 1(b)) and detailed in [2].

Besides environment perception abilities, the multimodal interface has been undertaken within the demonstration scenario. This is a household situation in which Jido executes human-friendly collaborative tasks (coordinated displacements and object exchange) ordered by its disabled user. Given both verbal and gesture commands, this person



(a) (b)

Fig. 1. Jido and its software architecture.

is allowed to make the robot change its position in the environment, marks some objects the robot must catch and carry, etc. A typical set of commands is for instance: *"come on", "take this bottle on the table", "bring it to me", "go over there"*. Given this scenario, the paper focuses on the multimodal components (Figure 1(b)), namely GEST, RECO and FUSION respectively for the visual perception of the user, the speech interpretation and the fusion from user and object perception with speech interpretation. The functionalities encapsulated in these modules are presented in the following sections.

III. Visual Perception of the Robot User

A. 3D tracking of heads and hands

Our system dedicated to the visual perception of the robot user includes 3D face and two-hand tracking. Particle filters (PF) constitute one of the most powerful framework for view-based multi-tracking purpose [12]. In the robotic context, their popularity stems from their simplicity, modeling flexibility, and ease of fusion of diverse kinds of measurements. Two main classes of multiple object tracking (MOT) can be considered. While the former, widely accepted in the Vision community, exploits a single joint state representation which concatenates all of the targets' states together, the latter uses distributed filters, namely one filter per target. The main drawback of the centralized approach remains the number of required particles which increases exponentially with the state-space dimensionality. The distributed approach, which is the one we have chosen, suffers from the well-known "error merge" and "labelling" problems when targets undergo partial or complete occlusion. In the vein of [12], we develop an interactively distributed MOT (IDMOT) framework which is depicted in Table I. The aim is to approximate the probability density function $p(\mathbf{x}_t^i|z_{1:t})$ of the state vector $\mathbf{x}_t^i$ for body part i at time t given the set of measurements $z_{1:t}$ and the cloud of "particles" indexed by n with likelihood -or "weight"- $\omega_t^{i,n}$. When targets do not interact on each other, the approach performs like multiple independent trackers. When they are in close proximity, magnetic and inertia likelihoods (annotated $\varphi_1(.)$ and $\varphi_2(.)$) are added in each filter to handle the aforementioned problems (see [12] for more details). Our IDMOT particle filter is improved and extended in three ways. First, the conventional CONDENSATION [5] strategy is replaced by the genuine ICONDENSATION one whose importance function $q(.)$ in step 3 permits automatic (re)-initialization when the targeted human body parts appear or re-appear in the scene. The principle consists in sampling the particle according to visual detectors $\pi(.)$, dynamics $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, and the prior p_0 so that, with $\alpha, \beta \in [0;1]$

$$q(\mathbf{x}_t^{i,n}|\mathbf{x}_{t-1}^{i,n}, z_t^i) = \alpha\pi(\mathbf{x}_t^{i,n}|z_t^i) + (1-\alpha)p(\mathbf{x}_t^i|\mathbf{x}_{t-1}^{i,n}). \quad (1)$$

Secondly, the IDMOT particle filter, devoted initially to the image-based tracking of multiple objects or people, is here extended to estimate the 3D pose of multiple deformable body parts of a single person. The third line of investigation concerns data fusion as our observation model is based on a robust and probabilistically motivated integration of multiple cues. Fusing 3D and 2D (image-based) information from the video stream of a stereo head - with cameras mounted on a mobile robot - enables to benefit both from reconstruction-based and appearance-based approaches. The aim of our IDMOT approach, named IIDMOT, is to fit the projections all along the video stream of a sphere and two deformable ellipsoids (resp. for the head and the two hands), through the estimation of the 3D location $\mathcal{X} = (X, Y, Z)'$, the orientation $\Theta = (\theta_x, \theta_y, \theta_z)'$, and the axis length [1] $\Sigma = (\sigma_x, \sigma_y, \sigma_z)'$ for ellipsoids. All these parameters are accounted for in the state vector $\mathbf{x}_t^i$ related to target i for the t-th frame. With regard to the dynamics model $p(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i)$, the 3D motions of observed gestures are difficult to characterize over time. This weak knowledge is formalized by defining the state vector as $\mathbf{x}_t^i = [\mathcal{X}_t, \Theta_t, \Sigma_t]'$ for each hand and assuming that its entries evolve according to mutually independent random walk models, viz. $p(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i) = \mathcal{N}(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i, \Lambda)$, where $\mathcal{N}(.|\mu, \Lambda)$ is a Gaussian distribution in 3D with mean μ and covariance Λ being determined *a priori*. Our importance function $q(.)$ followed by some consideration about the measurement function $p(z_t^i|\mathbf{x}_t^i)$ are given here below. Recall that α percent of the particles are sampled from detector $\pi(.)$ (equation (1)). These are also drawn from Gaussian distribution for head or hand configuration deduced from skin color blob segmentation in the stereo video stream.

[1] To take into account the hand orientation in 3D.

TABLE I

OUR IIDMOT ALGORITHM.

IF $t = 0$, **THEN** Draw $\mathbf{x}_0^{i,1}, \ldots, \mathbf{x}_0^{i,j}, \ldots, \mathbf{x}_0^{i,N}$ i.i.d. according to $p(\mathbf{x}_0^i)$, and set $w_0^{i,n} = \frac{1}{N}$ **END IF**

IF $t \geq 1$ **THEN** $\{—[\{\mathbf{x}_{t-1}^{i,n}, w_{t-1}^{i,n}\}]_{n=1}^{N}$ being a particle description of $p(\mathbf{x}_{t-1}^i | z_{1:t-1}^i)—\}$

"Propagate" the particle $\{\mathbf{x}_{t-1}^{i,n}\}_{n=1}^{N}$ by independently sampling $\mathbf{x}_t^{i,n} \sim q(\mathbf{x}_t^i | x_{t-1}^{i,n}, z_t^i)$

Update the weight $\{w_t^{i,n}\}_{n=1}^{N}$ associated to $\{\mathbf{x}_t^{i,n}\}_{n=1}^{N}$ according to the formula $w_t^{i,n} \propto w_{t-1}^{i,n} \frac{p(z_t^i | \mathbf{x}_t^{i,n}) p(\mathbf{x}_t^{i,n} | \mathbf{x}_{t-1}^{i,n})}{q(\mathbf{x}_t^{i,n} | \mathbf{x}_{t-1}^{i,n}, z_t^i)}$, prior to a normalization step so that $\sum_n w_t^{i,n} = 1$

Compute the conditional mean of any function of $\hat{x}_t^i$, *e.g.* the MMSE estimate $\mathrm{E}_{p(\mathbf{x}_t^i | z_{1:t}^i)}[\mathbf{x}_t^i]$, from the approximation $\sum_{n=1}^{N} w_t^{i,n} \delta(\mathbf{x}_t^i - \mathbf{x}_t^{i,n})$ of the posterior $p(\mathbf{x}_t^i | z_{1:t}^i)$

FOR $j = 1 : i$, **DO**

  IF $d_{ij}(\hat{\mathbf{x}}_{t,k}^i, \hat{\mathbf{x}}_{t,k}^j) < d_{TH}$ **THEN**

    Save link(i,j)

    FOR k=1:K iterations, **DO**

      Compute φ_1, φ_2

      Reweight $w_t^{i,n} = w_t^{i,n}.\varphi_1.\varphi_2$

      Normalization step for $\{w_t^{i,n}\}_{n=1}^{N}$

      Compute the MMSE estimate $\hat{\mathbf{x}}_t^i$

      Compute φ_1, φ_2

      Reweight $w_t^{j,n} = w_t^{j,n}.\varphi_1.\varphi_2$

      Normalization step for $\{w_t^{j,n}\}_{n=1}^{N}$

      Compute the MMSE estimate $\hat{\mathbf{x}}_t^j$

    END FOR

  END IF

END FOR

At any time or depending on an "efficiency" criterion, resample the description $[\{\mathbf{x}_t^{i,n}, w_t^{i,n}\}]_{n=1}^{N}$ of $p(\mathbf{x}_t^i | z_{1:t}^i)$ into the equivalent evenly weighted particles set $[\{\mathbf{x}_t^{(s^{i,n})}, \frac{1}{N}\}]_{n=1}^{N}$, by sampling in $\{1, \ldots, N\}$ the indexes $s^{i,1}, \ldots, s^{i,N}$ according to $P(s^{i,n} = j) = w_t^{i,j}$; set $\mathbf{x}_t^{i,n}$ and $w_t^{i,n}$ with $\mathbf{x}_t^{(s^{i,n})}$ and $\frac{1}{N}$

END IF

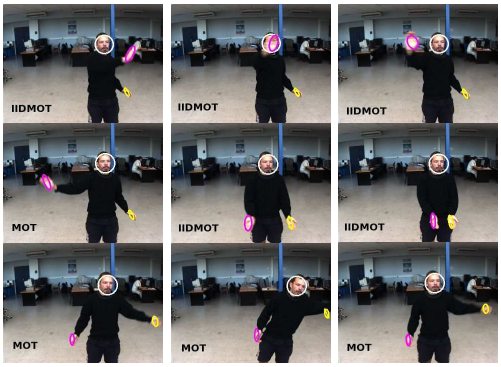


Fig. 2. Tracking scenario involving occlusion and out-field of sight with our IIDMOT filter.

The centroids and associated covariances of the matched regions are finally triangulated using the parameters of the calibrated stereo setup. For the weight updating step, each ellipsoid defined by its configuration $\mathbf{x}_t^i$ is then projected in one of the two image planes. The measurement function fuses skin color information but also motion and shape cues (see [3] for more details).

Prior to their integration on Jido, experiments on a database of 10 sequences (1214 stereo-images) acquired from the robot are performed off-line in order to: (i) determine the optimal parameter values of our strategy, (ii) characterize its performances. This sequence set involves variable viewing conditions, namely illumination changes, clutter, occlusions or out-field of sight. Figure 2 shows snapshots of a typical run involving sporadic disappearances of the hands. For each frame, the template depicts the projection of the MMSE estimate for each ellipsoid. Our strategy, by drawing some particles according to the detector output, permits automatic re-initialization and aids recovery after transient loss.

TABLE II

QUANTITATIVE PERFORMANCE AND SPEED COMPARISONS.

Method	MIPF	IDMOT	IIDMOT
FR_p	29%	18%	4%
FR_l	9%	1%	1%
Speed (fps)	15	12	10

Quantitative performance evaluation have been carried out on the sequence set. Since the main concern of tracking is the correctness of the tracker results, location as well as label, we compare the tracking performance quantitatively by defining the false position rate (FR_p) and the false label rate (FR_l). As we have no ground truth, failure situations must be defined. No tracker associated with one of the target in (at least) one image plane will correspond to a position failure while a tracker associated with the wrong target will correspond to a label failure. Table II presents the performance using multiple independent particle filters (MIPF) [5], conventional IDMOT [12] strategy, and our IIDMOT strategy with data fusion. The experiments, performed on an on-board 3 GHz Pentium PC, consider 100 particles to track each body part. Our IIDMOT strategy is shown to outperform the conventional approaches for a slight additional time consumption. The MIPF strategy suffers especially from "labelling" problem due to lacking modeling of interaction between trackers while the IDMOT strategy doesn't recover the target after transient loss.

B. Gesture interpretation

Gesture interpretation is reported briefly as this is not our key research goal while associated evaluations are currently performed. The typical motions pattern of eight reference gestures are classically modeled by dedicated HMMs. Five gestures serve for deictic references depending on the hold hand and the coarse pointed direction. The three last ones

TABLE III

EXAMPLES OF REQUESTS INTERPRETED BY THE ROBOT.

User starting interaction by introducing himself to the robot	"Hi robot X I'm Paul"
Basic movement orders / more advanced movement including deictic	"Turn left" / "Come here"
Guidance request in the human environment	"take me to the reception"
Interaction for object exchange including anaphora	"Give me this bottle"
Agreement / disagreement / thanks	"yes" / "no" / "thank you"

correspond to symbolic gestures, namely *"stop"* and *"come on"*[2]. The two-hands features used as the observations are derived from the tracked head position while each HMM has been found to perform best with 3-state model each. Preliminary evaluations on sequence dataset issued from a commercial human motion capture highlights that our gesture recognizer scored at about 91% sensitivity and 92% selectivity. Evaluations dedicated to our IIDMOT multi-tracker output, and so on noisy observations, will follow.

IV. SPEECH INTERPRETATION

The speech understanding system must recognize continuous speech, or even spontaneous, and must handle some linguistic phenomena ordinarily used in conversational speech and multimodal communication. We present hereafter how our robot can perceive and understand the information conveyed by the spoken message, how it can infer that a gesture event is necessary to complement speech or how gesture can strengthen speech in case of recognition errors.

A. Integration of a speech recognition module on the robot platform

To fulfil the platform architecture and software requirements, we have chosen to use a grammar-based recognizer. Julian, is a version of Julius developed by the Continuous Speech Recognition Consortium [1] which is itself an open source speech recognition engine. To process French utterances, a set of acoustic models (for phonetic units), a phonetic lexicon of words and a set of language models must be provided.

B. Linguistic resources for speech recognition

The acoustic models stem from previous work on large vocabulary speech transcription. They are HMM-based (37 phoneme and one short and one long pause, each one is a 3-state model with 32 Gaussians per state) and have been trained using the HTK toolkit on about 31 hours of Broadcast News recorded on French radios. Though speech recognition in a human-robot interaction context is a different task from the initial one, these acoustic models have not been adapted yet to this new applicative context, while the lexicon and the grammars have been specifically designed for it. The lexicon with 246 words and their different pronunciations (corresponding to 428 phoneme sequences) have been drawn up from the French lexical database BDLEX [8]. This vocabulary has been selected according to different subtasks as shown in the table III[3]. In order to focus on the multimodal aspect of human-robot communication, we will take a particular interest in recognizing and interpreting deictic and anaphora.

The language models, which are implemented through different context free grammars related to the above subtasks, describe an overall set of 2334 well-formed sentences.

C. Speech interpretation

This part of the RECO module processes speech recognition outputs in order first to extract the semantic units that are relevant in the user utterance and then to build the appropriate interpretation. It is based on a semantic lexicon specifically designed which associates relevant words with their interpretation in the context of the aforementioned subtasks. Some words are related to actions while others are related to objects, object attributes like color or size as well as location and robot configuration parameters (speed, rotation, distance). A semantic analysis step combines word semantic information and builds a global interpretation which is compared with available interpretation models. If one of them is compatible with the utterance interpretation, we consider that a valid command can be generated and sent to the robot supervisor in order to be executed.

D. Interpreting deictic and anaphora

Deictic words (here, there, ..) are defined in our semantic lexicon as related to a location which will be given by means of a gesture. This is specified by a semantic feature (location = Gesture_location?) . For example, if the verbal designation of an object or a location is precise enough (*"Put the bottle on the table"*) the parameters are directly extracted from the sentence according to the relevant words and their semantic information. In our semantic lexicon, the word "put" is associated with the meaning "put something somewhere" which is represented by the set of semantic features (action = put ; object = What? ; location = Where?). The sentence analysis instantiate the missing parameters (What? and Where?) and the underlying command can be generated (put(object=bottle, location = on_table)). But in deictic case (*"Put the bottle there"*), the semantic analysis will mark the interpretation as "must be completed by the gesture result" and a late and hierarchical fusion strategy will be applied to complete the command that has been generated (put(object=bottle, location=Gesture_location?)) (see section V). In the case of an anaphoric sentence (*"Take this glass"* (action = take; ref_object = (object = glass ; ref_location = Gesture_location?))) and other human-dependent commands such as (*"Come on my left-side"* (action = go ; relative_location = (ref_location = User_position? ; side = left))) the same kind of strategy will be applied. For the moment, only location reference are taken into account. In a human-robot dialog prospect anaphora could also be

[2]From single or two-handed gestures.

[3]Examples are given in English for an illustration purpose.

TABLE IV

EXPERIMENTAL RESULTS OF THE SPEECH RECOGNITION COMPONENT ALONE AND OF THE GLOBAL SPEECH INTERPRETATION SYSTEM.

subtask	COR_W	ACC	COR_S	COR_COM
starting/closing interaction	88.34%	81.97%	67.19%	71.88%
basic movement orders	89.63%	81.72%	65.10%	70.05%
basic object manipulation orders	86.41%	80.62%	62.50%	66.25%
deictic	94.79%	90.77%	82.81%	83.33%
guidance request	83.30%	78.66%	48.75%	71.25%
complete order for object exchange	86.41%	78.80%	61.25%	66.88%
anaphoric order for object exchange	85.62%	69.38%	47.92%	48.96%
agreement/disagreement	94.12%	89.34%	79.38%	83.75%
robot status	81.44%	77.34%	75.00%	75.00%
overall results	84.15%	75.84%	66.19%	71.69%

solved by means of an history, which is not taken into account yet.

E. First evaluation of speech recognition and interpretation

In order to evaluate the RECO module, a list of 50 well-formed sentences related to the different tasks described above has been drawn up. Each one has been uttered 32 times so our first evaluation corpus counts 1600 utterances. Fourteen different speakers were involved in these experiments. These first results are given in the table IV : percentages of correct words (COR_W), accuracy at word level (ACC), correct sentences (COR_S) and correct commands (COR_COM). A command has been generated from each valid interpretation of a speech recognition result and then compared with the corresponding reference command.

General comments can be made about these results. For each subtask, COR_COM is greater than COR_S (or equal in the last case). The speech recognition errors, at the word level, have less impact on the command than on the sentence. If a word omitted, inserted or substituted by another one, is not semantically relevant, this will not have a real impact on the command generation, but the sentence will be considered as completely wrong. This explains the COR_COM higher rates. The results for deictic orders are correct unlike the anaphoric ones, especially for the sentence (*"Take this"*). Only the best recognition output is taken into account at the moment. At mid-term, the N-best results will be considered at the fusion level. If a gesture has been interpreted, and if the recognized sentence does not need a complementary gesture, we can detect an incoherence and we could propose another interpretation. Further developments will consider such a multiple hypothesis strategy while the acoustic models will be adapted to the robotic context.

V. THE MULTIMODAL INTERFACE AND LIVE EXPERIMENTS

A. Vision and audio fusion

Vision and audio data are merged using a rule based approach. The speech is used as the main channel : the RECO module, thanks to its semantic knowledge, identifies actions needing a gesture disambiguation. Vision is used in a late and hierarchical fusion strategy to complete this input information.

For deictic commands, like *"put the bottle there"*, and its interpretation (put(object=bottle, location=Gesture_location?) the non instantiated parameters (here, the bottle position) are specified by the FUSION module via the line of sight between head and the hold hand extracted by the GEST module. In these cases, we assume that we can use head and hands 3D positions at the end of the speech utterance to extract the pointed direction, knowing that speech and gestures are strongly correlated in time. For human-dependent commands such as *"come on my left-side"* and its interpretation (action = go ; relative_location = (ref_location = User_position? ; side = left))) , the same kind of strategy is applied, extracting the human position from the head location.

B. Live experiments

The integration of the multimodal interface on Jido enables us to perform online experiments in our lab. Figure 3 illustrates a typical run of the scenario. For each step, the main picture depicts the current H/R situation, while the subfigure shows the tracking results of the GEST module.

The robot succeeds to interpret a sequence of commands by melting multimodal features in the FUSION module. The entire video and more illustrations are available at the URL `www.laas.fr/~bburger`.

More globally, the robot succeeds to execute the scenario in the majority of runs with Jido successfully bringing the bottle to its human user. The principal failures are attributable to the precision of pointing gesture which decreases with the angle between the head-hand line and the table. The multimodal interface is shown to be robust enough to allow continuous operation for the long-term experimentations that are intented to be performed.

VI. CONCLUSION

In this paper, we propose a scenario for Human-Robot interaction based on mutual assistance between speech and vision which rely on three modules integrated on a robotic platform. Before integration on the platform, each module and the underlying methods implemented are described, followed by some results provided by a step of quantitative evaluation of the module performances. The first contribution describes a fully automatic distributed approach for tracking two-handed gestures and head tracking in 3D. The amended particle filtering strategy allows to recover automatically from transient target loss while data fusion principle is shown to improve the tracker versatility and robustness to clutter. Speech recognition and interpretation constitutes the

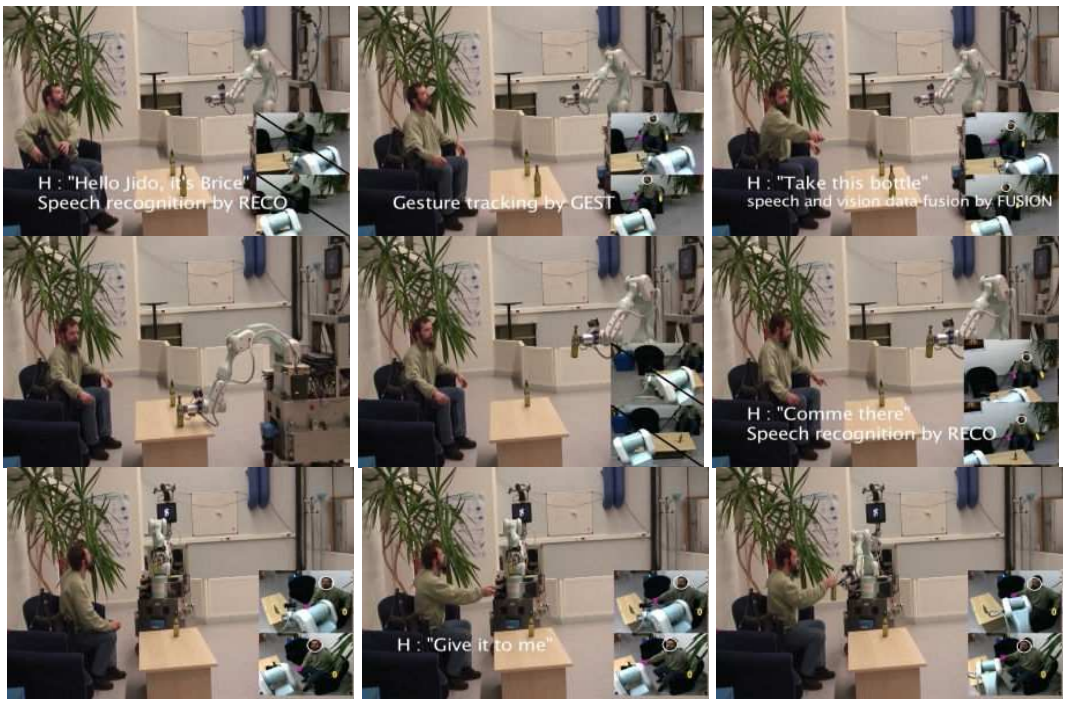


Fig. 3. From top-left to bottom-right : GEST module -left-, virtual 3D scene (yellow cubes represent hands) -middle-, current H/R situation -right-.

second contribution, focusing on the interpretation of utterances related to predefined subtasks and more particularly on deictic and anaphoric commands requiring fusion with gesture events. Then, in order to specify parameters for location references and object/person IDs and complement verbal statements, we present the outlines of the late fusion performed from both speech and gesture analysis. As shown by the scenario execution, these preliminary robotic experiments are promising even if speech recognition performances must be improved and quantitative performance evaluations still need to be carried out. These evaluations are expected to highlight the robot capacity to succeed in performing multimodal interaction. Further investigations will also be to : (i) process more natural and flexible utterances about object manipulation tasks, (ii) estimate the head orientation as additional features in the gesture characterization. Our robotic experiments report strongly evidence that person tend to look at pointing targets when performing such gestures. Dedicated HMM-based classifiers will be developed to filter more efficiently pointing gestures. Another investigation line will be to study other fusion methods based on the conjoint modelling of speech and gesture.

Acknowledgements: The work described in this paper was partially conducted within the EU Projects COGNIRON ("The Cognitive Robot Companion" - www.cogniron.org) and CommRob ("Advanced Robot behaviour and high-level multimodal communication" - www.commrob.eu) under contracts FP6-IST-002020 Future and FP6-IST-045441.

REFERENCES

[1] T. Kawahara A. Lee and K. Shikano. Julius — an open source real-time large vocabulary recognition engine. In *European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1691–1694, 2001.

[2] R. Alami, R. Chatila, S. Fleury, and F. Ingrand. An architecture for autonomy. *International Journal of Robotic Research (IJRR'98)*, 17(4):315–337, 1998.

[3] Brice Burger, Isabelle Ferrané, and Frédéric Lerasle. Multimodal interaction abilities for a robot companion. In *Int. Conf. on Computer Vision Systems (ICVS'08)*, pages 549–558, Santorini, Greece, 2008.

[4] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42:143–166, 2003.

[5] M. Isard and A. Blake. CONDENSATION – conditional density propagation for visual tracking. *Int. Journal on Computer Vision (IJCV'98)*, 29(1):5–28, 1998.

[6] J. Maas, T. Spexard, J. Fritsch, B. Wrede, and G. Sagerer. A multi-modal topic tracker for improved human-robot interaction. In *Int. Symp. on Robot and Human Interactive Communication*, Hatfield, September 2006.

[7] K. Nickel and R. Stiefehagen. Visual recognition of pointing gestures for human-robot interaction. *Image and Vision Computing (IVC'06)*, 3(12):1875–1884, 2006.

[8] G. Pérennou and M. de Calmès. MHATLex: Lexical resources for modelling the french pronunciation. In *Int. Conf. on Language Resources and Evaluations*, pages 257–264, Athens, June 2000.

[9] O. Rogalla, M. Ehrenmann, R. Zollner, R. Becher, and R. Dillman. *Advanced in human-robot interaction*, volume 14, chapter Using gesture and speech control for commanding a robot. Springer-Verlag, 2004.

[10] M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, and W. Adams. Spatial language for human-robot dialogs. *Journal of Systems, Man, and Cybernetics*, 2(34):154–167, 2004.

[11] R. Stiefelhagen, C. Fügen, P. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel. Natural human-robot interaction using speech head pose and gestures. In *Int. Conf. on Intelligent Robots and Systems (IROS'04)*, Sendal, October 2004.

[12] Q. Wei, D. Schonfeld, and M. Mohamed. Real-time interactively distributed multi-object tracking using a magnetic-inertia potential model. In *Int. Conf. on Computer Vision (ICCV'05)*, pages 535–540, Beijing, October 2005.

[13] M. Yoshizaki, Y. Kuno, and A. Nakamura. Mutual assistance between speech and vision for human-robot interface. In *Int. Conf. on Intelligent Robots and Systems (IROS'02)*, pages 1308–1313, 2002.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
PERGAMON
Control Engineering Practice 11 (2003) 1413–1421
CONTROL ENGINEERING PRACTICE
www.elsevier.com/locate/conengprac

Comparison of structured light and stereovision sensors for new airbag generations ☆

S. Boverie[a], M. Devy[b,*], F. Lerasle[b]

[a] Siemens VDO Automotive SAS, BP 1149, avenue Paul Ourliac, 31036 Toulouse, Cedex1, France
[b] LAAS-CNRS, Robotics and Artificial Intelligence Group, 7 avenue Colonel Roche, 31077 Toulouse, Cedex, France

Received 30 August 2002; received in revised form 7 April 2003; accepted 14 April 2003

Abstract

Providing new generations of airbags with reliable information about the vehicle inner space occupancy in order to minimize inappropriate inflation is a real challenge. Within this paper two techniques to rebuilt the 3D cockpit scene, are presented; beside the well-known stereoscopic vision principles, 3D reconstruction based on matricial sensor combined with an infrared structured light emitter is described. From the 3D points acquired by these sensors a 3D description of the seat area is built, pertinent attributes are extracted, and using a previously learnt data base, the current seat situation is identified amongst the learnt situations (empty seat, baby seat in normal or rear position, occupant, ...). First promising results are depicted.

Keywords: Video sensing; Airbag; Classification; 3D reconstruction

1. Introduction

In the last few years, a special effort has been focused on the improvement of passive safety both by car manufacturers and suppliers all over the world. Frontal airbags for the driver and the passenger are now mounted in almost every new car. In brief lateral airbags will be massively installed. The most part of these airbag systems are operating in open-loop conditions. That means that whenever an impact is detected, the airbag is automatically inflated without any feedback from the seat occupant nature. This operating mode has caused some dramatic situations, for example: babies installed on a baby seat in rear position thrown to the back of the vehicle when airbag inflates, passengers in "out of position" situation who are injured by the airbag, etc.

In US, 97th statistics show that airbags saved about 1600 lives. It is also acknowledged that they killed 32 children and 20 small adults. New generations of airbags need to be more "intelligent" to provide appropriate inflations with regard to the vehicle inner space situation and thus to minimize the injury risk. The improvement of this function requires the introduction of new sensors to provide reliable information about the cockpit occupancy such as passenger nature, occupant morphology, occupant volumic distribution to detect "out of position" situations of the passenger (Fig. 1). This information will then be used by airbag electronic control units (ECUs) in order to take the appropriate decisions (Fig. 2).

Classical systems fuse heterogeneous physical measures provided by optical, pressure, capacity, thermical or weight sensors placed between the seat and the dashboard, for example the PDS from BMW, the IROPS from Siemens VDO or the Delphi's systems. Data fusion of such sensors increases of course the system reliability or performance but these devices remain inadequate to give pertinent informations such as the passenger posture for example.

More sophisticated techniques have been explored: time of flight sensor and well-known stereoscopic vision. A time of flight sensor (Boverie, Devy, Lequellec, Mengel, & Zittlau, 2000) is based on a NIR laser diode and a CMOS camera with an ultra short integration time; the intensity measured at the CMOS sensor

☆This project was supported and funded by the PREDIT II program of the French Ministry of National Education and Research.
*Corresponding author. Tel.: +33-5-61336331; fax: +33-5-61336455.
E-mail address: michel@laas.fr (M. Devy).

0967-0661/$ - see front matter © 2003 Published by Elsevier Ltd.
doi:10.1016/S0967-0661(03)00137-0

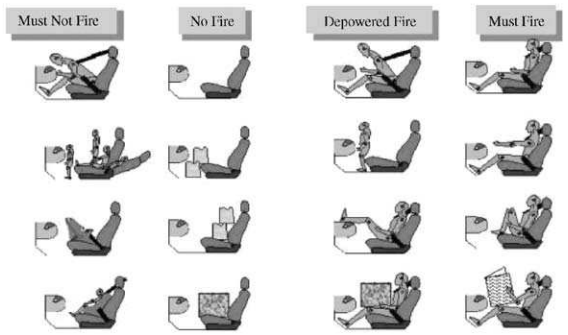


Fig. 1. Scenarii for occupant detection.

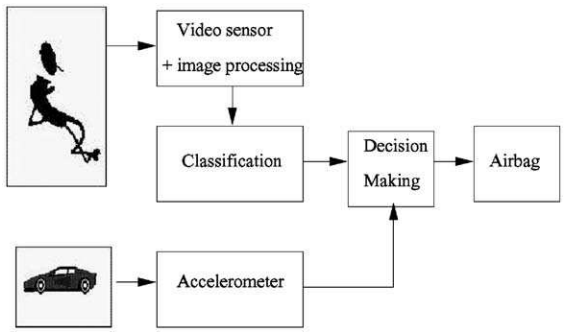


Fig. 2. System architecture.

depends on the distance and the surface reflexion. A recent improvement of such systems consists in determining the light propagation time by multiple double short time integration (Mengel, Doemens, & Listl, 2001). Advantages of this system are its cost, its insentivity to backlight and its high-integrated possibilities while a major drawback concerns the need of the short illumination time and the good synchronization between the camera shutter and the light source.

Stereoscopic vision is based on the principle that depth information can be computed by triangulation from two images with a common area in their field of view. For many years, many developments have led to mature 3D perception methods (Matthies, 1992; Granjean & Lasserre, 1995). Faber and Forstner (2000) use a dense stereoscopic method to detect the passenger presence. The basic principle consists in looking for the passenger head. Krumm and Kirk (1998) are looking for three classes, rear facing baby seat, empty seat and other situations. They show: (1) a first approach using monocular vision and ACP on an image base in order to find a minimal dimension of the space in which it is possible to separate the three classes, (2) a second approach takes advantage of the same classification methods but with data provided by a stereoscopic sensor. Results look less good but more stable in relation with light disturbances.

In our project, two different solutions have been developed. Beside the passive stereoscopic sensor, an active 3D vision system has been developed, based on a CCD camera combined with an IR structured light emitter. From one or another system, 3D reconstruction is reached and attribute vectors are then extracted and allow to identify the current seat situation among the learnt ones illustrated in Fig. 1.

These two vision-based methods will be presented and discussed in this paper, with respect to several criteria: density of the acquired depth images, acquisition speed, mechanical and safety constraints, etc. The first promising results with these different technologies will be presented.

2. Structured light-based approach

2.1. Generalities

This approach developed by Siemens VDO Automotive in collaboration with LAAS-CNRS and ONERA-DOTA is concerned with 3D active vision system based on a matricial sensor combined with an infrared structured light emitter. This principle is currently used in other application fields as robotics, architecture, etc. (Boyer & Kak, 1987; Rosenfeld & Tsikos, 1986; Stockman & Hu, 1989). A light pattern is projected onto the scene, the 2D deformation of this pattern in the image plane due to the objects contained in the scene is analyzed. Then once the sensor is calibrated, triangulation techniques allow to give 3D data on the observed scene. The latest step of this process is the extraction of specific characteristics from the 3D reconstruction and then the classification. This technique is usually applied with supervised environmental conditions (light control, etc.), low real-time constraints and no light power restrictions.

The approach is original in the sense that the system has to cope with specific automotive constraints and to present a good robustness with respect to an uncontrolled light disturbances and back-light. These constraints have required the development of specific measurement devices and algorithms in order to classify all the most current situations regarding occupant safety.

2.2. Sensor presentation

The efficiency of such a system depends on some characteristics like the resolution of the sensor and light emitter, the gap between the light emitter and the sensor, the calibration accuracy, the radiance of the beams, its location in the cockpit, etc. The gap between the light emitter and the sensor must be large enough to allow 2 cm accuracy on 3D points located at 1 m from the

camera. A too large gap will lead to mounting problems and to occluded beams. A 6 cm gap has issued good results. The light emitter is composed of a laser diode (830 nm wavelength), a Damman diffraction grid that split the original beam into several beams and a concave planar lens to spread the illumination on a conic field of $90^\circ \times 70^\circ$.

The light emitter resolution is quite critical in order to determine if there is a person or an object on the seat and to distinguish head, arms, etc. from a passenger. Current developments have been performed with a 11×11 array of beams that demonstrate a good compromise between 3D reconstruction accuracy and calculation time. Another prototype has been designed with 19×16 beams to have a meaningful reconstruction for demo-car (Fig. 3(a)). Each laser beam D_f is labelled by its row and column position in the array (Fig. 3(b)).

In order to brighten the dots on the image, the radiance of each beam should be maximized by optimizing the output power (limited by the eye safety requirements) and narrowing the beam diameter. The maximal output power is function of the wavelength of the emitted light and of the operating modes of the illuminator. As an example for an illuminator wavelength of 850 nm, operating in a pulse mode of 1 ms duration every 10 ms, the maximum permissible power considering, non-intentional vision is 0.78 mW (Standard IEC 825-1). Pulse duration of 100 μs would increase the permissible power by a factor 2.

Image acquisition is carried out using a single short focal length CCD camera with a 2.6 mm lens providing a field of view of $130^\circ \times 100^\circ$. The off-line calibration step determines the parameters corresponding to the well-known perspective projection camera model, taking into account intra and intersensor/illuminator and lens characteristics (Devy, Garric, & Orteu, 1997). At first, the camera calibration process consists in estimating in one global step both the intrinsic and distortion parameters from matchings between a set of points defined on a planar calibration target and their projections on the image plane. To reduce the measurement errors, the computation of the parameters is done with multiple pose of this target.

Then another calibration process is required to identify the laser beams equation parameters in the camera coordinate system. It uses multiple views of the same calibration target on which the lasers beams are projected (Fig. 3(c)).

The camera/illuminator device is located in the overhead console position which appears to be the most efficient position since it provides the best overview of the passenger seat even if it has two drawbacks. The first one is the need of a very wide angle of sight of the illuminator/camera ($> 90^\circ$). The second one is that the compactness requirement gets critical which could lead to mount the ECU in a distant location.

2.3. 3D reconstruction and classification

The occupant detection methodology can be decomposed in a 3D reconstruction process and then a classification one, both detailed hereafter. The 3D reconstruction process itself requires two steps. The *first step* concerns the extraction of light dots, intersections of laser beams with the image plane, by the way of conventional image processing techniques that have to deal with problems like light disturbances (shadows, direct sunlight, etc.), passenger movement, surfaces with low reflexivity, specular reflections, etc.

The *second step* is related to the dot labeling, e.g. the search of matchings between the light dots and the laser beams. The labeling process is based on the sequential application of the following constraints:

Epipolar constraint: For each dot i, after distortion correction, and each beam D_f, the distance in the image between the dot and the D_f projection (noted $D_{f_{proj}}$) is calculated. A matching (i,f) is discarded if the distance $d(i,f)$ exceeds a certain tolerance, generally three or four pixels in experiments. Fig. 4 shows light beams projections in the image plane. The different crosses represent the extracted light dots over time. Although the beams projections are very close to each other, thanks to this first constraints, the number of candidate beams D_f to correspond to a light dot i is drastically reduced (about 95%).

Depth constraints: For cockpit occupancy reconstruction, the triangulation must yield to 3D points whose

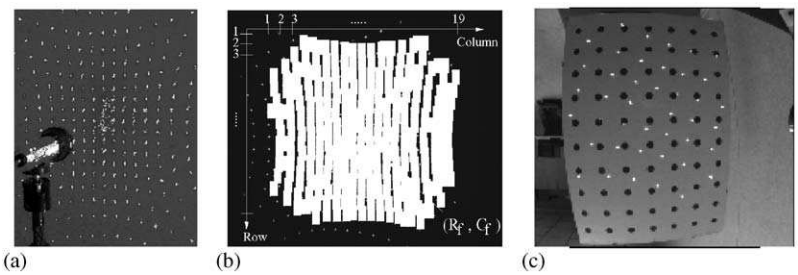


Fig. 3. (a) Laser beams; (b) associated labels; (c) calibration target.

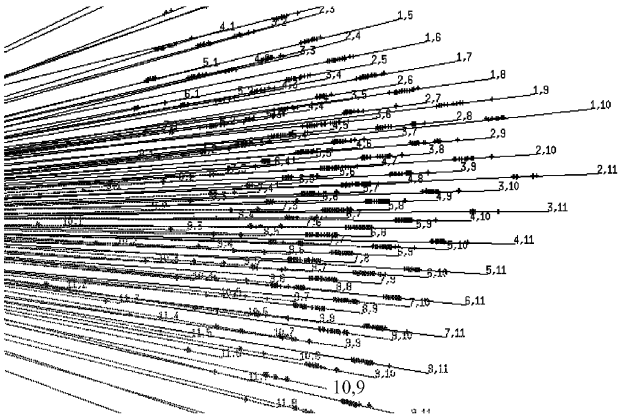


Fig. 4. Image dots positions and associated projected beams.

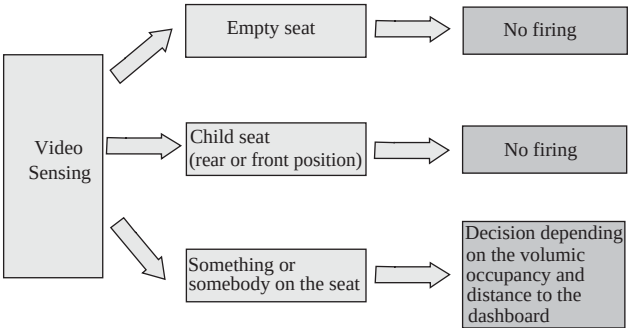


Fig. 5. The different scenarii and associated airbags operating conditions.

depth lies within the range [0, 1.5 m]. A matching (i,f) is discarded, if the depth of the resulting reconstructed 3D point, exceeds this threshold.

Topological constraints: The two following constraints allow to evaluate the matching confidence:

(a) *Uniqueness*: Each laser beam must be imaged with, at most, one light dot. So, a beam label can be associated to at most one dot and vice versa.

(b) *Order*: This constraint implying pairs of light dots, is applied only if their two labels belong to the same column or row in the array of beams (Fig. 3(b)). Given two light dots $p_i = (u_i, v_i)^{\mathrm{T}}$ and $p_j = (u_j, v_j)^{\mathrm{T}}$, two labels f (column C_f, row R_f) and g (column C_g, row R_g), the order constraint is:

if $R_f = R_g$ (resp. $C_f = C_g$) then (i,f) and (j,g) are consistent $\Leftrightarrow (C_f, C_g)$ (resp. (R_f, R_g)) and (v_i, v_j) (resp. (u_i, u_j)) are in the same order.

To apply these constraints, three different optimization techniques have been evaluated maximal cliques, continuous relaxation and discrete relaxation (Lerasle, Lequellec, & Devy, 2000). This last one has shown the better compromise between speed and matching performances. Its principle can be summarized as follows: a first pass, defined by epipolar and depth constraints, exhibits trustworthy and ambiguous labellings, depending on whether a dot i matches a unique beam f_0 or may be associated with several ones. Topological constraints are next checked for the remaining ambiguous matchings (Lequellec & Lerasle, 2000). The discrete iterative relaxation process is set up so as to eliminate the ambiguous matchings that are incompatible with confident ones regarding some uniqueness and order relationships. Consequently, the total number of confident matchings increases iteratively.

From the labelling process results and the off-line calibration phase, 3D coordinates can be computed. From the 3D points, the 3D shape is then represented by a triangular mesh (Fig. 7). Referring to the final decision related with airbag inflation, the different scenarii described in Fig. 1 can be grouped and reduced to a simpler set represented in Fig. 5.

The classification method is also based on a hierarchical process. The *first level* consists in detecting one of the three occupant categories: empty seat, baby seat, "something" on the seat. It is based, on the one hand, on the analysis of the motion of the objects located within the observed scene and, on the other hand, on the extraction of specific attributes related to each occupancy class and computed in some analysis areas defined in the vehicle longitudinal projection plane. The 3D reconstructed points are simply projected onto this particular plane to have a kind of scene profile characteristic of the occupancy class (Fig. 8). The analysis areas correspond to three horizontally oriented stripes in the longitudinal plane and are located as follows: one which envelops the seat sitting and the two other located just above the seat sitting. The classification is made by the method of the K nearest neighbors, from the results of the attributes learnt in a database built from images acquired on the different seat occupancy classes.

In case of "something" is detected on the seat, a *second processing level* estimates the position of the occupant with respect to two operating zones of the airbag. This very simple technique is based on the analysis of the number of light dots in each zone. These zones are defined according to an emerging normalization about the critical volume close to the dashboard (Fig. 6): the critical out of position (COOP) and out of position (OOP) volumes. The only criterion for the definition of these zones, consists in the distance to the dashboard (invariant w.r.t. the seat position). Finally, this information is then fused in order to provide a final decision to the airbag ECU.

2.4. Results

The different algorithms have been implemented in C code on a 400 MHz PC. 3D reconstruction is performed

each 50 ms for a 11×11 array of beams. Fig. 7 shows two examples of reconstruction provided with the 19×16 array. The right sub-figures represent the triangular meshes generated from the 3D points. The results show that the system is able to give a very good approximation of the volumetric distribution of the occupancy within the observed scene. In addition a very good estimation of the distances between the dashboard and the occupant is achieved (± 2 cm).

The classification even if it is based on very simple principles gives promising results. The tests that have been performed in real situations have proved that it is able to make the distinction between the most important occupancy classes listed before. In addition this technique allows to make the distinction between the different dashboard proximity situations. In a further step, the analysis of the dynamic evolution of the situation along an image sequence, will lead to an improvement of the classification robustness.

Fig. 8 shows two scenarii examples and their projections in the longitudinal plane. In the first case, some reconstructed points are close to the airbag, so the airbag must not fire. In the second case, the airbag firing is compatible with the analyzed situation.

3. Stereovision-based approach

Concerning the stereoscopic vision, developments take profit of well-known principles and aim to adapt them to the constrained automotive context: real time performances, dense and accurate reconstruction, low cost technology, etc.

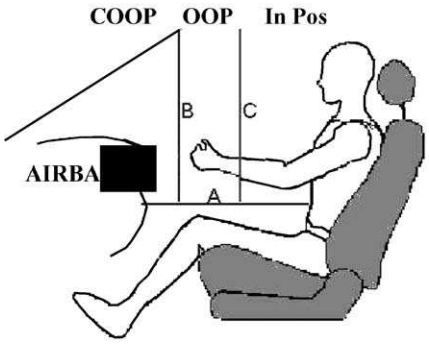


Fig. 6. OOP and COOP zones.

3.1. 3D acquisition from stereovision

The pixel-based stereo algorithm aims to match pixels between left and right images (Gautama, Lacroix, & Devy, 1999) acquired by the stereo cameras. The stereovision process includes several steps. An off-line calibration determines the parameters of the stereo sensor: camera models, inter camera situations, lens distortions etc. it allows computing the epipolar geometry between the cameras.

From on line acquisitions, the rectification process (Granjean & Lasserre, 1995) corrects the original images, to perform a perfect virtual alignment of the two cameras and their epipolar lines; two matched pixels must be on the same line of the rectified images. Rectification and distorsion correction could require complex computations: these functions are performed in the same loop, using pre-computed tables to find the rectified (u, v) coordinates from the real ones, and using simply a bilinear interpolation.

Then, for every line of the right and left images, the correlation process (Fig. 9) must match pixels, with respect to a similarity measurement based on windows centered on the compared pixels: several similarity measurements (sum of squared differences (SSD), zero normalized cross correlation (ZNCC), census transform (CT), etc.) have been evaluated, using typically 11×11 windows. Every pixel (u, v) on the left image, is compared to all potential matched pixels of the right image, located on the same u line, from the position $v + d_{min}$ to $v + d_{max}$; a score function $score = f(d)$ is obtained, where d is the disparity between the corresponding left and right pixels. A good optimum must be found from the *score* function. Several criteria can be used to filter false matchings: strength, uniqueness and form of the correlation peak (Fig. 10), and right–left validation. This last step inverses the role of left and right images and considers as valid only those matches

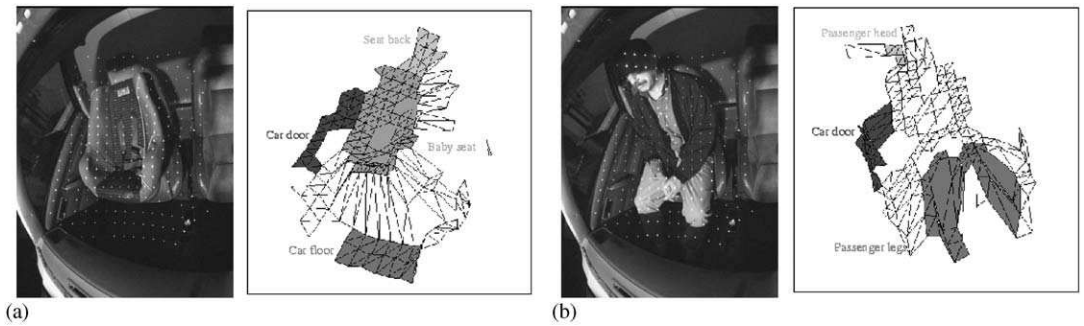


Fig. 7. 3D reconstruction of scenarii: (a) empty baby seat; (b) passenger on the seat.

for which the reverse correlation has fallen on the initial point in the left image (Granjean & Lasserre, 1995).

Sub-pixel estimates are obtained by fitting a parabole to the correlation values surrounding the optimum. The quality is accessed not only by the correctness of the estimate, but also by the ability to filter out false matchings using the validity criteria.

The better similarity measurement has been selected in order to improve the robustness of the process with respect to environmental conditions; the CT score gives more matches, less artifacts in the disparity map and a better reconstruction than the classical scores. Nevertheless, matching can be found only on textured areas of the images. In Fig. 11, the disparity map is presented for an image of a passenger in the cockpit: white points correspond to unmatched pixels, in homogeneous areas of the original image; generally, the correlation is very good on the passenger head or hands (skin, hair); CT score is better than ZNCC score.

At last the 3D reconstruction process based on triangulation techniques and on the calibration results, computes a depth map from the disparity map.

The performances of the stereovision method are good enough to fulfill real-time and accuracy requirements. The complete algorithm is executed in 250 ms on 128×128 images; it provides a 3D dense reconstruction on the passenger seat (between 3000 and 5000 3D points) so that a large variety of situations (e.g. passenger in advanced or extended position, different objects, etc.) could be sufficiently characterized to provide good inputs for a classifier. With such a frequency, the airbag requirements are not satisfied, especially for fast passenger motions (fast transition between safe and unsafe situations); the stereo frequency could be increased easily to 20 Hz using a multi-resolution approach.

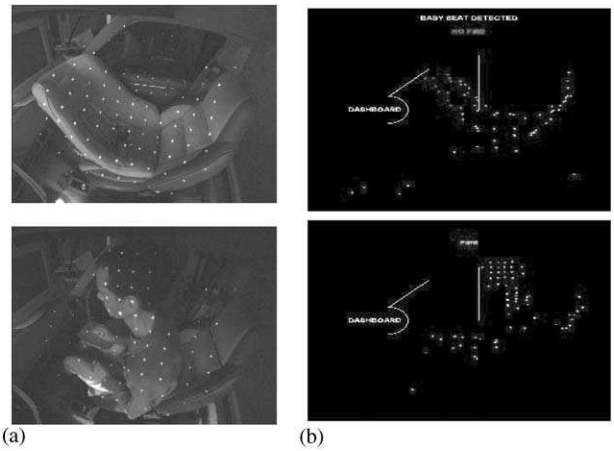


Fig. 8. (a) Original images; (b) reconstruction projection in the vehicle longitudinal plane and associated airbags operating conditions.

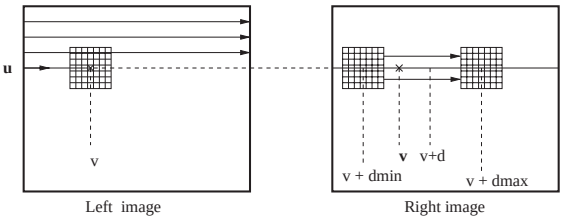


Fig. 9. Left to right correlation.

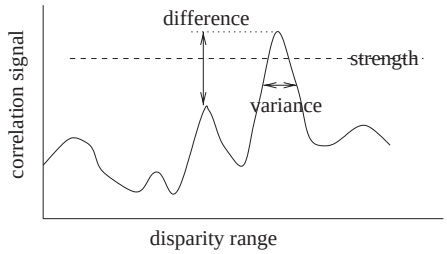


Fig. 10. Detection of the correlation peak.

3.2. Classification from stereovision

By now, the passenger seat classification is performed using a classical case-based approach. During a supervised learning step, a lot of prototypes are recorded for each class to be identified. Fig. 12 presents some images of the large data base built in order to learn some characteristic configurations of the passenger seat. With respect to the NHTSA requirements concerning the firing conditions of the airbag, the work has been focused on identifying the following classes: (1) empty

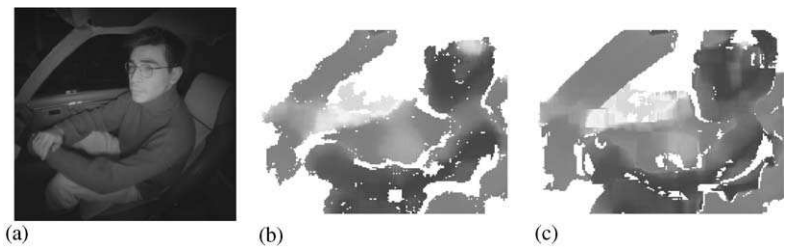


Fig. 11. Stereo matching examples with a passenger on the seat: (a) original image; (b) disparity map with CT score; (c) disparity map with ZNCC score.

Fig. 12. Some images: (a) empty seat; (b) child in a booster; (c) standing child.

seat, (2) passenger in normal position, (3) passenger out of position, (4) empty booster, (5) child in a booster, (6) front facing baby seat, (7) rear facing baby seat, (8) object(s) on the seat.

The out-of-position configurations of a passenger (Fig. 13) are mainly defined from the head position with respect to the airbag. Three areas are defined by two vertical planes parallel to the dashboard (Fig. 6); if some significative parts of the passenger are detected in the critical out-of-position area, then, the airbag cannot be inflated, while in the intermediate area, only a depowered inflation is desirable.

One difficult issue consists in identifying these situations without any false alarm or misclassification: the real-time constraints are very severe, because the configuration changes could occur very fast (for example, a quick head motion in order to switch on the radio). A trade-off must be found between the computation time required by the data acquisition and analysis, against the classification capabilities of the system: with very few data processing, it is possible to detect the presence of something in the critical out-of-position area, but it is more than likely that mistakes or false alarms will occur.

The classification strategy uses a scene description as a set of local attributes (a specific detail located in a precise location: for example, number of points acquired in a given area of the cockpit (Fig. 14), that could allow to identify a rear facing baby seat) or global ones (for example, number of points that belongs to planar faces, that could be significative to recognize an empty seat). These attributes must be *discriminant* (they must allow to distinguish between the classes), but also, *generic*, so that a generalization can be automatically obtained from the large learning data set, and *invariant* with respect to the possible modifications of the cockpit geometry, mainly the seat translation and the orientation of the seat back. This invariance issue is important, because all global attributes could fail, if they are not adapted with respect to the seat position and orientation.

An iterative method has been designed in order to find these seat parameters: from an initial position given by the lower point located in the left side of the 3D image and the higher point located in the right side, intermediate points belonging to the sitting or to the back part of the seat are integrated iteratively to the seat boundary, using some shape constraints that must be verified by this boundary (maximum translation, maximum orientation, planarity, etc.). Fig. 17 shows the seat configuration for some images presented in Figs. 12 and 13.

Fig. 13. Out-of-position passengers.

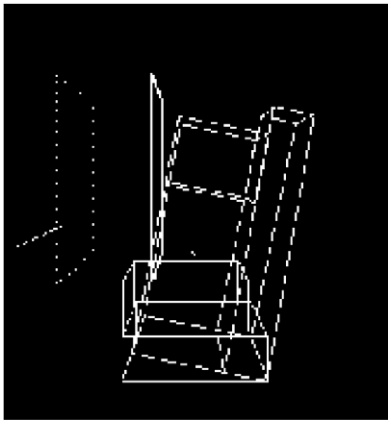


Fig. 14. Pertinent areas in the cockpit.

Once this seat configuration has been computed, some specific areas are defined in Fig. 13: the segment on the left corresponds to the dashboard. The two security areas are limited by planes and five parallelipipedic

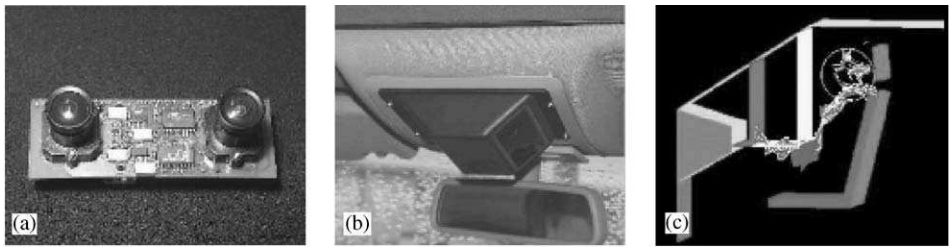


Fig. 15. (a) Camera prototype; (b) vehicle integration; (c) cockpit representation.

boxes are defined on the seat to classify the passenger seat occupancy: two boxes for the seat volume (on the seat sitting and the seat back), one box above the sitting area and two boxes with variable heights along the seat back (the lower one for the passenger body, the higher one for the head). A preliminary classification, using as attributes, the number of 3D points which belong to these seven areas, gives good results, but it is not sufficient to deal with all possible configurations.

Within the context of passenger safety, the head plays a central role. Locating the head with respect to the dashboard is an important issue and fast, robust techniques need to be developed. Some segmentation methods have been explored based either on the density or the curvature estimate on each 3D point; such an approach could be performed only on a ROI corresponding to some boxes described here before. From the depth map, using a density criterion to extract the head when a passenger is detected, an accurate enough positioning of the head within the cockpit ($\simeq 2$ cm) can be estimated.

3.3. Stereovision prototype

A low-cost and compact stereo head, developed for this application, consists in two synchronized low-cost sensor, mounted on the same board (Fig. 15(a)). It integrates a wide angle NIR illuminator and shutter/gain can be controlled by means of a RS232 interface, with respect to the intensity computed in some areas of the images. Each sensor is equipped with a low cost and compact optical lens (focal: 2.1 mm); such optic has high vignetting and distortion. In order to limit these drawbacks, only the central part of the images is processed; in this configuration, the perception field is bounded to a 110° view angle. In Fig. 16, an image of the calibration pattern (located 25 cm in front of the stereo head) shows clearly the importance of the calibration step required to correct the distortion in further processings.

This stereo prototype has been integrated in the cockpit of a demo car (Fig. 15(b)). The different algorithms have been implemented in "C" code within a Windows NT environment; a dedicated man–machine interface allows to evaluate the method results (Fig. 15(c)). First classification algorithms have been tested successfully (Fig. 17), but more intensive validations must be performed.

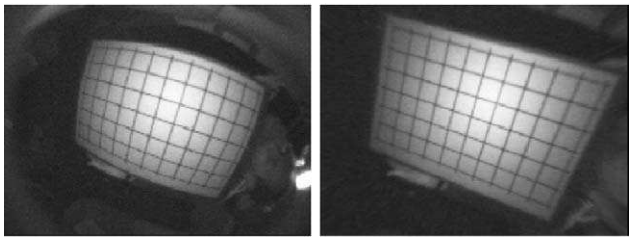


Fig. 16. Distortion correction (2.1 mm lens).

4. Comparison of the two approaches

3D vision based on structured light sensor shows some very good advantages for automotive application:

- The sensor can be realized in CMOS technology, cheap and high integrated solution can be viewed.
- First results show that 3D reconstruction using a restricted number of information allows a quite fast process with respect to the airbag requirements but some limitations exists in terms of resolution.
- Intensity images are still available and can be used for complementary processing.
- An homogeneous distribution of the light dots in the scene allows very good reconstruction capabilities.

Nevertheless, some improvements must be brought concerning the compactness of the measurement device (camera + light emitter) and with respect to uncontrolled light disturbances and backlight.

Advantages of stereoscopic vision are mainly related with the very good resolution of the reconstructed image useful for classification purposes. In addition it is weakly influenced by backlights. As it consists in a passive system, eye safety requirements are not concerned with. Some remaining drawbacks concern the computation time and the non-homogeneous distribution of the reconstructed 3D image, so that it cannot be guaranteed that the field of interest will be always described.

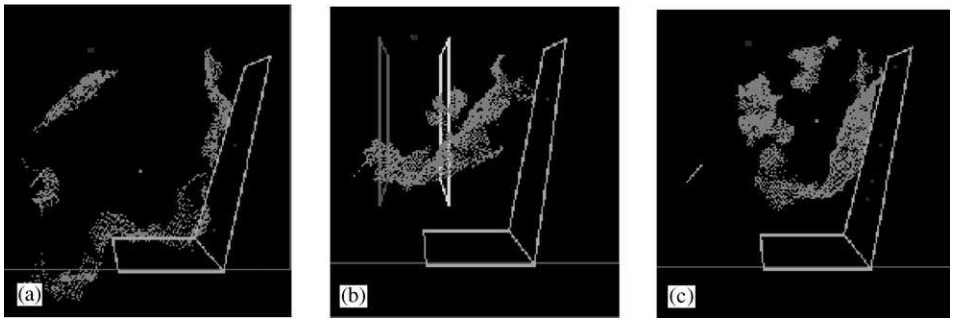


Fig. 17. 3D points and seat configuration for images presented in Fig. 12.

5. Conclusion

New airbag generations will need more and more information about the automobile cockpit occupancy: nature of the occupant, distances to the dashboard… . It is now obvious that 3D reconstruction of the cockpit inner space is necessary. Uses of video sensing for extracting this information look to be one of the real potential solutions. This technology is supported by the drastic reduction of the sensor cost but also by the exponential improvement and cost reduction of the image processing hardware. Within Siemens VDO Automotive, in collaboration with some labs and institutes, both in France and Germany, several video-based approaches have been evaluated for performing 3D reconstruction: conventional stereoscopic vision, 3D vision based on structured lightning, and time of flight techniques.

The results presented within this paper show the very high potentialities of the two first mentioned techniques. Beside the well known stereoscopic technique which gives a very high reliability of the reconstructed 3D images, but still remains slightly heavy in terms of computational time, uses of a structured light sensor looks to propose an interesting compromise between accuracy and speed. A combination of the two techniques should give optimal results: the presence of light dots on non-textured areas should give 3D data where stereo by itself could not find points.

A strong interest of cars manufacturers, customers, governmental organizations has been identified for such 3D perception concepts. New "smart airbag" generation including partial inflation capabilities, should take into account information provided by such intelligent perception systems.

References

Boverie, S., Devy, M., Lequellec, J. M., Mengel, P., & Zittlau, D. (2000). 3D Perception for vehicle inner space monitoring. In *Advanced microsystems for automotive applications* (pp. 230–242).

Boyer, K. L., & Kak, A. C. (1987). Color-encoded structured light for rapid active ranging. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *9*(31).

Devy, M., Garric, V., & Orteu, J. J. (1997). Camera calibration from multiple views of a 2D object, using a global nonlinear minimization method. In *International conference on intelligent robots and systems*, Vol. 3 (pp. 1583–1589).

Faber, P., & Forstner, W. (2000). A system architecture for an intelligent airbag deployment. In *IEEE intelligent vehicles symposium*, Detroit, USA.

Gautama, S., Lacroix, S., & Devy, M. (1999). On the performance of stereo matching algorithms. In *Erlangen workshop*: *Vision*, *modeling and visualization* (8 p.).

Granjean, P., & Lasserre, P. (1995). Stereo vision improvements. In *IEEE international conference on advanced robotics*, September 1995 (pp. 679–685).

Krumm, J., & Kirk, G. (1998). Video occupant detection for airbag deployment. In *Workshop on applications of computer vision* (pp. 30–35), *IEEE Computer Society Press*, Los Alamitos, USA.

Lequellec, J. M., & Lerasle, F. (2000). Car cockpit 3D reconstruction by a structured light sensor. In *IEEE intelligent vehicles symposium* (pp. 87–92).

Lerasle, F., Lequellec, J. M., & Devy, M. (2000). Relaxation vs. maximal cliques search for projected beams labeling in a Structured Light Sensor. In *International conference on pattern recognition*, Vol. 1 (pp. 782–785).

Matthies, L. (1992). Stereovision for planetary rovers: Stochastic modeling to near real-time implementation. *International Journal on Computer Vision*, *8*(1), 71–91.

Mengel, P., Doemens, G., & Listl, L. (2001). Fast range imaging by CMOS sensor array through multiple double short time integration. In *International conference on image processing* (pp. 169–172).

Rosenfeld, J. P., & Tsikos, C. J. (1986). High-speed space encoding projector for 3D imaging. In *Optics*, *illumination and image sensing for machine vision* (pp. 146–151).

Stockman, G., & Hu, G. (1989). 3D surface solution using structured light and constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(4), 390–402.