N° d’ordre : 2526

THÈSE

présentée

pour obtenir

LE TITRE DE DOCTEUR DE L’INSTITUT NATIONAL POLYTECHNIQUE DE TOULOUSE

École doctorale : Informatique et Télécommunications

Spécialité : informatique

Par M Maxime Cottret

Titre de la thèse : Exploration Visuelle d’Environnement Intérieur par Détection et Modélisation d’Objets Saillants

Soutenue le 26 oct. 07 devant le jury composé de :

M.	Alain Ayache	Président
MM.	Michel Devy	Directeur de thèse
	Etienne Colle	Rapporteur
	François Brémond	Rapporteur
	Ryad Chellali	Membre
	Raja Chatila	Membre

Remerciements

Voila, je me retrouve enfin à écrire ma page de remerciements, clôturant ainsi cette grande aventure que furent les 4 années de mon doctorat.

Je tiens donc à remercier en premier lieu Raja pour la confiance qu'il m'a accordé en me proposant une bourse à la fin de mon stage de DEA. Ensuite vient naturellement Michel pour son encadrement, sa compréhension de mes idées souvent embrouillées, sa disponibilité (surtout sur la fin) et l'absence de contrainte sur la voie à suivre. Je tiens également à remercier plus particulièrement Etienne Colle et François Brémond pour m'avoir accordé un peu de leur temps à la lecture et critique de mon manuscrit, Alain Ayache pour avoir accepter la présidence de mon jury et Ryad Chellali pour avoir fait le déplacements depuis l'Italie pour assister à ma soutenance (et être passé à travers les mailles des grèves AirFr...).

Je remercie l'ensemble des permanents et thésards du feu groupe RIA :-) pour la très agréable ambiance de travail, avec une petite attention particulière Fred Lerasle pour ses retours sur mon manuscrit, Nizar pour son aide sur le bundle ajustement malgré le timing serré, les occupants du bureau SLAM Thomas et Abdelatif et enfin mes compères accros de caféine et de poker Seb, Nico, Alex, Fred Py, Léo et Tony.

Et pour finir, de manière non exhaustive, je remercie tout ceux qui m'ont soutenu (ou supporter, au choix) pendant ces 4 ans: Stéphanie, Julie, Juliette, Wassila, Julia, Julien E, Gillo, Gaétan, PH et toute la bande de c.....ds de Toulouse et navarre, mes deux groupes HSOS et Soulmate, les versaillais et assimilé: Sophie, Jérôme, Seb et Etienne (Maintenant, je vais être plus dispo) et bien entendu toute ma petite famille.

Table des matières

Table des figures

Chapitre 1

Introduction

1.1 vers la Robotique personnelle

Notre contribution concerne des fonctions perceptuelles nécessaires au développement de ce que sera peut-être demain le Robot Personnel ou Compagnon de l'Homme. Un tel robot doit rendre service à un homme, dans son lieu de vie, au domicile. De nombreux résultats ont déjà été obtenus dans le contexte de la robotique d'assistance aux personnes handicapées ou agées [Hoppenot 01]: les travaux en ce domaine ont surtout porté sur la robotisation des fauteuils roulants, c'est-à-dire l'intégration sur un tel système, de fonctions robotiques permettant de rendre plus facile la commande du fauteuil par son utilisateur: fonctions évoluées de commande référencée capteur (suivi de murs, franchissement de portes...), fonctions de détection d'obstacles ...cela en tenant compte des contraintes d'acceptabilité par l'utilisateur.

Notons que ces contraintes d'acceptabilité excluent généralement l'instrumentation de l'environnement avec des capteurs; les méthodes de surveillance étudiées dans le contexte du *lieu intelligent* ou de l'*Intelligence Ambiante*, notions très populaires depuis quelques années (voir les actes des nombreuses conférences spécialisées dans cette thématique: EUSAI pour *European Symposium on Ambient Intelligence*, PETS pour *Performance Evaluation for Tracking and Surveillance*...), ne sont pas ou sont peu exploités dans ce contexte, les personnes ne souhaitant pas être observées.

Plusieurs systèmes robotiques développés pour le Handicap, sont des fauteuils équipés d'un bras manipulateur: citons notamment le bras MANUS utilisé par de nombreux centres de recherche ou le projet *Clickrog: Object grasping for disable persons*, conduit par le CEA LIST et l' IRISA, qui a pour objet d'aider une personne à saisir un objet avec un bras embarqué sur un fauteuil . Dans des projets plus récents, ce n'est pas le fauteuil qui est robotisé, mais un robot personnel mis au service de la personne handicapée: par

exemple, le projet ASSIST (projet ANR PSIROB 2007, impliquant LIRMM, CEA LIST, LAAS-CNRS et LISIR), a pour but de développer un robot mobile, doté de deux bras mobiles, capable d'exécuter des actions simples non réalisables par la personne du fait de son handicap, par exemple, saisir un objet tombé au sol, aller chercher un objet dans une autre pièce du lieu de vie, objet éventuellement rangé dans un placard ou un réfrigérateur ...

De nombreux projets concernent donc le développement de robots personnels, compagnons de l'homme, dotés de fonctions spécifiques pour tenir compte des capacités sensorimotrices limitées de l'Homme. Dans le cas général, le développement d'un robot personnel nécessite de résoudre de nombreux problèmes, en particulier:

- planification et exécution de tâches à proximité ou même, au contact de l'homme (par exemple, pour donner un objet à l'homme),
- intégration dans le système robotique, d'un Interface Homme-Machine (IHM) permettant (1) à l'Homme d'indiquer la tâche à exécuter par le Robot et de préciser les modalités de cette tâche, et (2) au Robot de percevoir l'Homme si la tâche nécessite une intéraction sensorielle directe (le Robot doit suivre l'Homme, doit lui donner un objet ...) ou une reconnaissance des comportements de l'Homme pour détecter des situations à risques, pour comprendre ses intentions ...
- apprentissage de l'environnement partagé entre l'Homme et le Robot.

Nos travaux ont porté sur la perception de l'environnement: comment le Robot peut-il construire des représentations de son environnement de travail? quelles représentations? comment l'Homme peut-il intervenir pour, si nécessaire, introduire des informations contextuelles et plus généralement, pour rajouter des informations sémantiques (nommage de lieux, d'objets...) nécessaires dans le dialogue Homme-Robot?

Il existe de nombreux travaux en Robotique mobile, portant sur la construction de modèles de l'environnement. Comment décliner cette problématique dans le cas d'un robot personnel, compagnon de l'homme dans son lieu de vie?

- ce lieu de vie est typiquement un appartement, donc un ensemble de pièces meublées, contenant de nombreux objets, supposés statiques.
- ces objets peuvent être fixes (affiches, étagères... accrochées sur un mur ...) ou peuvent être déplacés par l'Homme (verres, bouteilles... posés sur une table; chaises autour de la table ...).
- chaque pièce, meuble et objet sera associé à un symbole donné par l'Homme: nous parlerons de *nommage* (pièce *Cuisine*, meuble *Table*, objet *Bouteille*...)

– l'Homme intervient donc au niveau sémantique; il peut aussi guider le Robot durant l'apprentissage de la représentation, pointer des objets intéressants ...
– par contre, les représentations construites doivent permettre l'exécution de tâches en mode autonome par le robot: *Aller-à-Lieu*, *Saisir-Objet*, *Chercher-Objet*...

Dans ce contexte d'environnement humain, il est intéressant de faire un parallèle entre l'apprentissage des représentations par un Robot, et celle d'un humain découvrant ce même appartement (par exemple, un enfant ou une personne en charge du service). Intéressant car un humain parvient très vite à se repèrer dans un nouvel environnement et que les mécanismes mis en oeuvre pour l'apprentissage de l'espace par l'Homme, sont au coeur de très nombreuses études dans les Sciences Cognitives. Néanmoins, il est clair que l'Homme exploite (1) de nombreuses modalités sensorielles, dont sa Vision, et (2) surtout, un substrat très important de connaissances contextuelles, accumulées depuis l'enfance. Or, la Vision Humaine est incomparablement plus performante que la Vision Artificielle dont est doté notre Robot, et il est impossible de doter notre Robot de toutes les connaissances acquises, même par un Homme enfant.

Ces limitations étant bien posées, nous nous sommes inspirés des études sur la Vision et la Cognition Humaine dans nos travaux sur la Vision en robotique: on parle dans la littérature, de *Cognitive Vision*.

1.2 Approche générale

Notre contribution porte donc sur l'apprentissage de représentations spatiales de l'environnement par un Robot exploitant uniquement des capteurs visuels. Les principales caractéristiques de notre approche sont décrites ci-dessous.

◇ Précisons tout d'abord que nos travaux s'inscrivent dans le cadre du projet européen COGNIRON, dédié à la conception et au développement d'un Robot Compagnon de l'Homme. Nous avons participé au lot de travail *Spatial cognition and multimodal situation awareness*. Le Robot apprend une représentation hiérarchique de l'environnement, qui comprend plusieurs niveaux: chaque partenaire s'est intéressé à un sous-ensemble de ce modèle. Pour notre part, nous avons participé aux travaux sur la construction d'un modèle qualitatif de type *Graphe d'Objets*, chaque objet étant représenté par son apparence selon plusieurs points de vue.

◇ Notre principale contribution concerne la phase d'exploration durant laquelle le

Robot apprend ce graphe d'objets. Pour ce faire, il se déplace dans l'environnement: soit il erre de manière aléatoire, soit il est piloté par l'Homme. Nous proposons une approche autonome, inspirée de la Vision Humaine, comportant deux processus visuels asynchrones, dits pré-attentif et attentif:

- avec le processus pré-attentif, notre Robot va rechercher des objets caractéristiques qui serviront de noeuds dans la représentation spatiale de l'environnement. Nous exploitons une caméra large champ (éventuellement, une caméra omnidirectionnelle), pour détecter, puis suivre durant les déplacements du robot, des régions d'intérêt. Chaque région contient potentiellement un objet isolé du fond (un crayon sur une table) ou un agencement de tels objets (un crayon, une gomme et un journal); nous appellerons une telle structure locale caractéristique, un **proto-objet**. Ces structures sont grossièrement localisées relativement au robot, puis introduites dans une *mémoire courte durée*.
- le processus attentif a pour mission de modéliser successivement chacun des proto-objets détectés par le processus pré-attentif. Nous exploitons pour ce faire une caméra PTZ, qui sera à terme asservie en orientation et zoom, sur le proto-objet en cours d'analyse. Pour tous les aspects significatifs, nous stockons l'apparence du proto-objet dans la *mémoire longue durée* du système: un aspect est significatif s'il est nouveau, dans le sens où sa ressemblance avec un aspect déjà perçu est inférieure à un seuil donné. Le proto-objet ainsi modélisé est rajouté au Graphe décrivant l'environnement.
- si le temps d'analyse du proto-objet courant est trop long, ou si la distance parcourue par le robot est trop importante, les régions d'intérêt non encore analysées, stockées en mémoire courte durée, sont oubliées.

◇ Une fois l'exploration terminée, plusieurs étapes, non traitées dans nos travaux, seraient nécessaires pour améliorer la représentation de l'environnement. L'Homme pourrait vérifier la cohérence du modèle: suppression des structures locales non rémanentes, décomposition des proto-objets en objets isolés, nommage de ces objets... Une phase de catégorisation supervisée ou automatique, permettrait de faire émerger des modèles d'objets génériques, donc de regrouper sous une seule représentation, plusieurs instances d'objets de même classe (des chaises, des crayons, des verres ...).

◇ En phase d'exploitation, la représentation de type Graphe d'Objets servira pour planifier et exécuter des déplacements: la représentation pourrait être mise à jour pour tenir compte des évolutions de l'environnement (suppression des objets non retrouvés à leur place, rajout d'objets connus, mais déplacés) ou pour apprendre les modèles de nouveaux

objets.

1.3 Organisation de ce mémoire

Nous avons étudié la phase d'exploration et proposé une première implémentation des processus pré-attentif et attentif sur des images acquises hors-ligne depuis une des plate-formes robotiques disponibles dans le pôle Robotique et IA du LAAS-CNRS.

Dans le chapitre suivant, nous préciserons les enjeux, les besoins et les contraintes de la robotique mobile de service. Nous définirons quelques notions essentielles de la perception de l'environnement et de la la cognition spatiale. Nous rappellerons les principales approches exploitées pour décrire des environnements intérieurs en robotique: approches géométriques, qualitatives et écologiques. Dans tous les cas, l'exploitation de représentations spatiales pour exécuter des déplacements, nécessite d'appliquer des techniques de reconnaissance de formes, pour identifier des primitives sensorielles, des amers ou des lieux; aussi, nous rappellerons les méthodes existantes pour la reconnaissance d'objets en vision artificielle.

Le chapitre 3 introduit plus précisément la représentation qualitative sur laquelle nous avons travaillé dans le cadre du projet européen COGNIRON, dédié à la robotique personnelle. Avant COGNIRON, nous avons participé au développement du robot RACKHAM, robot guide de musée évalué pendant plusieurs mois à la Cité de l'Espace de Toulouse; nous décrivons les représentations spatiales de nature métrique, topologique et sémantique introduites sur ce robot pour satisfaire les besoins et contraintes de la robotique en milieu public. Nous décrivons ensuite la hiérarchie des représentations développées par l'ensemble des partenaires de COGNIRON, avant de focaliser sur le modèle qualitatif que nous avons développé. Cette représentation est construite par une approche inspirée de la Vision humaine, exploitant un processus pré-attentif (*early vision*, *peripheral vision*...) pour détecter des régions saillantes, puis un processus attentif pour focaliser l'attention visuelle sur ces régions et pour en décrire l'apparence selon divers aspects.

Le chapitre 4 décrit le processus pré-attentif, qui exploite une caméra large champ ou omnidirectionnelle, pour détecter des régions saillantes, dans la séquence d'images acquises en continu pendant les déplacements du robot; nous avons implémenté le concept de carte de saillance proposé par L.Itti. Nous avons ensuite évalué plusieurs méthodes

(corrélation, CAMSHIFT...) pour suivre les régions détectées dans les images suivantes de la séquence, cela pour pouvoir grossièrement déterminer la position et la taille 3D de ces régions relativement à la position courante du robot.

Le chapitre 5 décrit le processus attentif: chaque région saillante correspond à une structure locale, appelée par la suite *proto-objet* car de ces structures, émergeront les objets appris par le système. Après focalisation et asservissement d'une caméra PTZ sur la région d'intérêt, nous apprenons un modèle d'apparence de ce proto-objet, sous la forme des points d'intérêt visibles et appariés sur les vues successives; la focalisation permet de faire abstraction du fond de la scène. Nous montrons comment ces modèles sont exploités pour la reconnaissance par des techniques d'indexation.

Enfin, dans la conclusion, nous résumerons quelles ont été nos contributions, et nous évoquerons les nombreuses pistes laissées ouvertes par ce travail; au delà de l'intégration sur un robot afin de traiter des images en ligne, nous sommes conscient qu'il reste de nombreux progrès à faire en Vision, pour qu'un Robot Cognitif apprenne une représentation spatiale de l'environnement qui s'apparente aux représentations exploitées par les humains.

Finalement, deux annexes fournissent des détails techniques sur la plateforme robotique RACKHAM que nous avons exploitée pour valider une partie de nos travaux, puis sur l'environnement de développement JAFAR que nous avons construit avec notre collègne doctorant Thomas Lemaire, pendant notre séjour dans le pôle Robotique et Intelligence Artificielle du LAAS-CNRS.

Chapitre 2

Compréhension de l'environnement

La robotique personnelle ou en milieu public, pour améliorer en particulier les conditions de vie des personnes handicapées ou agées, est un défi pour notre communauté scientifique, défi déjà relevé dans les pays développés où le problème du vieillissement de la population est le plus aigu, comme Japon et Allemagne. Nous allons d'abord rappeler les besoins et contraintes de ces applications de la robotique.

Pour être accepté par l'Homme, un Robot qui doit intervenir dans un milieu humain, doit se comporter de manière prévisible pour l'Homme. Accessoirement il peut avoir une apparence humaine, d'où l'attrait de notre communauté pour la robotique humanoïde, ou pour l'expréssivité du robot; citons le projet Kismet du MIT, dont le but était "seulement" de développer une tête stéréo active, munie de degrés de liberté actionnés en fonction de l'état du l'IHM: compris/non compris, content/non content ...

Il est tentant alors de s'inspirer des Sciences Cognitives, des connaissances sur la cognition humaine, pour concevoir des systèmes artificiels amenés à partager l'environnement de l'Homme. Pour bien comprendre la démarche entreprise, il est intéressant de rappeler le contexte de développement de la *robotique cognitive*. En particulier notre sujet porte sur la perception de l'environnement: nous allons donc rappeler les différents modèles proposés sur la cognition spatiale et sur la reconnaissance d'objets, pour l'Homme et pour le Robot.

2.1 Robotique mobile de service

Il y a encore quelques années, les principales recherches en robotique tendaient à élaborer des systèmes précis et rapides, adaptés à une tâche spécifique dans un environnement connu. Poussés par les besoins industriels de performances et de réductions des coûts, ces systèmes étaient destinés à automatiser des tâches fastidieuses réalisées jusqu'alors par

l'homme: robotique de manipulation sur les chaînes d'assemblage, mais aussi robotique mobile pour la logistique dans des ateliers ou des grands dépots de marchandises. Souvent l'environnement est instrumenté pour simplifier les fonctions robotiques (suivi de lignes au sol, localisation par un réseau de capteurs déployés sur le site...)

Même avec de tels aménagements, nous savons que l'exploitation d'un robot mobile, a fortiori d'une flotte de robots, est encore difficile, du fait de la nécessité de prendre en compte toutes les incertitudes, tous les événements non prévus... afin d'arriver à des taux d'échec très faibles. Chacun connait bien sûr, une *Success Story*; citons le déploiement de robots de service dans un hopital de New-York par l'entreprise SwissLog, exploitant le savoir-faire de l'institut FZI de Karlruhe, mais chacun connait aussi les difficultés de telles applications. Aujourd'hui, l'évolution technologique et le développement de la recherche en "Intelligence Artificielle" permettent d'entrevoir de nouveaux champs d'utilisations des robots et la conception d'une véritable robotique de services. Citons quelques perspectives d'utilisation de robots, et analysons les besoins et contraintes à satisfaire en robotique.

2.1.1 Nouvelles Perspectives d'utilisation

Nous ne prétendons pas donner ci-dessous une liste de tous les domaines qui sont ou qui pourraient être de nouveaux champs d'application en Robotique. Nous donnons les domaines qui nous semblent les plus significatifs.

1. **sécurité civile:** la sécurité civile regroupe sous une même bannière un ensemble d'organismes comme les sapeurs-pompiers, les militaires d'intervention civile ou les démineurs qui sont appelés à intervenir sur des sites où des vies humaines sont menacées suite à une catastrophe naturelle ou l'écroulement d'un immeuble. Ces sites peuvent se montrer particulièrement hostiles et instables (site radioactif de Tchernobyl, les décombres du World Trade Center ou plus simplement l'incendie d'un immeuble) et donc rendre les interventions difficiles et dangereuses. Un robot d'intervention doté de capacités d'évolution, de compréhension du terrain et de manipulations permettrait d'assister efficacement les sauveteurs et même d'intervenir à des endroits hors d'atteinte de l'homme.

   On connait les échecs de la robotique à Tchernobyl, mais c'était en 1986 et la technologie n'était pas prête. Plus recemment, l'armée américaine a utilisé des robots similaires au *PackBot* commercialisé par iRobot aux Etats-Unis, pour détecter des présences humaines, dans les décombres du World Trade Center, ou dans les grottes en Afghanistan: face à la difficulté du contexte, ces expériences n'ont pas non plus été concluantes.

Ce sont des applications très exigeantes, pour des réseaux de robots et de capteurs déposés dans des sites dangereux. L'homme n'est pas très loin, les délais de communication sont faibles et les bandes passantes sont élevées en communication sans fil: donc l'intéraction avec des opérateurs experts permet de limiter la complexité des systèmes robotiques à développer.

2. **exploration extra-planétaire:** depuis des millénaires, l'Homme a les yeux tournés vers le ciel et rêve de pouvoir explorer notre Univers. Cette volonté s'est concrétisée depuis les années 50, d'abord grâce à l'utilisation de satellites puis par le développement des vols spatiaux habités. Néanmoins, ces deux modes d'exploration ne sont pas forcément utilisables dans le cadre d'explorations planétaires: pas de navigation en surface possible pour les premiers, contraintes de développement et de mise en place extrêmes pour les deuxièmes. La robotique mobile autonome se place donc comme une alternative de choix; sans parler d'autonomie, la NASA a déjà réalisé quelques expérimentations fructueuses sur Mars avec les robots *Sojourner, Opportunity et Spirit*.

   Dans ce contexte, l'homme est loin, et les délais de commuication importants (8mn sur Mars). La téléopération ou la supervision par l'Homme sur terre, est encore une obligation du fait de la faible capacité de calcul embarquée sur les robots d'exploration planétaires: un déplacement de quelques mêtres pour déployer un capteur auprès d'un rocher à analyser, prend plusieurs heures. On voit néanmoins se développer des fonctions autonomes qui permettront à terme, de minimiser le rôle de l'opérateur expert.

3. **loisirs:** Bien que créée pour fournir un cadre d'expérimentation de travaux sur les approches coopératives multi-robots, la "Robocup"[1] donne une idée de ce que pourrait être une forme de jeu de demain, où les robots prendraient la place des acteurs virtuels des jeux vidéos d'aujourd'hui. De même, le succès du robot chien AÏBO© de SONY ouvre une voie vers la robotique domestique ludique.

   Si ces applications sont intéressantes pour analyser le comportement de l'Homme (adulte ou enfant) face à un robot pouvant imiter des comportements biologiques, le manque de tâches précises à exécuter par le robot en limite l'intérêt.

4. **robot compagnon:** C'est l'extension de la robotique ludique. Et si ce robot à domicile, pouvait exécuter des tâches utiles pour l'homme? Qui n'a jamais rêver d'avoir un robot capable de faire le ménage, le repassage et le café ? Plus sérieusement, l'introduction d'un robot assistant "intelligent" pourrait rendre de grands services, notamment aux personnes âgées ou handicapées.

1. http://www.robocup.org/

C'est à cette application que nous nous sommes confronté. Il n'y a pas ici d'aspect multi; l'opérateur est un simple utilisateur, non expert en robotique, mais toujours le même; le Robot pourrait progressivement, par des techniques d'apprentissage, adapter son IHM aux besoins et préférences de l'Homme. L'environnement est typiquement un appartement.

5. **robot de service en lieu public**: c'est encore une extension du cas précédent. Sortons notre robot de son appartement, et mettons le au service du public dans un musée (robot guide), un centre commercial (caddy robotisé), dans des rues piétonnes en site urbain (*Person Mover*, logistique...). Il interagit avec un utilisateur principal non expert, mais qui change à chaque session; il pourrait reconnaître les besoins spécifiques de cet utilisateur et adapter en conséquence son comportement. En plus il doit se coordonner avec ses semblables ou avec les autres humains pour partager un espace d'évolution éminement dynamique et évolutif: imaginez une flotte de caddys robotisés dans votre grand magasin favori aux heures de pointe!

2.1.2 Besoins et Contraintes

Ces nouveaux systèmes robotiques quittant leur niche déterministe de "l'Automatisme", ils doivent satisfaire de nouveaux critères et de nouvelles contraintes.

Autonomie: les contraintes liées aux développements de tels robots sont multiples et rendent l'intégration d'un tel système particulièrement complexe: - environnements partiellement ou totalement inconnus, avec un contexte a priori plus ou moins important (un environnement urbain contient un certain nombre de caractéristiques invariantes, connues alors qu'un site d'exploration martien ou de catastrophe naturelle ne peut être pré-modélisé). Dans tous ces cas de figures, le robot est appelé à agir plus en coopération avec l'homme que guidé par lui, dans des situations beaucoup plus variées et imprévisibles, demandant un système au fonctionnement dynamique, adaptatif et ouvert.

Nous parlerons alors de robot autonome, autonomie que nous pouvous définir sous deux angles distincts:

- le point vue technologique, qui fait principalement référence à sa durée d'utilisation, en exploitant ses sources d'énergie internes ou une énergie tirée de l'environnement naturel (énergie solaire), sans recours à des sources d'énergie externes (recharge sur le réseau électrique ou ravitaillement en carburant).
- le point de vue plus philosophique qui se résume comme suit : *c'est la faculté d'un individu de se déterminer soi-même, de choisir et d'agir librement, suivant sa propre*

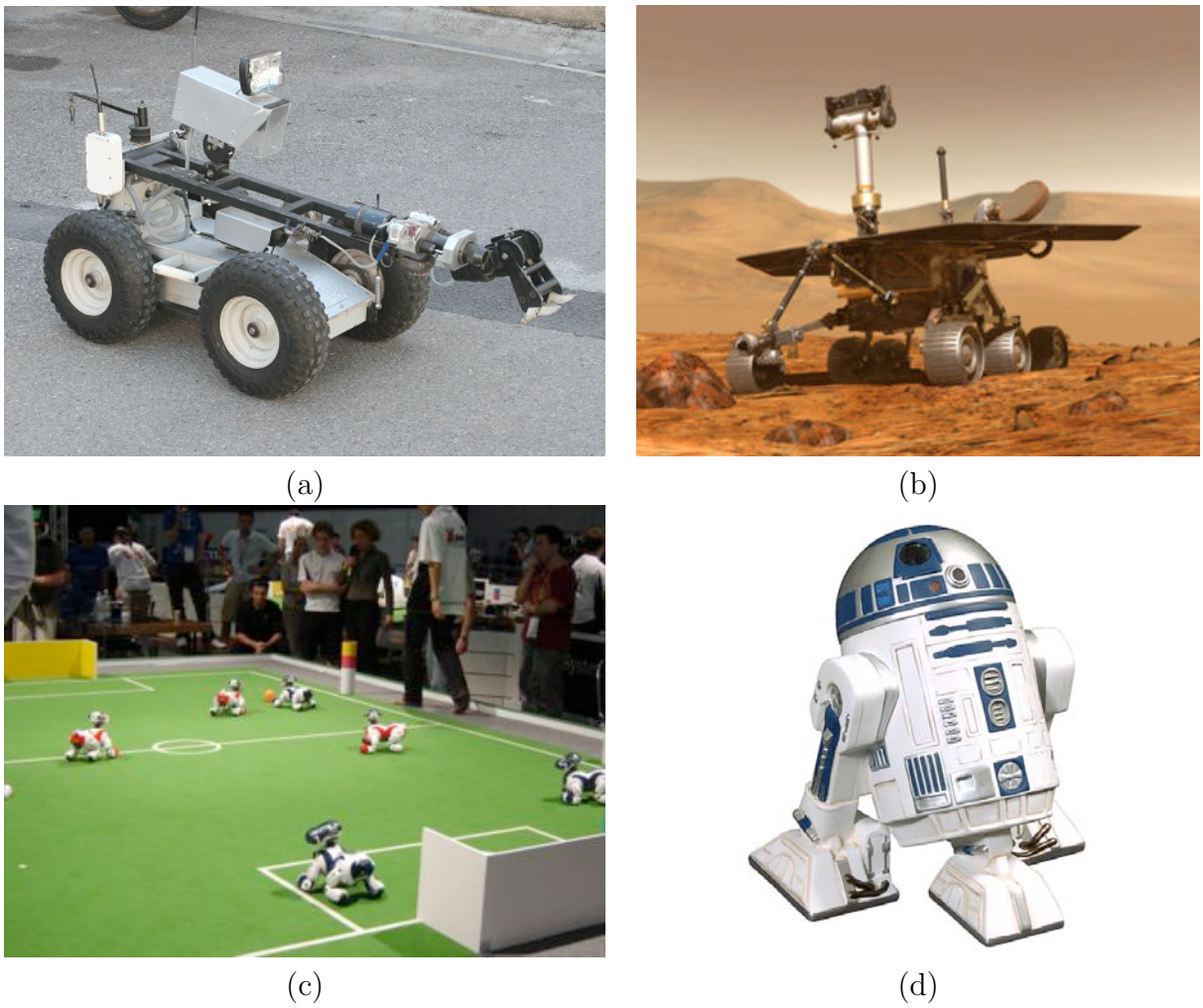

FIG. 2.1 – *(a) robot d'intervention civile pour le déminage - (b) SPIRIT, robot d'exploration martienne de la NASA - (c) Match entre robots AÏBO© de SONY lors de la Robocup - (d) le célèbre robot compagnon R2D2 de la saga StarWars*

loi. Cela se traduirait en robotique par des capacités accrues d'adaptations à toutes situations, de prises de décision, d'actions ou de perceptions sans aucune intervention d'un opérateur humain.

Une manière réaliste en milieu intérieur (pas de capteurs solaires), de traiter de l'autonomie énergétique, est de doter le robot de la capacité d'évaluer ses ressources courantes (état de charge des batteries) et de déplacement autonome vers la station de recharge.

Dans ce contexte, les **approches cognitives** tendent à donner au robot l'autonomie d'action, par l'étude et la compréhension des mécanismes de la pensée humaine, animale ou artificielle, et plus généralement de tout système cognitif, c'est-à-dire tout système complexe de traitement de l'information capable d'acquérir, conserver, et transmettre des

connaissances.

Il est difficile d'évaluer les progrès dans ce domaine, de par la compléxité des systèmes et concepts mis en oeuvre. Ainsi, depuis quelques années, un certain nombre de travaux sur les méthodes d'évaluation de l'autonomie de robots de services ont vu le jour comme [Kamsickas 03, Huang 04, Lampe 06]. Pour les applications militaires aux Etats-Unis en particulier (mais la DGA a repris ce concept en France) ont été définis des **niveaux d'autonomie**; à chaque niveau d'autonomie du robot, il correspond aussi un niveau de complexité de l'IHM, et donc des contraintes sur le réseau de communication entre le robot et une station de contrôle opérateur.

intégration sociale: Les robots compagnons étant amenés (à plus ou moins long terme) à intégrer l'environnement quotidien de l'homme, il devient indispensable de se pencher sur les questions de la socialisation du Robot et de son acceptabilité par l'Homme

Le premier point important que le robot doit acquérir est la notion de proxémie[2], distance qui s'établit naturellement entre deux individus en interaction, afin de faciliter son acceptation par l'homme. Cette distance varie en fonction de l'appartenance culturel, du niveau social et du niveau d'intimité des individus (figure 2.2) ainsi que de l'environnement dans lequel prend place l'interaction Le robot doit donc être capable d'apprendre les différentes manières d'approcher un individu et de maintenir une certaine distance avec celui-ci. Idéalement, il devra être capable d'adapter, au fur et à mesure des interactions, la distance sociale préférentielle d'un individu ou d'un groupe d'individus.

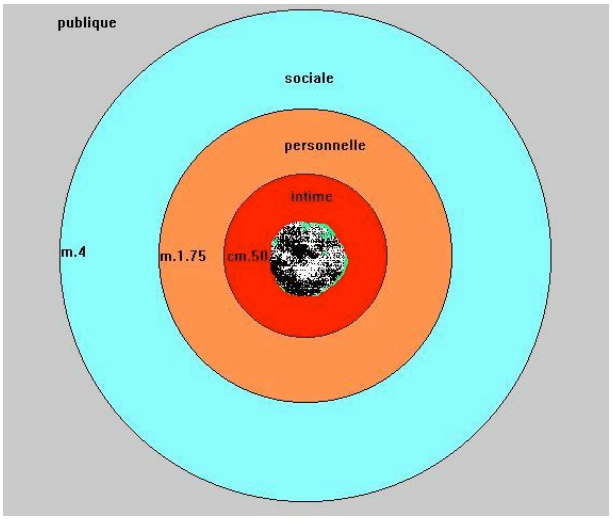

FIG. 2.2 – *Distances d'interaction en fonction du degré d'intimité établit par E. Hall*

2. terme introduit en 1963 par l'anthropologue américain Edward Hall

Le deuxième point porte sur l'engagement d'une interaction avec un individu. Le robot compagnon doit se comporter de manière socialement acceptable par l'individu tout en menant à bien sa tâche. Pour cela, il doit pouvoir comprendre l'activité de l'individu, ses intentions et ainsi être prêt à interagir. A l'inverse, l'individu doit pouvoir comprendre facilement les intentions et les actions du robot. Ce sont ces contraintes qui justifient les nombreux travaux actuels sur la reconnaissance de l'expression (le Robot doit savoir si l'Homme est content, non content, interrogatif ...) et du comportement de l'Homme (le Robot doit analyser la position ou les gestes de l'Homme afin d'adapter son comportement ou d'inférer comment il peut rendre un service), mais aussi sur l'expressivité du Robot (nous avons cité Kismet, mais certains chercheurs au Japon ont déjà conçu des robots imitant les expressions humaines de manière très réaliste).

Des ressources limitées: Nous avons donc à faire à des systèmes de plus en plus complexes, gérant de multiples capteurs et devant traiter un nombre d'informations exponentielle. Malgré l'augmentation constante de la puissance des calculateurs, la puissance embarquée sera toujours limitée par les capacités de charges utiles plus ou moins critiques que peuvent embarquer les robots. Par exemple, dans le cadre du drône Lhassa du LAAS-CNRS, la charge utile est seulement de 6 kg.

Dans de nombreuses applications, il existe une station de contrôle, station SOL en robotique spatiale par exemple; des opérateurs experts sont dans cette station pour surveiller le comportement du robot à distance, et pour éventuellement le commander en téléopération ou téléprogrammation. La capacité de communication entre le robot et la station de contrôle est également une ressource critique souvent limitée soit en débit soit en temps (problème des fenêtres de transmissions des expéditions martiennes par exemple). De plus, les transmissions sans fils peuvent rencontrés de nombreux obstacles pouvant empêcher momentanément une transmission. Il est donc nécessaire de développer non seulement des outils de télécommunications robustes mais également des capacités d'adaptation du comportement du robot pour gérer efficacement ses ressources en toute situation.

Nous allons nous intéresser à la robotique personnelle; il n'existe pas en ce cas d'opérateur expert. Même dans cette configuration, il existe des solutions technologiques pour donner à un robot physique, un niveau élevé d'autonomie sans embarquer à bord toutes les capacités fonctionnelles de perception ou de décision: c'est le concept de *Remote-Brained Robot* proposé en particulier à l'Université de Tokyo. Le robot physique est connecté à une station de contrôle; les algorithmes complexes de décision (planification de tâches ou de trajectoires) ou de perception (analyse d'images) peuvent être exécutés sur cette

station non soumise aux contraintes propres aux systèmes embarqués (énergie, compacité, robustesse aux chocs et vibrations...). Cette station peut aussi exploiter des données sensorielles acquises par des capteurs disséminés dans l'environnement, simplifiant d'autant les capacités de perception embarquées (capteurs et puissance de calcul associée).

Dans nos travaux, nous avons considéré que le système robotique se réduit au robot physique: il doit avoir toutes ses capacités fonctionnelles et décisionnelles embarquées. L'environnement n'est pas instrumenté (le robot n'est pas dans un lieu intelligent); cela répond aux contraintes déjà évoquées d'acceptabilité de la robotique dans un lieu social, en particulier par des personnes handicapées.

Comment un tel système représente l'environnement dans lequel il doit se déplacer, ou les objets qu'il doit saisir? Comment exploite-t'il ces représentations pour naviguer, se localiser, reconnaître des objets? Nous allons donner l'état de l'art sur ces problèmes, dans les deux sections suivantes de ce chapitre.

2.2 La perception de l'environnement en robotique mobile

Comme nous l'avons vu dans la section précédente, le robot de service doit avoir un comportement complexe, adapté à son environnement et à sa tâche. Pour une grande partie, son étude et son développement s'inspirent fortement des sciences dites "cognitives": nées à la fin des années 50, elles forment un ensemble de disciplines scientifiques visant à l'étude et la compréhension des mécanismes de la pensée humaine, animale ou artificielle, et plus généralement de tout système cognitif, c'est-à-dire tout système complexe de traitement de l'information capable d'acquérir, conserver, et transmettre des connaissances. Les sciences cognitives reposent donc sur l'étude et la modélisation de phénomènes aussi diverses que la perception, l'intelligence, le langage, le calcul, le raisonnement aux travers d'un certain nombre de fonctions de traitement de l'information.

Dans notre contexte, nous allons nous intéresser à l'apprentissage et à l'exploitation de représentations qui vont permettre au robot mobile d'exécuter des tâches, soit de navigation (déplacement de la plateforme mobile de la position courante à une position but dans l'environnement), soit de manipulation (contrôle d'un bras éventuellement monté sur la plateforme mobile, afin de saisir un objet). L'exécution de telles tâches requiert de nombreuses connaissances sur l'environnement et sur les objets. Depuis vingt à vint-cinq ans, les communautés Robotique et Vision ont proposé des modèles et des fonctions d'apprentissage associées; le problème majeur est de tenir compte des imprécisions des

données sensorielles et des incertitudes dans tons les processus décisionnels (mises en correspondance, identification...).

C'est plus récemment, dans les années 90, que sont apparues des convergences entre les communautés Robotique, Vision Artificielle et Neurosciences, donnant naissance aux approches bio-inspirées de la Robotique et aux opérateurs inspirés de la Vision humaine, tels que la recherche de primitives invariantes, les approches d'indexation d'images exploitées pour la reconnaissance d'objets...

Nous allons donc nous intéresser dans cette section aux modèles cognitifs liés à la perception spatiale de l'environnement et leurs homologues computationnels.

2.2.1 Notions de cartes cognitives

Tolman [Tolman 48], fût le premier à introduire le concept de représentations internes de l'environnement et à apporter les premières preuves des mécanismes internes du cerveau mis en jeu. En usant d'apprentissage par renforcement (que l'on peut assimiler à un conditionnement du comportement par une succession d'expériences) sur le comportement et l'orientation de rats dans un labyrinthe à la recherche de nourriture, il conclut que le simple postulat des approches comportementales visant à décrire un comportement comme une réponse à une suite de stimuli n'était pas suffisant pour décrire les capacités de navigation des rats. Ainsi, les rats doivent avoir appris une représentation interne, une "carte cognitive" qu'il définit comme suit:

Dans une carte sensorielle, l'environnement est représenté d'une manière formelle. Si le point de départ de l'animal ou les routes qu'il a empruntées précédemment changent, cette carte formelle lui permettra de continuer à s'orienter plus ou moins correctement et à choisir le bon chemin.

Tolman a également rapporté des cas expérimentaux où le rat, mis dans une configuration du labyrinthe autre que celle d'apprentissage, escaladait les murs pour sortir du labyrinthe et se rendre directement à l'endroit où était déposé la nourriture. Ces expériences ont donc permis de mettre en évidence le fait que les rats apprenaient plus qu'une simple succession d'associations stimuli-réponse et élaboraient une représentation de l'environnement qui leur permettait de généraliser les chemins appris.

2.2.2 Erreur systématique et nature multimodale des représentations cognitives de l'environnement

Il est difficile d'élaborer un système artificiel cognitif sans s'appuyer sur les diverses études de la cognition humaine. Dans le cadre qui nous intéresse de la perception de l'envi-

ronnement, de nombreuses expériences ont montré que l'homme a tendance à commettre des erreurs systématiques lorsqu'il agit dans son environnement ou lorsqu'il tient des raisonnements sur la spatialité de cet environnement, indiquant clairement dans certains cas que les représentations cognitives de l'espace répondent à une structure particulière.

Dans une expérience bien connue de [Stevens 78], il fut demandé à un ensemble d'étudiants de situer relativement San Diego (Californie) et Reno (Nevada). La grande majorité des étudiants placèrent San Diego à l'ouest de Reno, alors qu'en réalité, elle se trouve largement au sud et plutôt à l'est (cf figure 2.3).

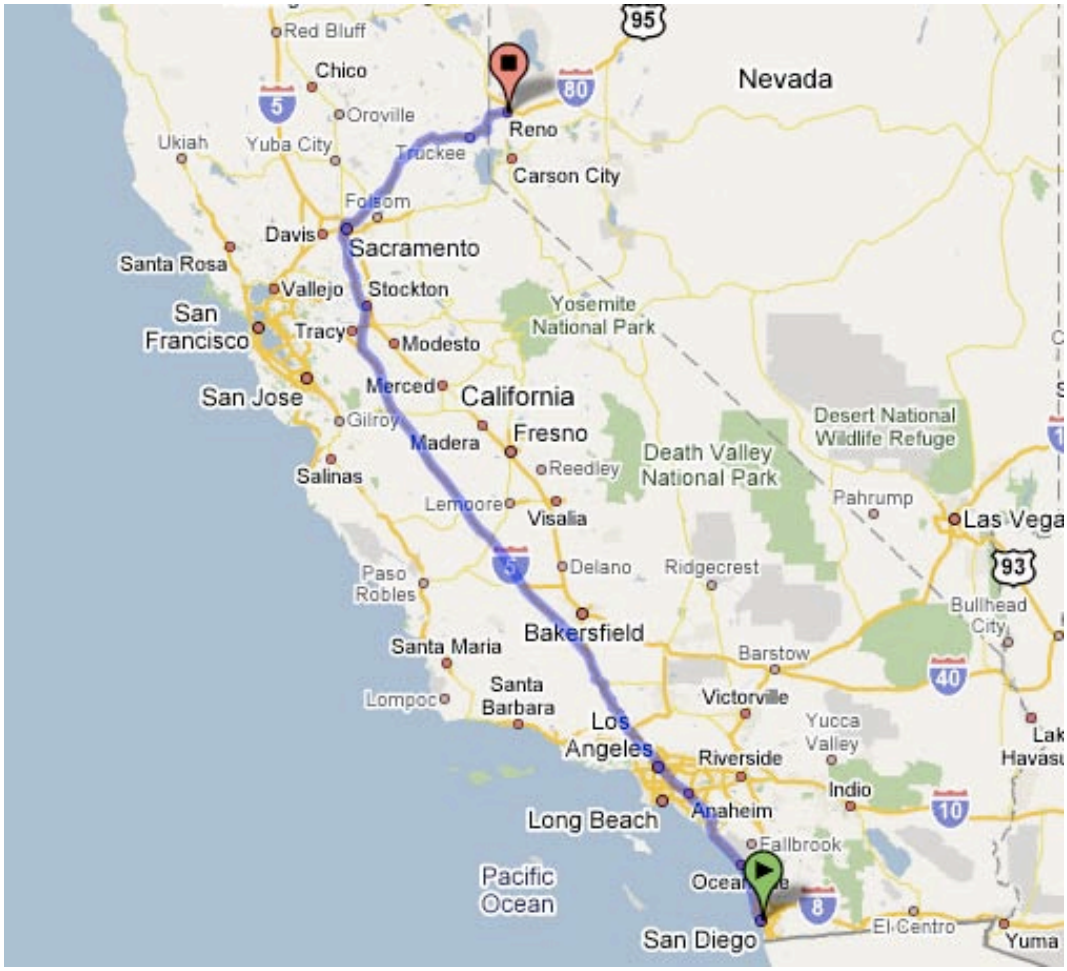

FIG. 2.3 – *trajet San Diego - Reno*

Cela suggère que d'une certaine manière, la mémoire spatiale est organisée hiérarchiquement: plutôt que d'intégrer la position absolue d'une ville, l'homme va la retenir comme appartenant à un état et la positionner relativement à cet état. Ainsi, la Californie étant situé globalement à l'ouest du Nevada, l'homme aura toujours tendance à situer

de prime abord toutes les villes de Californie à l'ouest des villes du Nevada. D'autres expériences montrent qu'une prise de position - par exemple, raisonner sur l'espace en s'imaginant dans une position particulière-, engendre systématiquement des distorsions dans le jugement des distances, de telle façon que la zone autour de la position est estimée plus large que la même zone située à une plus grande distance. De même, l'homme décrivant et mémorisant une position en fonction de points de repère de la zone environnante, il induit une distorsion asymétrique dans son appréciation des distances: la distance d'un objet quelconque à un point de repère apparaîtra toujours plus petite que la distance du point de repère.

Plusieurs types d'erreurs systématiques relatives au raisonnement spatial ont été mises en évidence et dans la majorité des cas, appréciations des distances et des directions se retrouvent perturbées par des influences cognitives de "plus haut niveau", comme le point de référence (réel ou imaginaire), la structure hiérarchique d'un souvenir, etc... Cela tend à montrer que la représentation spatiale cognitive chez l'homme ne se résume pas uniquement à une représentation métrique, mais rassemble un ensemble d'informations de natures différentes.

Selon [Tversky 93], deux types de représentations cognitives semblent émerger des diverses expériences menées sur le sujet. (1) Dans la plupart des situations, lorsque la connaissance de l'environnement est relativement grossière, la représentation cognitive de l'espace apparaît plus comme une collection éparse d'informations de différentes natures que comme une carte classique; elle y intègre des données de type local et relationnel - comme le fait qu'un point de repère soit à gauche d'un autre - ou des indices visuels relatifs à un point de repère. (2) Dans le cas d'un environnement simple ou bien connu, elle se rapprocherait des cartes que nous avons l'habitude de manipuler, c'est-à-dire faisant preuve d'une modélisation plus on moins précise, permettant de se situer, de raisonner, d'inférer.

2.2.3 Application à la robotique mobile

En robotique mobile, de nombreuses représentations de l'environnement, plus ou moins inspirées des cartes cognitives vues précédemment, sont utilisées dans le cadre de planification de mouvement, de contrôle d'exécution ou de raisonnement géométrique. Nous n'évoquons ici que les modèles adaptés aux environnement intérieurs, celui dans lequel va agir notre Robot compagnon de l'Homme: le sol est plat (nous éludons le cas des escaliers ou des différents niveaux qui peuvent exister dans un appartement). Le problème est totalement différent pour un robot de service en milieu extérieur, par exemple une voiturette suivant un joueur de golf devra exploiter un modèle 3D du terrain, devra discriminer la

pelouse des massifs floraux ...

Par ailleurs nous évoquons ici les modèles exploités pour la navigation: amers de localisation, espace navigable au sol, obstacles... Les modèles nécessaires à la manipulation sont du même ordre, mais sont forcément 3D et locaux: objets sur une table...

Il existe un lien étroit entre:

- le type de représentations: discrète/continue, éparse/dense, métrique/qualitative...
- les fonctions d'apprentissage qui permettent de les construire,
- et les fonctions de navigation qui les exploitent.

Pendant l'apprentissage le robot utilise pour se localiser et pour générer des trajectoires, les fonctions de navigation sur la portion du modèle déjà appris ; l'environnement étant dynamique et évolutif, le modèle doit être remis à jour en permanence; de ce fait, on peut considérer que les fonctions d'apprentissage et de navigation coexistent en permanence sur le robot, l'apprentissage étant activé uniquement quand des différences sont détectées entre modèle courant et données capteur.

La représentation classique des roboticiens se fonde sur la géométrie, donc sur des données métriques; le robot doit reconstruire son environnement en 3D à l'aide de sa vision ou disposer de capteurs tridimensionnels (télémétrie-laser, capteurs optiques à temps de vol...). La représentation géométrique est généralement exprimée dans un repère unique, dit *repère du monde*, défini souvent par la première position du robot dans l'environnement lors de l'apprentissage des représentations.

Au fil du temps et des besoins, différents types de cartes ont fait leur apparition:

Grille d'occupation (figure 2.4): La grille d'occupation est probablement la représentation de l'espace libre la plus simple, basée sur une discrétisation en cellule élémentaire. Une information binaire est affectée à chaque cellule suivant qu'elle est occupée ou non par un obstacle. Pour rendre compte des erreurs des capteurs et des erreurs de cartographie, il est plus courant d'utiliser un score d'occupation probabiliste. Ce type de modèle fût popularisé par [Moravec 88, Elfes 89], puis étendu dans sa forme probabiliste par [Thrun 96, Fox 98]: cela donne essentiellement une représentation de l'espace libre dans lequel le robot pourra se déplaçer. Lui sont associées les fonctions de planification sur *bitmap*, les procédures d'évitement sur cartes glissantes...

Carte de primitives géométriques (figure 2.5): Très utilisées depuis les débuts de la robotique mobile, les cartes de primitives géométriques représentent l'environne-

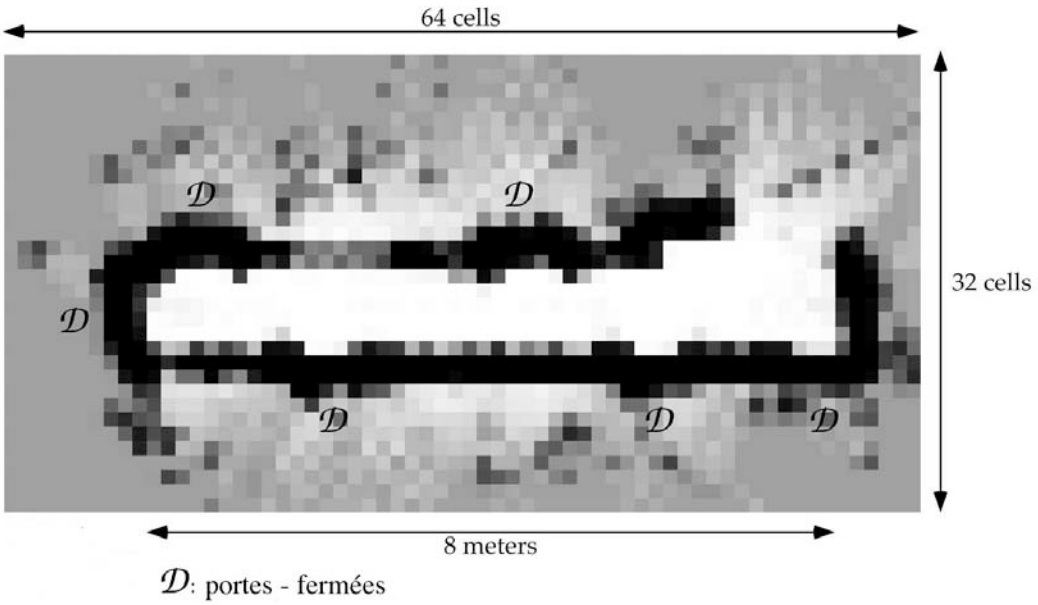

FIG. 2.4 – *Grille d'occupation : les cellules noires caractérisent une probabilité de 1 qu'elles soient occupées par un obstacle*

ment comme un ensemble de primitives (points, lignes [Jensfelt 01], régions de profondeur constante [Leonard 92], *etc*) dont les coordonnées sont exprimées dans un repère fixe donné. Une stratégie de mise en correspondance et d'alignement des données perçues par les capteurs avec une carte définie a priori est alors utilisée pour se localiser et naviguer. Leur principal avantage est d'être très performante, aussi bien au niveau calculatoire qu'en taille mémoire. C'est pourquoi elles sont principalement utilisées dans les travaux sur la localisation et la cartographie simultanée ou SLAM[1]. [Moutarlier 89, Leonard 91, Castellanos 01, Davidson 02]: ce sont des cartes éparses, sans aucune information sur l'espace libre, même si les cartes de segments laser ont pu être exploitées dans les travaux de Moutarlier ou Bulata par exemple au LAAS, pour générer un modèle de l'espace libre sous la forme d'un ensemble de polygones ou d'un *bitmap*.

Carte topologique (figure 2.6): Une carte topologique est une modélisation qualitative discrète de l'environnement - c'est-à-dire sans considération géométrique. Généralement, une carte topologique est représentée sous forme d'un graphe de lieux, rendant ainsi compte de certaines relations "topologiques" (connectivité, relation de tout à partie, ...) entre zones caractéristiques de l'environnement. Le plan de métro de Paris forme un exemple typique de carte topologique. Les premières applications à la robotique appa-

1. en anglais: Simultaneous Localization And Mapping

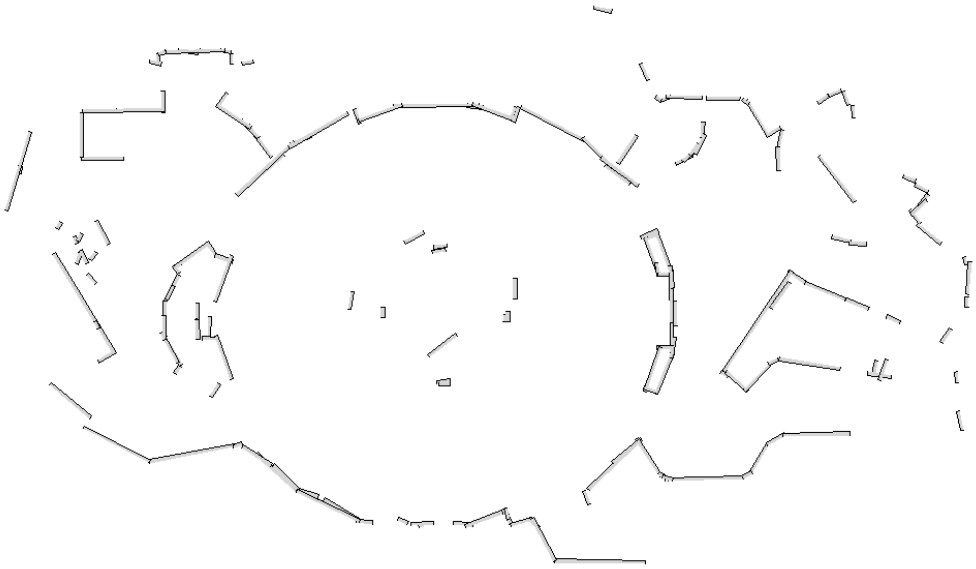

FIG. 2.5 – *Carte de primitives géométriques - ici utilisation de segments laser 2D*

raissent dans [Korthenkamp 94] et ont été popularisées plus récemment par [Choset 01], grâce notamment à l'utilisation des graphes de Voronoï généralisés (ou GVG); pour être exploitées, ces représentations nécessitent d'embarquer sur le robot, des primitives de mouvement sensori-motrices, telles que *Suivre-Couloir*, *Suivre-Mur*, *Aller-vers-Objet*...

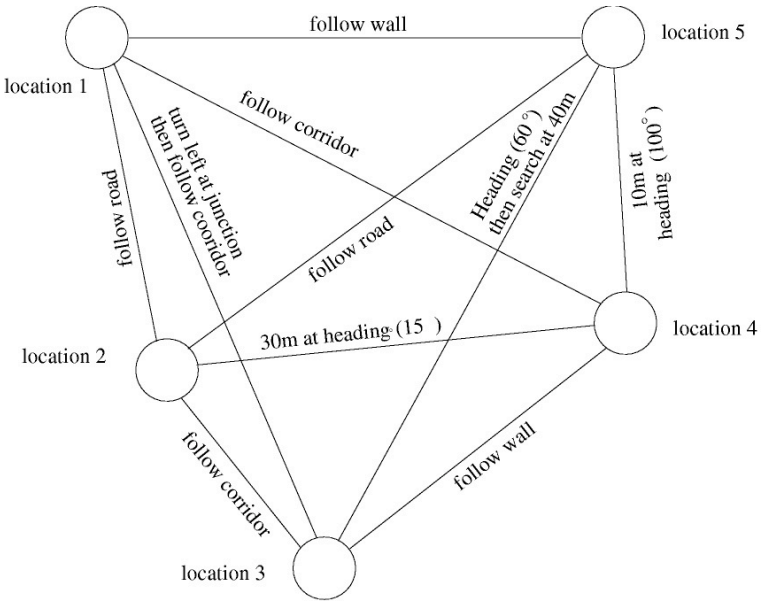

FIG. 2.6 – *Carte topologique d'un ensemble de lieux*

Carte d'apparence (figure 2.7): Pour la modélisation, les cartes d'apparence [Crowley 98, Kröse 01, Gonzalez-Barbosa 05] s'appuient directement sur les données capteurs, mémorisées suivant un référentiel global lié à l'environnement. L'apparence peut être aussi bien visuelle (issue d'une caméra) que télémétrique (issue d'un laser ou d'un sonar). Leur utilisation s'appuie sur la même stratégie que pour les cartes de primitives, la mise en correspondance pouvant être avantageusement faite directement sur les données brutes des capteurs. Cette représentation s'apparenterait le plus aux méthodes cognitives de l'Homme; elle suscite un grand nombre de travaux en cette période. Pour satisfaire les contraintes de ressources limitées des systèmes embarqués, elle doit être rendue éparse durant la phase d'apprentissage, et doit être analysée avant exploitation pour la navigation [Remazeilles 05]:

- seules sont conservées des images-clé, correspondant à des changements importants des scènes perçues (par exemple lorsque le robot franchit une porte).
- pour chaque image, seule est conservée l'information qui permettra l'indexation de la base: typiquement un vecteur global d'attributs (histogrammes...) ou un ensemble de primitives invariantes (points d'intérêt: Harris, SIFT....

Cette représentation peut être uniquement exploitée pour la localisation du robot entre plusieurs lieux, ou peut permettre de gérer les déplacements par asservissement visuel entre images successives de la base.

Chaque modèle de carte robotique présenté ci-dessus ne contient qu'un seul type d'information, métrique pour les cartes d'occupation et les cartes d'amers, qualitative pour les cartes topologiques ou les cartes d'apparence. Mais comme nous l'avons mentionné dans les paragraphes précédents, les modèles spatiaux cognitifs vont bien au delà de la simple représentation géométrique et de complexes processus cognitifs permettent l'intégration d'informations perceptuelles avec une connaissance plus qualitative du monde. Pour revenir sur le modèle cognitif de [Tversky 93], il faut dissocier les aspects globaux et locaux:

- les cartes topologiques ou des cartes d'apparence éparses peuvent décrire des environnements de grande dimension; elle peuvent être exploitées pour générer des sous-buts lors des grands déplacements (séquence de lieux à traverser);
- les cartes métriques, vu leur encombrement, ne peuvent être que locales; elles permettent la planification et l'exécution de trajectoires dans des environnements de faible dimension, éventuellement encombrés (traversée d'une pièce, franchissement d'une porte...).

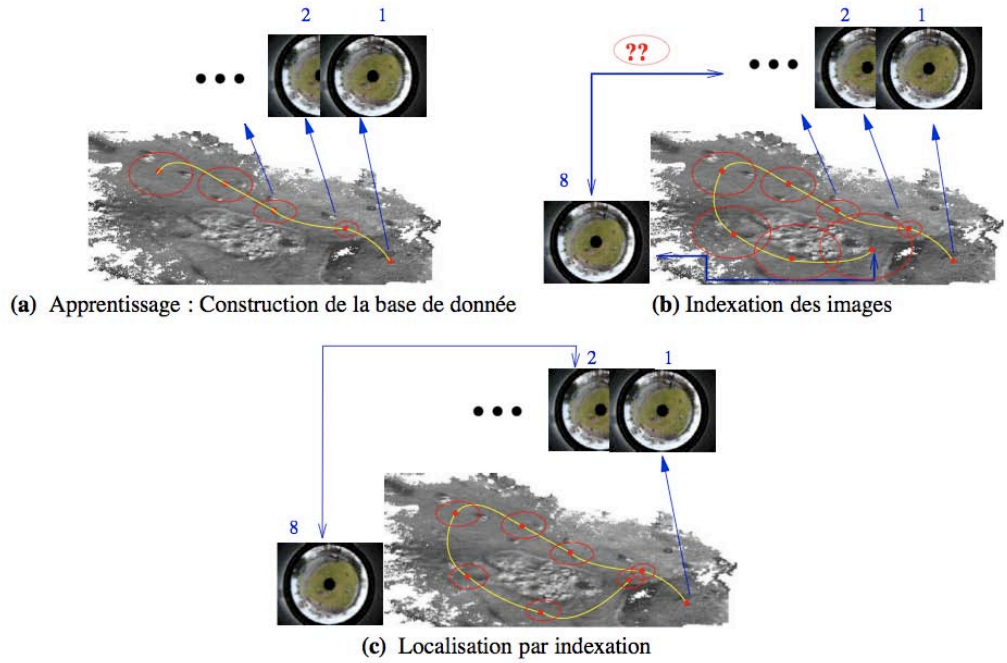

FIG. 2.7 – *Carte d'apparence par indexation d'images panoramiques: exemple de localisation*

Cette nécessité de combiner les représentations a généré beaucoup de travaux. Ainsi, [Thrun 98] proposa un modèle hybride métrique-topologique. Dans les "espaces conceptuels" de [Gärdenfors 00], les informations cognitives et les primitives géométriques associées sont intégrées et traitées dans un même espace de grande dimensionnalité. La carte ainsi créée gagne en homogénéité, au dépend d'une complexité accrue des fonctions de comparaison devant traiter informations géométriques et informations qualitatives avec la même métrique. Enfin, l'un des concepts de représentations de l'espace en robotique mobile les plus complets est la Hiérarchie de sémantique spatiale ou SSH[3] proposée par [Kuipers 77, Kuipers 00]. Ici, les informations qualitatives et métriques sont organisées dans deux hiérarchies complémentaires suivant des couches capteurs, contrôle, causalité, topologie et géométrie. Chaque couche répond à des besoins spécifiques liés au raisonnement et/ou à l'intégration perception-action. Jusqu'à aujourd'hui, le modèle SSH n'a effectivement été utilisé que pour des tâches de navigation dans un contexte intérieur.

2.2.4 Approche écologique

L'évolution des sciences cognitives dans le domaine de la perception a donné naissance à un contre-courant appelé approche écologique ou approche de perception directe

3. en anglais: Spatial Semantic Hierarchy

[Gibson 50, Gibson 79, Turvey 81], qui cherche à mettre en évidence la richesse des informations disponibles directement au travers de la perception du monde. En particulier, elle se base sur le postulat que l'organisme (vivant ou artificiel) et les éléments extérieurs sont virtuellement toujours en mouvement. Ainsi, les propriétés particulières de l'environnement seraient directement "accessibles" depuis le flux de perception, sans nécessité de représentions internes complexes du monde.

Un exemple d'application de cette approche est l'estimation du "temps avant contact" d'un objet approchant dans le champ visuel de l'organisme: le temps de contact peut être directement déduit de la taille de l'objet dans l'image perçue et de sa vitesse d'expansion sans pour autant devoir calculer explicitement la distance à l'objet.

Plus généralement, l'approche écologique de Gibson s'attache aux changements de structures de l'impression lumineuse,c'est-à-dire le flux optique, et le rapport avec le mouvement dans la scène (figure 2.8). Cet intérêt à la nature dynamique de la perception de l'approche gibsonnienne marque la différence avec la majorité des approches cognitives en perception qui considèrent plus généralement des situations statiques.

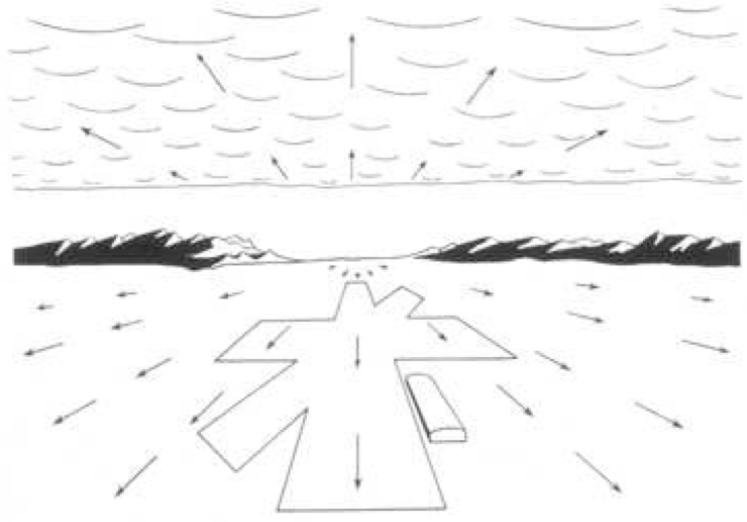

FIG. 2.8 – *Représentation schématique du flux optique d'un avion ou d'un oiseau volant droit*

De plus, l'approche écologique, contrairement aux approches cognitives, prône la perception pour l'action et non la perception pour la construction d'une représentation interne de l'environnement. En cela, les principaux travaux cherchent à exhiber comment des informations pertinentes pour l'action peuvent être extraites du flux de perception, comme par exemple pour l'estimation du "temps avant contact". On parle alors d'affordances: cela reflète l'ensemble des possibilités d'actions et d'interactions de l'organisme vis-à-vis

de son environnement. Par exemple, le sol possède pour un homme les affordances "tenir sur" et "marcher sur", ou pour un robot mobile à roues "tenir sur" et "rouler sur".

2.3 Reconnaissance visuelle d'objets appliquée à la robotique

La reconnaissance visuelle d'objets est un thème important des sciences cognitives. En robotique ou en vision artificielle, cette thématique a fait l'objet de très nombreux travaux; un robot doit reconnaître un objet dans deux situations assez différentes

- pour la navigation, l'objet est perçu dans son environnement, éventuellement très complexe (poster sur un mur texturé...); cela peut être un objet générique (un poster, une chaise, une table...). Un objet peut être un amer de localisation ou peut être une consigne dans une primitive de mouvement de type *Aller-vers-Objet*. La reconnaissance peut n'être que qualitative (localisation qualitative, asservissement visuel 2D) ou peut nécessiter l'estimation de la position relative Robot-Objet (localisation métrique, asservissement 3D);
- pour la manipulation, l'objet est perçu de près, généralement isolé, posé sur un fond uniforme; c'est un objet spécifique, dont un modèle géométrique 3D a été appris; on parle d'instance d'objets génériques (la tasse *Oxford*, la bouteille de *Maury*...). Il doit être localisé en métrique, pour sélectionner une position de prise et exécuter un mouvement de l'effecteur vers cette position.

La question soulevée par les différents travaux sur la reconnaissance d'objets, est de comprendre comment les "objets" sont représentés et reconnus à partir de la seule donnée visuelle: il est nécessaire de prendre en compte la capacité du système cognitif à reconnaître un objet sous différents points de vue et différentes illuminations, sans pour autant perdre la faculté de généralisation à partir de plusieurs exemples d'une même catégorie.

2.3.1 Approches Objet-centrée *vs* Vue-centrée

Depuis le début des années 80, de nombreuses études psychologiques et neurologiques ont exploré ce domaine pour essayer d'extraire un modèle computationnel de l'extraction et de la catégorisation d'objets 3D des informations visuelles par l'homme (et les primates en général), dont la première approche fut proposée par [Marr 78]. De ces études sont apparues et se sont opposées deux approches principales.

Approche Objet-centrée ou structurelle L'approche structurelle repose sur l'idée qu'un ensemble de primitives sont systématiquement extraites des données visuelles et combinées suivant une certaine hiérarchie pour fournir une reconstruction en trois dimensions de l'objet observé. Le postulat de base de cette théorie est que la représentation dépend uniquement de l'objet [Marr 78], rendant ainsi compte de la capacité de reconnaissance suivant n'importe quel point de vue et permettant d'éviter l'explosion combinatoire qu'engendrerait l'utilisation d'un modèle dépendant du point de vue.

L'un des principaux défenseurs de cette approche est Biedermann [Biederman 87] et sa "reconnaissance par composants" ou RBC[4]. Son modèle propose une reconstruction volumétrique des objets à l'aide de composants de base ("geoms") liés par un ensemble de relations spatiales.

Néanmoins, les expérimentations psychologiques et neurologiques illustrent difficilement cette théorie et tendraient plutôt à prouver que l'homme utilise des représentations dépendantes du point de vue.

Approche Vue-centrée L'approche vue-centrée propose de modéliser un objet par un ensemble d'indices locaux situés dans différentes images, ou vues, de cet objet. Ces indices locaux aident au processus de reconnaissance par leur présence dans la vue perçue [Riesenhuber 99]. Cependant, cette approche doit également permettre la généralisation de la reconnaissance aux instances d'une même classe et pas seulement le même objet sous différents points de vue.

[Tarr 98] donne une très bonne vision d'ensemble de ces deux approches. Néanmoins, même si les études psychologiques et neurologiques tendent à aller dans le sens d'une modélisation et reconnaissance par vues, elles sont toujours matière à controverses et elles ne permettent pas d'expliquer seules le processus de reconnaissance visuelle. L'approche structurelle (et pas seulement l'approche RBC) permet facilement la généralisation du modèle et la catégorisation, mais est trop déterministe. L'approche par vue traite, elle, difficilement de la catégorisation, est par nature relativement sensible au point de vue et nécessite des mécanismes de normalisation pour prendre en compte les invariances. Mais des techniques permettent de réduire ces difficultés, et elle est souvent plus adaptée à la reconnaissance d'une instance donnée d'un objet.

4. en anglais: Recognition By Components

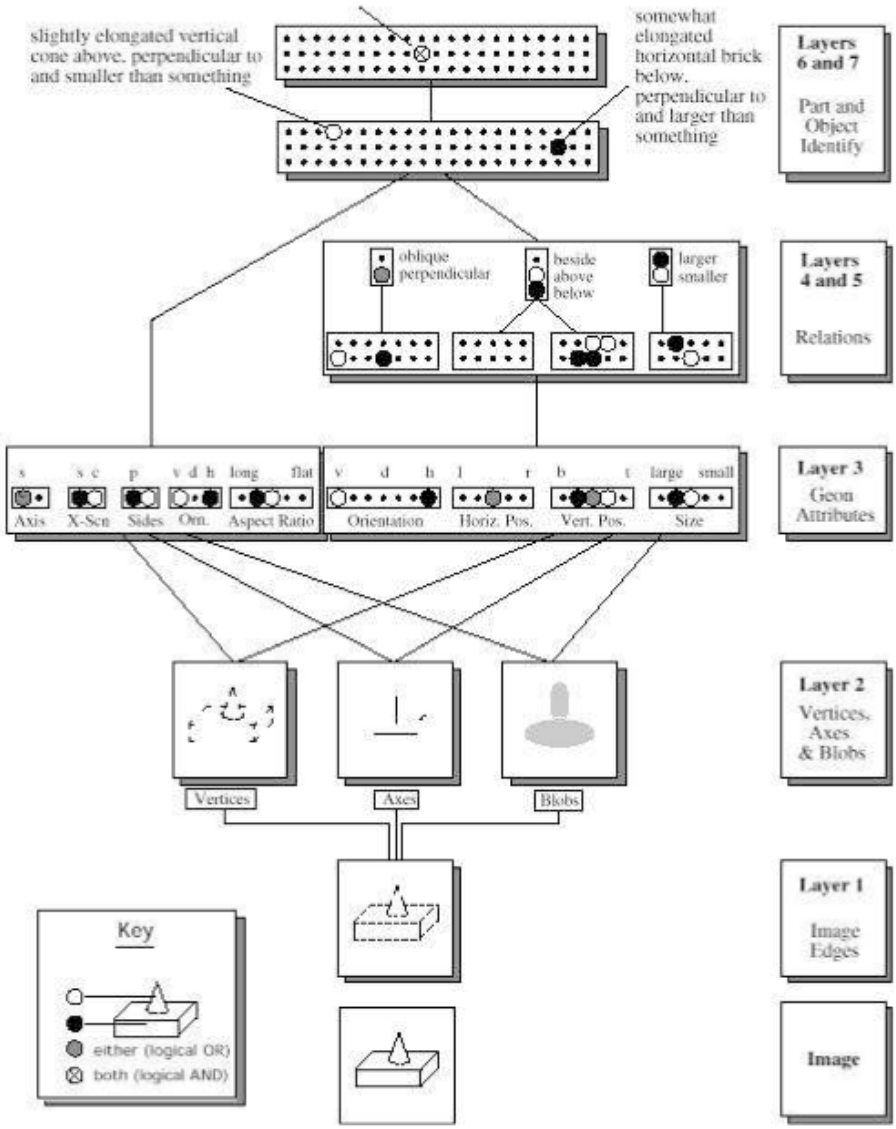

FIG. 2.9 – *Reconnaissance de forme de type RBC*

2.3.2 Modèles computationels

Cette distinction entre approche structurelle et approche par vues se retrouve plus ou moins dans les méthodes de vision par ordinateur.

Nous retrouvons l'approche structurée chez [Chella 97, Chella 01, Edelman 99, Demerci 04], où les objets sont représentés comme un ensemble de primitives géométriques simples associées par des relations spatiales ou topologiques. Dans son expression la plus simple, un objet est composé d'un simple squelette, mais il existe également des modèles de sphères, cylindres, quadriques, *etc.* Cette méthode, très attractive au premier abord dû à sa ri-

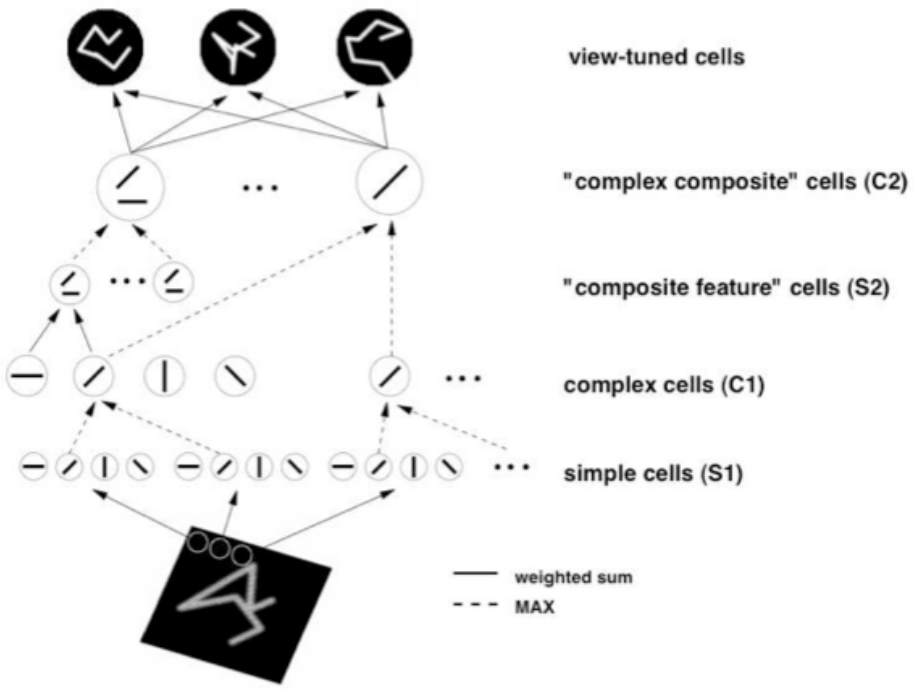

FIG. 2.10 – *Exemple de modèle vue-centrée*

gueur conceptuelle et à son invariance vis-à-vis du point de vue, souffre néanmoins de l'abscence de méthodes robustes d'extraction des structures à partir des images brutes. De plus, certains objets, comme une chaussure, sont difficiles à modéliser avec un graphe de primitives ([Edelman 97] résume assez bien l'ensemble des difficultés d'implementation liées à cette approche).

L'approche par vues a inspiré la plupart des méthodes de reconnaissance d'objets visuelles. L'une des plus connues est la méthode d'indexation d'images, qui consiste à mémoriser un objet comme un ensemble discret de ses points de vues. Diverses méthodes d'analyses comme l'Analyse en Composantes Principales (ACP) [Leonardis 02] permet de réduire la dimensionnalité de l'espace d'analyse des vues. Une autre approche vue-centrée fait appel à des contraintes géométriques [Pope 00]: ici, un objet est représenté par un petit ensemble d'indices visuels et leurs positions géométriques relatives. A l'aide de méthodes d'extraction robuste aux changements de points de vues ou l'illumination, cette méthode s'est révélée très performante [Mikolajczyk 02, Lowe 04].

Un certain nombre de méthodes issues de la mouvance vue-centrée reposent sur l'utilisation d'espaces de caractéristiques. Ici, un objet est représenté par un vecteur de valeurs caractéristiques. Une caractéristique représente un type d'information discrète ou conti-

nue comme la teinte, taille, moment d'ordre x, *etc*. Ces caractéristiques forment ainsi un espace de grande dimension sur lequel peuvent être appliquer des outils d'analyses statistiques pour essayer de mettre en évidence une structure particulière de l'espace des caractéristiques représentatives d'un ensemble d'exemples de l'objet considéré. Ces caractéristiques sont en général faciles à détecter et à extraire des images brutes. Cependant, la structure et les propriétés liées à la catégorie d'objet sont implicites à la représentation, contrairement à l'approche structurelle. De ce fait, il est nécessaire d'appliquer un traitement supplémentaire pour extraire des informations de plus haut niveau comme par exemple dans [Schiele 00]. La simplicité d'implémentation de la reconnaissance et sa robustesse dépendent notamment de la structure de la représentation des caractéristiques (histogramme, point d'un espace de dimension n, *etc*).

La représentation par vecteurs de caractéristiques est confrontée à trois problèmes. Tout d'abord, une caractéristique doit être à la fois discriminante et robuste aux transformations et déformations visuelles (Nous voulons pouvoir reconnaître un visage donné quelles que soient les conditions de prises de vues, sans pour autant le confondre avec un autre visage). Ensuite, dans le cas d'un apprentissage dans un espace à grandes dimensions, un certain nombre d'exemples sont nécessaires pour pouvoir extraire une région de l'espace représentative de l'objet. Plus la dimensionnalité de l'espace est grande, plus le nombre d'exemples nécessaires pour l'apprentissage est important, et certains objets ne permettent même pas d'extraire cette région par analyse statistique ou réduction de la dimensionnalité. Enfin, ces caractéristiques ne contiennent en général pas d'informations de haut niveau de type abstrait ou sémantique, nécessitant la conception de modèles hybrides à l'instar des représentations de l'environnement vues dans la section 2.2.

2.3.3 Des contraintes supplémentaires

Les études récentes en psychophysique ont permis d'exhiber de nouvelles contraintes dans les modèles cognitifs de reconnaissances d'objets, principalement la contrainte temps et la contrainte attentionnelle.

Thorpe et al. [Thorpe 96] mesurèrent le temps mis par le système visuel humain pour analyser une image naturelle complexe. Dans le cas d'une catégorisation binaire (par exemple, l'image contient ou non un animal), les expériences ont montré que l'analyse dure 150 ms, même avec une stimulation visuelle de seulement 20 ms. Ces expériences tendent à prouver que le système visuel humain ne présente pas de boucle de retour d'information. Thorpe décrit dans [Thorpe 98] un modèle de réseaux de neurones reflétant ce processus linéaire, qui sera implémenté dans le cadre de reconnaissance de visages, traitement du flux optique, *etc*Ce sont les travaux de la société *Spikenet*, qui ont inspiré aussi les travaux

de doctorat de W.Paquier et N.Dohuu [Huu 05, Huu 06] dans le pôle Robotique et IA du LAAS-CNRS sur l'apprentissage visuel distribué d'objets.

La seconde contrainte est liée aux études sur la vision active et le phénomène d'attention. Rensink et al. [Rensink 97, Rensink 00] ont mis en évidence que le système visuel humain peut ne pas détecter un changement dans la scène si celui-ci correspond à une saccade ou un clignement des yeux. Cela tend à montrer que le cerveau ne traite pas l'information visuelle dans sa globalité avec une acuité accrue mais use d'un processus d'attention et de focalisation pour construire une représentation stable de l'objet perçu. L'implémentation d'un système attentionnel permettrait de simplifier et d'accélérer les mécanismes de reconnaissances liés aux modèles cognitifs vue-centrés.

2.3.4 Représentation multimodale des objets

Les différentes implémentations de reconnaissance purement visuelles ont permis de mettre en évidence un défaut majeur de ces approches: l'information visuelle seule ne permet pas d'aller au delà de la détection et de la reconnaissance. Le concept d'objet chez l'homme ne repose qu'en partie sur les informations visuelles: la plupart des catégories reposent sur un ensemble de labels linguistiques ou un ensemble de fonctionnalités.

De plus en robotique, la reconnaissance visuelle n'est qu'un des besoins liés à la modélisation d'objets: un robot cognitif doit pouvoir manipuler les objets, comprendre leur utilisation suivant leurs fonctionnalités et leurs limites, et, dans le cadre d'un robot de service en interaction avec l'homme, pouvoir communiquer sur leur nature.

A terme, le modèle d'objet devra intégrer des informations visuelles, des informations haptiques ainsi que des descriptions sémantiques pour permettant au robot de se rappeler un objet à partir d'une description linguistique ou à l'inverse de générer une telle description à partir des perceptions de l'objet.

2.4 Conclusion

Nous avons dans ce premier chapitre, rappelé les besoins et les contraintes des applications de la robotique en milieu humain, ainsi que les modèles proposés pour la cognition spatiale de l'Homme, et les représentations de l'environnement exploitées par un système robotique qui doit gérer des déplacements dans un environnement inconnu a priori.

Notre travail s'est inscrit dans deux principaux projets. Nous avons d'abord participé au développement du robot guide de musée RACKHAM, ce qui nous a permis d'appréhender les représentations nécessaires sur un robot pour planifier et exécuter des trajectoires

dans un environnement intérieur spécifique, partagé avec de nombreuses personnes. Puis nous avons été impliqué dans le projet COGNIRON, dédié à l'étude du robot personnel: en ce cas le Robot intervient dans un environnement intérieur classique, et interagit avec un seul Homme. Nous allons décrire ces deux projets, puis proposer notre représentation de l'environnement et notre approche pour la construire de manière autonome.

Chapitre 3

Vers une exploration qualitative

Les travaux présentés dans ce manuscrit s'inscrivent dans deux projets, mais plus particulièrement, dans le projet COGNIRON [1] - Le robot cognitif compagnon -: il s'agit d'un projet européen du programme FP6-IST visant aux "développements de robots cognitifs dont la fonction est de servir l'homme en tant qu'assistant ou "compagnon". De tels robots doivent être capables d'acquérir de nouvelles capacités et de nouvelles connaissances à l'aide d'un apprentissage ouvert et actif et d'évoluer en constante interaction et coopération avec l'homme" [2].

Nous allons donc nous intéresser plus particulièrement aux problèmes de compréhension de l'environnement des robots autonomes en milieu humain. Avant de préciser notre contribution en ce domaine, nous proposons de découvrir les résultats obtenus dans les deux projets auxquels nous avons participé:

- nous avons participé au développement du robot guide de musée RACKHAM; la démonstration à la cité de l'Espace de Toulouse a été mise en place par le pôle robotique du LAAS-CNRS, au printemps 2004, et est restée active en présence du public jusqu'à fin 2005.
- nous avons ensuite été impliqué dans le projet COGNIRON, de 2004 à aujourd'hui.

3.1 Rackham: un exemple de robot de service

Dans le cadre d'une coopération entre le LAAS-CNRS et la "Cité de l'Espace", un musée scientifique de Toulouse, un démonstrateur interactif jouant le rôle de guide a été développé [Clodic 06]. D'autres robots guides ont également déjà fait parler d'eux : citons

1. http://www.cogniron.org
2. extrait du programme de recherche de la Commission Européenne "Beyond Robotics" ("Au délà de la robotique")

les projets Rhino en Allemagne [Burgard 99] et Minerva aux Etats-Unis [Thrun 00].

Une présentation technique du robot Rackham est donnée dans l'annexe A.

3.1.1 Contexte expérimental

Mission Biospace était une exposition temporaire de la "Cité de l'Espace" à Toulouse proposant d'immerger le visiteur dans ce que pourrait être, au regard des technologies d'aujourd'hui et des sciences de demain, le premier vaisseau de colonisation spatial, au travers de quatorze animations interactives. L'exposition simule, tant d'un point de vue visuel que sonore, l'intérieur d'un vaisseau de 25x10 mètres dénommé "Tsiolkovski".

Cet environnement n'a pas été conçu dans le but d'accueillir un robot mobile et se montrait a priori fort contraignant pour la mise en place de la démonstration:

- Une bande sonore ambiante rendant difficile la reconnaissance vocale et la compréhension de la synthèse vocale.
- Une pièce sombre avec des changements de couleurs sur les parois du vaisseau, mais avec des bandes verticales de couleur bleue pour séparer les stands dans les animations.
- Des parois du vaisseau souples (tissus) et arrondies difficiles à modéliser par une approche classique à base de segments 2D extraits depuis une coupe horizontale acquise par télémétrie-laser.
- Des obstacles proéminents au niveau du sol et à hauteur du casque du robot, invisibles aux capteurs de proximité.
- Des obstacles transparents non détectés par le télémètre laser.
- Des passages étroits nécessitant un positionnement de précision pour avoir une navigation sûre.
- Une foule parfois très dense, non habituée à la présence d'un robot mobile.

Néanmoins cette exposition *Mission Biospace* s'est avérée être une parfaite mise en situation réelle pour notre robot de service, c'est-à-dire un environnement dynamique, pas forcément adapté a priori aux capteurs du robot, avec une forte présence humaine. Il faut noter que cette collaboration fut une aubaine pour une équipe de robotique comme la notre, ce genre de conditions "réelles" étant difficiles à mettre en oeuvre dans un laboratoire.

FIG. 3.1 – *Rackham en promenade dans le vaisseau "Tsiolkovski"*

3.1.2 Modalités de localisation et de navigation

Pour évoluer dans le vaisseau, Rachkam utilise une carte 2D de segments construite a priori, comme présenté sur la figure 3.3. Un opérateur, tuteur du Robot, promène grâce au *joystick*, le robot dans l'environnement: les segments détectés depuis les données sensorielles acquises par le télémètre-laser sont intégrés à une carte stochastique 2D au fur et à mesure des déplacements, par une approche SLAM classique fondée sur le filtrage de Kalman. La localisation est assurée par le module SEGLOC par mise en correspondance à fréquence élevée des segments actuellement perçus par le laser et ceux de la carte 2D. Le laser, placé à l'avant du robot, avec un champ perçu de 180°, fournit suffisamment d'informations pour la localisation mais pas pour la navigation réactive qui nécessiterait un champ de vision complet à 360°. Un deuxième module, ASPECT, donne à chaque instant une estimée locale des obstacles entourant le robot grâce aux segments courant et à ceux précédemment perçus. Il permet également, grâce à SEGLOC, de déterminer la présence d'objets inconnus ou dynamiques présents dans le champ de vision du laser.

Enfin, pour pallier à certains problèmes dûs à l'utilisation d'un laser 2D et pour aider la supervision pour l'interaction, un ensemble de zones ont été manuellement définies sur la carte, comme indiquées sur la figure 3.3. Il existe trois types de zones: *obstacle*, dans lesquelles le robot de peut naviguer, *cible*, indiquant la proximité d'une des quatorze expériences de l'exposition et *spécial*, pour les entrées, sorties, intersections potentiellement encombrées par le public. Le module ZONES, se basant sur les données de localisation,

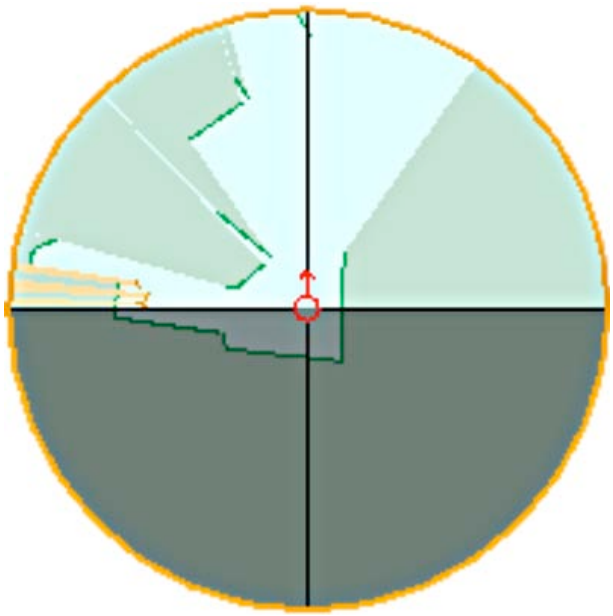

FIG. 3.2 – *Carte d'aspect. les segments verts représentent les segments reconnus par le module SEGLOC. les segments oranges représentent les objets inconnus et éventuellement dynamiques*

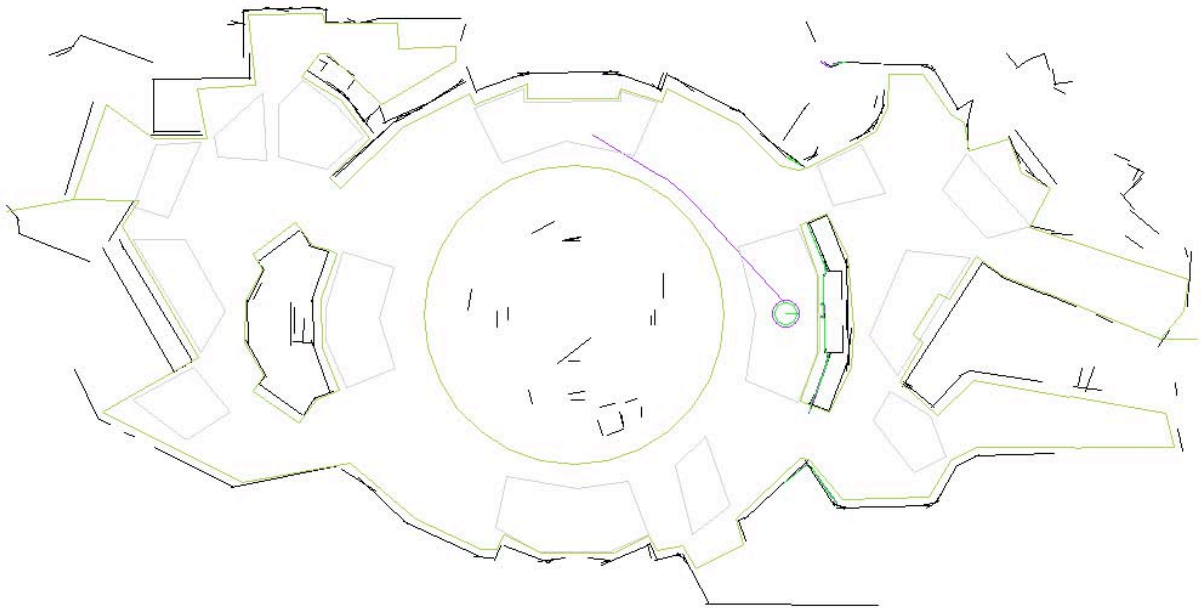

FIG. 3.3 – *Carte de segments 2D de l'exposition et les différentes zones:* cible *et* spécial *en gris,* obstacle *en vert*

surveille et enregistre les entrées et sorties du robot de chaque zone.

L'interface avec les utilisateurs se fait à travers un écran tactile, illustré en figure 3.4:

le robot détecte et suit l'utilisateur qui ouvre une session sur le robot via une caméra PTZ; l'image acquise par cette caméra est montrée en haut à droite. Le robot exploite un module de synthèse de la parole pour inviter l'utilisateur à choisir une destination sur la carte sémantique affichée à gauche de l'écan; le clone en bas à droite, intégré des travaux du laboratoire ICP de Grenoble, donne une apparence humaine pour cet IHM.

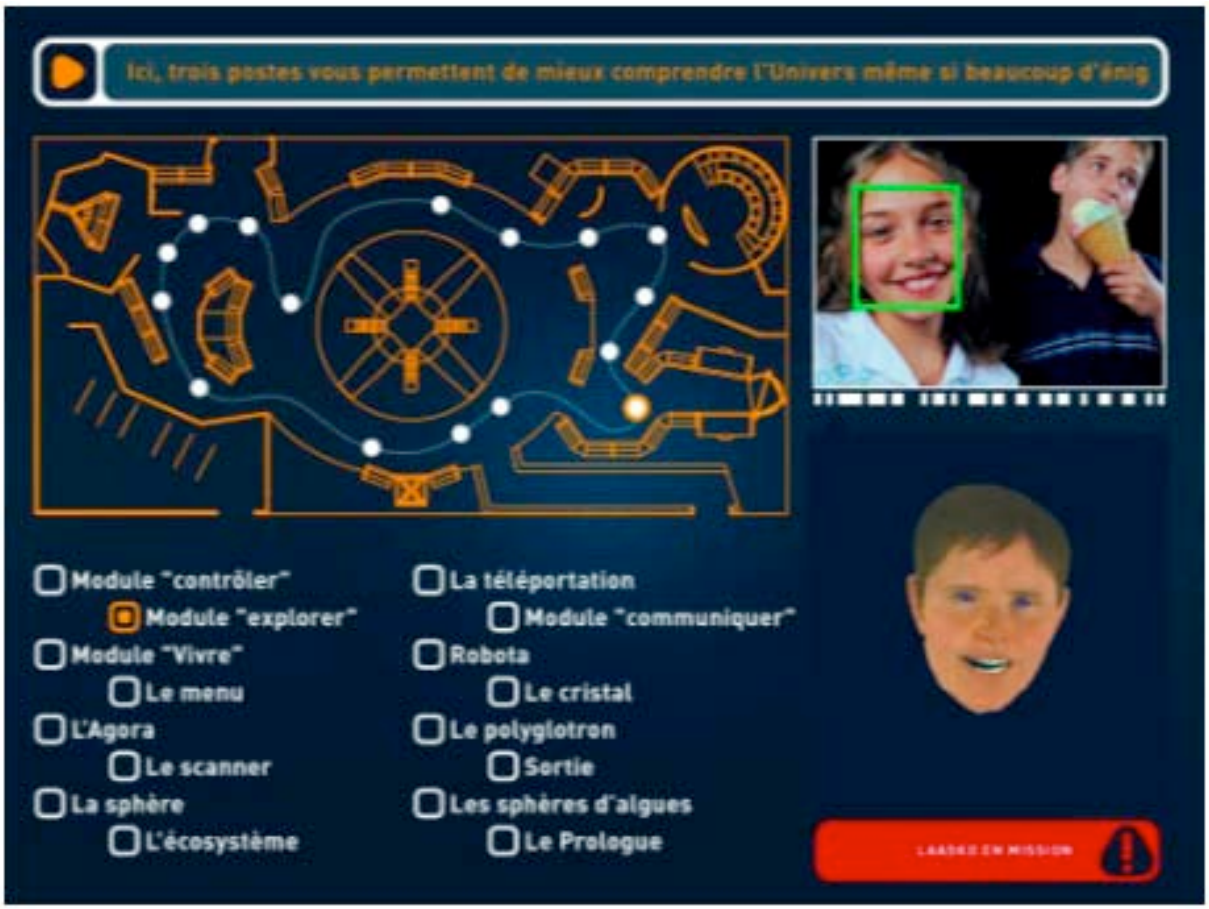

FIG. 3.4 – *Ecran exploité pour l'interface Homme-Robot, avec affichage de la carte sémantique de l'environnement, contenant les noeuds du modèle topologique.*

La figure 3.5 montre en haut une image acquise depuis le robot dans le vaisseau: la couleur verte sur les murs change graduellement pour devenir rouge, puis bleue. Par contre la couleur des néons verticaux, et des panneaux présentant des informations aux visiteurs, reste stable: des travaux préliminaires ont donc été menés pour exploiter ces bandes et ces panneaux comme des amers visuels; la figure 3.5 montre en bas une image des segments associés par une technique de relaxation, aux bords des bandes et des panneaux. La couleur des régions horizontales, et le nombre des bandes bleues verticales permettent de les associer sans ambiguité aux panneaux existants dans le vaisseau: une approche SLAM

aurait du être développée pour ajouter ces amers visuels dans la carte, mais ce travail n'a pas été achevé car de fait, malgré l'encombrement du site par le public et les occultations qui en résultent, la localisation sur les segments-laser s'est révélée suffisamment robuste.

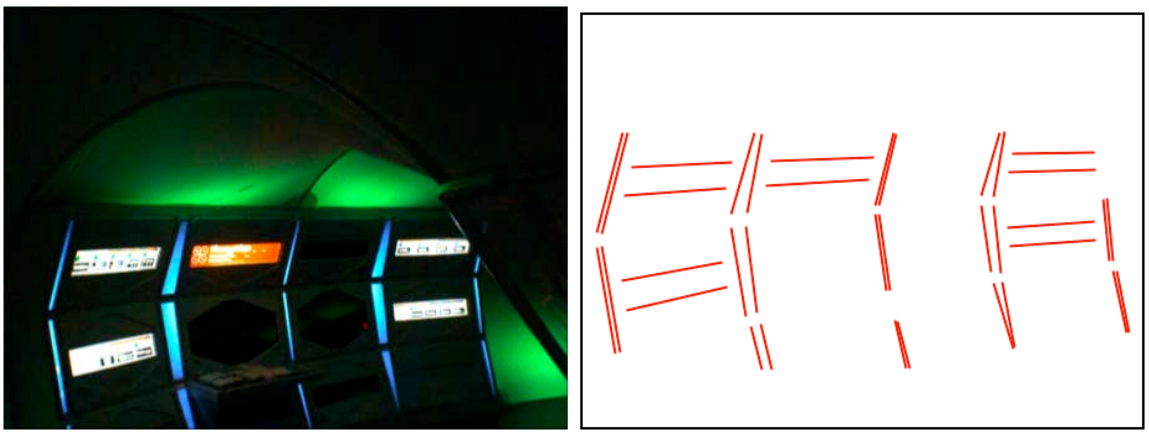

FIG. 3.5 – *Fonctions visuelles pour l'extraction d'amers visuels spécifiques à cet environnement.*

3.1.3 Discussion

Commençons par des résultats généraux sur l'expérimentation. Entre mars 2004 et novembre 2006, Rackham a passé près de 100 jours à la "Cité de l'Espace", répartis en 9 séjours (les périodes de retour au laboratoire ont permis d'apporter de nombreuses améliorations sur sa sûreté de fonctionnement, sa robustesse et sa mise en oeuvre). Cela représente 200 heures effectives en "mission", suite aux sollicitations d'environ 3000 visiteurs (Rackham était également capable de proposer "spontanément" ses services après un certain temps d'inactivité) et 66 kms. lors des derniers séjours, environ 2% seulement des missions furent interrompues suite à une détection d'erreur par le superviseur. A noter également qu'il n'y eut aucune collision avec des visiteurs ou le décor.

Néanmoins, Rackham est très loin d'être un robot "intelligent" capable de s'adapter à n'importe quelle situation ou environnement, et ce en grande partie par le manque de processus d'apprentissage évolués et le manque de flexibilité de sa représentation interne de l'environnement. En effet, la carte finalement utilisée par le robot ne repose que sur un seul type d'attributs, les segments 2D laser, sémantiquement pauvres et dont la structure locale peut être difficile reconnaissable; par exemple, dans un long couloir, ce mode de localisation fonctionne très mal par manque d'une structure locale des segments (il n'existe que deux segments parallèles en ce cas). Notons la thèse de H.Bulata [Bulata 96]

qui proposait de créer des amers à partir des segments-laser, amers définis comme une structure locale de segments connexes (coins, objets disposés contre les murs...) suffisants pour déterminer la situation relative Amer-Robot; ces travaux ont bien montré le faible contenu sémantique des segments-laser.

Les quelques informations sémantiques disponibles ont été tracées à la main par un opérateur externe et leur pertinence dépend uniquement de la qualité de la localisation, qui ici pouvait être fortement dégradée par la foule entourant le robot et donc occultant les segments laser. Ce problème est particulièrement délicat pour les zones obstacles, marquées non navigables car il y existe des obstacles non perceptibles pour le capteur laser, comme la grande bulle transparente au centre du vaisseau (visible sur la figure 3.1) ou de manière plus courante, comme des baies vitrées, ou des marches vers le bas que le capteur laser ne détecte pas; il faudrait pour détecter ces situations, d'autres capteurs (nappe laser pour les marches par exemple).

Enfin, la représentation du site de l'exposition a été figée une fois fini l'apprentissage; elle a été ensuite exploitée pour la navigation (localisation et détection des obstacles). Il n'était pas permis au robot de mettre à jour automatiquement sa carte en cas de déplacement d'une démonstration ou ajout d'une nouvelle aile à l'exposition. Ce choix se justifie en ce cas, car les modifications d'une exposition dans un musée sont exceptionnelles: dans le contexte du robot personnel, l'environnement est modifié sans cesse par l'Homme habitant l'environnement, et le découplage entre les phases d'apprentissage et d'exploitation du modèle, n'est plus possible: la représentation doit être mise à jour en ligne.

3.2 Apprentissage de l'environnement par un robot personnel

Projetons nous dans N années: l'Homme acquiert son Robot personnel, un robot mobile doté de un ou deux manipulateurs afin de réaliser des tâches complexes, et de nombreux capteurs pour percevoir l'Homme et l'environnement. Pour satisfaire des contraintes de coût, nous supposons ici que le capteur essentiel sur un tel robot, sera la Vision, avec plusieurs modalités possibles: monoculaire ou stéréo, multi-focale, avec éventuellement une combinaison caméra omnidirectionnelle/caméra perspective active (PTZ). Eventuellement la vision stéréo pourra être remplacée avantageusement par une caméra PMD, ou capteur optique à temps de vol (tel que les capteurs commercialisés par Canesta aux Etats-Unis ou CSEM en Suisse).

Ce Robot est doté d'un grand nombre de fonctions lui permettant de se déplaçer, de reconnaître des objets, de les saisir..., mais aussi de capacités d'apprentissage et d'auto-

adaptation, lui permettant une fois activées, d'une part de construire une représentation spatiale de l'environnement de l'Homme et d'autre part, de sélectionner un comportement le plus adapté aux caractéristiques de l'Homme.

Le projet COGNIRON, *The Cognitive Robot Companion*, est un projet intégré FP6, coordonné par le pôle Robotique et IA du LAAS-CNRS, commencé en Janvier 2004 et devant s'achever en Février 2008. En sus du LAAS, sept autres partenaires académiques sont impliqués. Il s'agit donc d'un projet de recherche, sélectionné sur un appel à projets de type FET *Future and Emergent Technologies*, intitulé *Beyond Robotics*. Son objectif est d'étudier les thématiques liées au développement d'un Robot personnel, compagnon de l'Homme. Plusieurs enjeux ont été identifiées: le dialogue Homme-Robot multi-modal, la détection et l'interprétation des activités de l'Homme, le comportement social du Robot, l'apprentissage de tâches, la cognition spatiale pour la reconnaissance de situations et le système décisionnel auto-adaptatif. Afin d'évaluer les concepts proposés dans ses thématiques, les partenaires participent à trois expérimentations:

- la démonstration *Robot Home tour*, est dédiée à la navigation, de l'apprentissage de l'environnement par le Robot guidé par l'Homme jusqu'à l'exécution de déplacements dans cet environnement.
- la démonstration *Curious Robot*, est dédiée à l'interaction Homme-Robot à travers des objets: de l'apprentissage des objets jusqu'à la saisie de ces objets pour les donner à l'Homme.
- enfin la démonstration *Learning Skills and Tasks* concerne l'apprentissage de tâches par imitation.

La figure 3.6 présente les deux principaux démonstrateurs du projet: le robot Biron, intégré à l'Université de Bielefeld (Allemagne) est le support pour la démonstration *Home Tour*: il apprend la carte de l'environnement par intéraction avec l'Homme. Le robot Jido, intégré au LAAS-CNRS, est le support pour la démonstration *Curious Robot*: il peut modéliser, reconnaître et saisir des objets pour les donner à l'Homme.

Nous avons participé aux travaux sur la cognition spatiale; ce lot de travail, coordonné par B.Krose de UVA (Université d'Amsterdam,Pays-Bas), a mobilisé aussi des collègues de KTH (Université de Stockholm, Suède), de UKA (Université de Karlsruhe, Allemagne), de l'IPA (Stuttgart, Allemagne) et de EPFL (Lausanne, Suisse):

- IPA et UKA avec d'autres collègues du LAAS-CNRS, ont traité de la modélisation des objets: construction d'un modèle hybride, fondé sur l'apparence et la géométrie, exploité pour reconnaître, localiser et saisir des objets. IPA [Kubacki 05a,

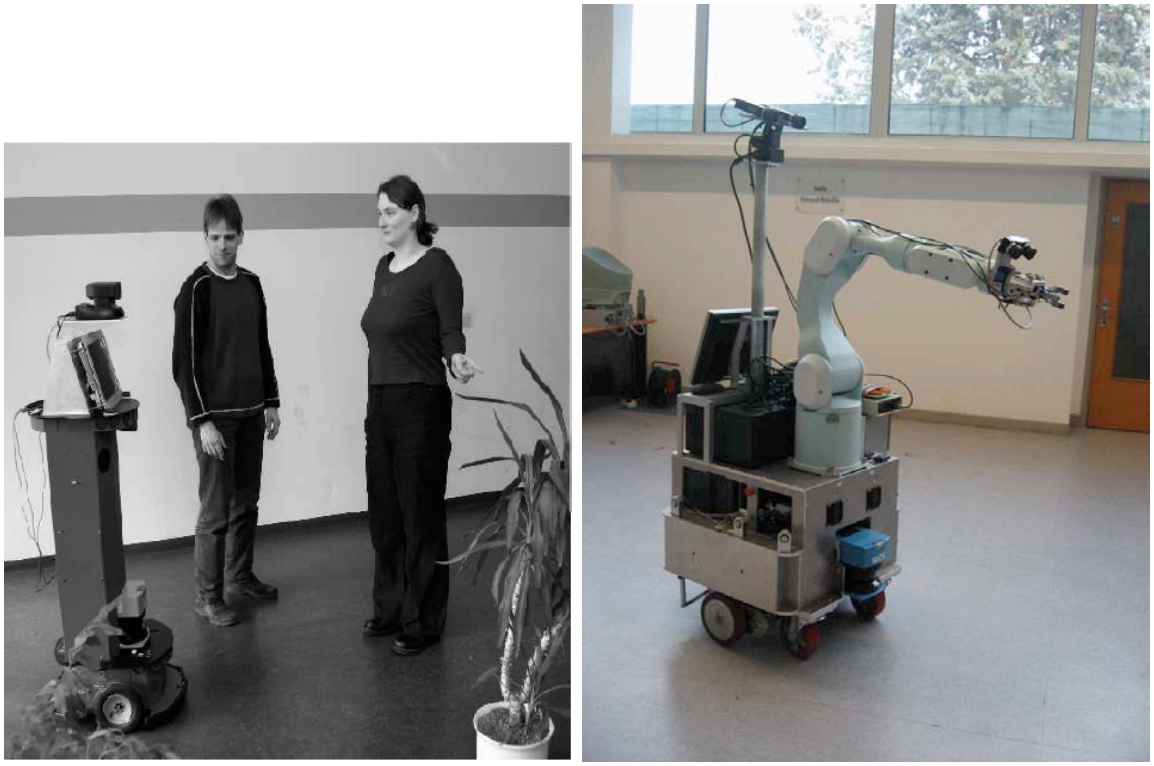

FIG. 3.6 – *Les deux principaux démonstrateurs du projet COGNIRON: Biron à gauche, Jido à droite.*

Kubacki 05b] a développé des méthodes de modélisation par l'apparence et de reconnaissance d'objets, proches de notre contribution décrite en chapitre 5.

– UVA, KTH, EPFL et nous-même au LAAS-CNRS, nous avons ensemble défini un modèle de l'environnement comme une hiérarchie de représentations spatiales. Puis chaque partenaire a proposé des approches spécifiques pour construire et exploiter ce modèle, à un ou plusieurs niveaux.

Nous décrivons ci-dessous la représentation proposée collectivement, et les méthodes d'apprentissage associées.

3.2.1 Représentation hiérarchique

Comme nous l'avions déjà évoqué dans la section 2.2.3, il est nécessaire de développer une carte multimodale pour répondre aux besoins et contraintes liés à la compréhension de l'environnement et aux fonctions de navigation: localisation, sélection d'une stratégie pour aller à un lieu donné (choix d'une route, définie comme une séquence de lieux à traverser), planification de trajectoires dans l'espace libre de chaque lieu...

Nous parlons alors de cartes hybrides respectant une certaine hiérarchie, fondée sur

le degré d'abstraction des représentations. Cette carte contient l'ensemble des représentations spatiales nécessaires pour la navigation d'un robot mobile dans un environnement humain tel qu'un appartement. La configuration la plus utilisée de cartes hybrides est la configuration métrique/topologique, fondée sur des grilles d'occupations [Thrun 98], des segments 2D [Gasos 99] ou des modèles d'apparence [Zivkovic 05]. La carte topologique est soit extraite en fonction de configurations caractéristiques de la carte métrique (par génération d'un graphe de Voronoï par exemple) soit comme liens entre différentes cartes métriques locales.

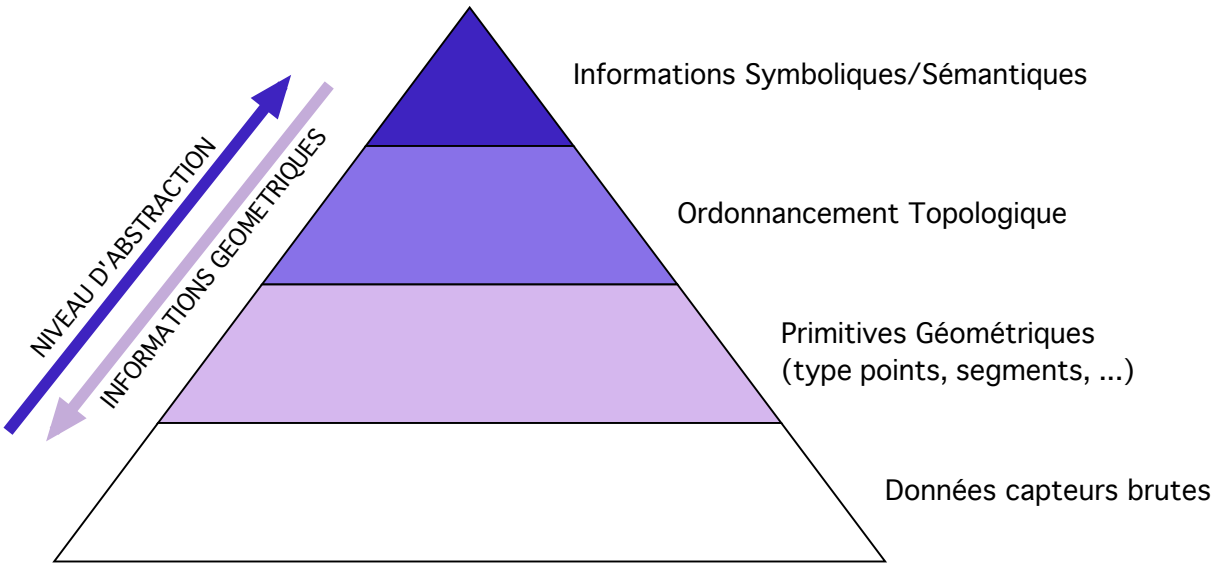

FIG. 3.7 – *Hiérarchie de représentations spatiales*

Le projet COGNIRON propose aussi une carte hybride, sous la forme d'une représentation hiérarchique (figure 3.7), qui part des données capteurs pour atteindre le niveau sémantique, avec interaction de l'Homme pour donner un nom symbolique aux entités du modèle. Plusieurs instances d'une telle représentation hiérarchique ont été développées dans COGNIRON:

- UVA d'Amsterdam et EPFL ont chacun, mais d'une manière différente, exploité une base d'images omnidirectionnelles acquises en des positions aléatoires dans l'environnement, pour apprendre un modèle topologique, sous la forme d'un graphe d'images. Une fois appris ce modèle, la navigation est traitée par asservissement visuel pour parcourir une trajectoire définie dans cette base d'images (approche similaire à celle proposée par A.Remazeilles [Remazeilles 05] à l'IRISA).
- KTH à Stockholm exploite des coupes-laser, afin de construire une carte stochastique de segments-laser, décomposée en sous-cartes pour faire émerger le concept

de lieux. Un tel modèle permet d'activer une navigation classique: génération d'une trajectoire dans l'espace libre, puis exécution en se localisant sur les segments-laser de la carte.

– Enfin EPFL et nous-même au LAAS, avons proposé une approche partant d'une base d'images acquises dans l'environnement. Des objets sont reconnus dans chaque image, ce qui permet la construction d'un graphe d'objets, décrivant l'agencement des objets reconnus depuis les différentes positions du robot. EPFL considère un objet spécifique, appelé *Porte*; les instances des portes rajoutées dans le Graphe d'Objets, servent à découper le graphe d'objets en lieux topologiques.

Dans chaque cas, le niveau sémantique est obtenu grâce au nommage des lieux, et pour la troisième approche, aux objets que le robot sait reconnaître.

Notre approche est de type objet. Cette représentation hiérarchique est bien adaptée aux environnements intérieurs partagés avec l'Homme; elle est plus proche du raisonnement humain. Comme montré en figure 3.8, elle peut être généralisée avec d'autres données sensorielles que de simples images; un lieu peut être défini par une combinaison de simples primitives (segments-laser, segments 3D reconstruits par stéréo ou à partir de N images acquises par la caméra montée sur le Robot...) et d'objets.

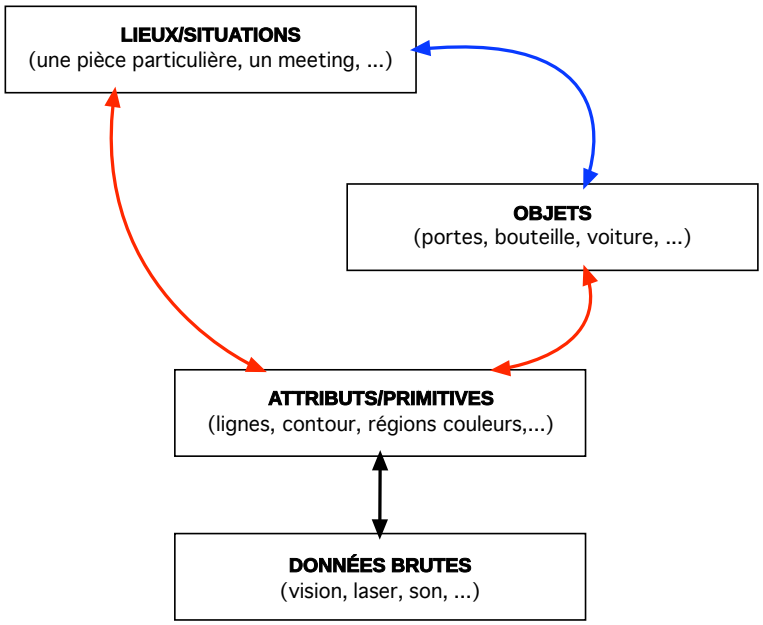

FIG. 3.8 – *hiérarchie en terme d'objets et lieux*

3.2.2 Méthodes d'apprentissage des représentations

Construire des représentations adaptées ne suffit pas à rendre le robot complètement autonome pour exécuter des déplacements à l'aide de sa perception de l'environnement. Il doit également être capable d'adapter, d'améliorer et d'étendre sa représentation interne.

Dans le cadre de la navigation métrique, ce problème est connu sous l'appellation SLAM, pour *Simultaneous Localization And Mapping* ou Localisation et cartographie simultanée. Au départ les méthodes SLAM servent essentiellement à construire des représentations métriques, sous la forme de cartes stochastiques de primitives ou d'objets; mais il est également possible de faire du SLAM topologique et hybride; par exemple, [Choset 01] génére un Graphe de Voronoï Généralisé au cours du déplacement du robot; l'un des intérêts du GVG est qu'il peut intègrer également des informations métriques. Dans COGNIRON, [Topp 06] de KTH, construit une carte de segments-laser, le Robot étant guidé par l'Homme dans l'appartement.

En dehors des approches SLAM, d'autres méthodes de construction automatique de cartes ont été développées, notamment fondées sur l'apparence. Dans COGNIRON, [Booij 06] de l'Université d'Amsterdam exploite une base d'images omnidirectionnelles non ordonnées; de chaque image sont extraits des points d'intérêt; pour chaque paire d'images de la base, des appariements entre points sont recherchés; un graphe est construit, deux images étant connectées par un arc, seulement s'il existe suffisament d'appariements entre les points qui y ont été extraits; un modèle topologique est généré en segmentant ce graphe en lieux par une méthode de type *Graph Cut*; après quelques étapes de filtrage, ces lieux correspondent aux pièces d'un appartement. Egalement dans COGNIRON, [Tapus 05] de EPFL, utilise un système d'empreintes visuelles pour caractériser les noeuds de la représentation topologiques et utilise le formalisme des POMDP (*Partially Observable Markov Decision Process*) pour effectuer les mises-à-jours.

Il existe par contre peu de travaux sur les explorations ou construction de cartes basés sur la notion d'objets.

– [Galindo 05, Vasudevan 07] représentent un environnement de type appartement, par un graphe comportant des arcs objets-objets, objets-lieux et lieux-objets. Dans COGNIRON, [Vasudevan 07] de EPFL, utilise un détecteur de portes pour détecter les changements de scènes (ou pièces). Dans chaque pièce, il cherche à reconnaître des objets préalablement appris pour construire un graphe d'agencements de ces objets, caractéristiques de la pièce visitée. la figure 3.9 montre des résultats obtenus à EPFL par S.Vasudevan: une carte des objets détectés dans un environnement fait de plusieurs pièces, et une partie du graphe d'agencement d'objets (pièce en bas à gauche).

– dans notre groupe au LAAS, [Hayet 03] avait proposé une approche similaire, mais il considérait uniquement une seule classe d'objets, reconnus et localisés en vision monoculaire: des quadrangles plans de type posters. Le robot, durant sa phase d'exploration recherchait au fur et à mesure des déplacements des objets quadrangulaires de ce type: il les exploitait soit en milieu ouvert pour construire par une méthode SLAM classique, une carte stochastique d'objets, soit dans un réseau de couloirs, pour annoter un graphe topologique de type GVG, construit par ailleurs à partir de segments-laser [Hayet 02].

Ces deux approches présentent néanmoins des lacunes: la première nécessite la construction a priori d'une base de données des objets présents dans les différents lieux. La seconde a la possibilité de découvrir les objets pendant l'exploration mais est restreinte aux objets de type posters.

L'approche que nous présentons dans ce document peut être vue comme une généralisation des travaux préalablement effectués dans notre groupe par [Hayet 03] et un complément aux travaux de [Vasudevan 07] sur les graphes d'agencements d'objets:

– S.Vasudevan fournit au robot, les modèles des instances d'objets (pas de classes d'objets) que le robot peut détecter lors de l'exploration. **Nous construisons en ligne les modèles des objets détectés par le robot** ·

– J.B.Hayet ne détectait que les objets appartenant à une classe pré-définie: les quadrangles planaires. Les objets détectés étaient appris en ligne, puis reconnus plus tard en phase de navigation. **Nous souhaitons apprendre tous les objets détectés dans les régions saillantes des images**.

Notre approche peut être vue aussi comme une variante de l'approche proposée par UVA:

– O.Booj décrit un lieu par un ensemble d'images panoramiques acquises en ce lieu, chacune décrite par un ensemble de points d'intérêt. Il reconnaîtra ce lieu par la recherche d'appariements entre points d'intérêt extraits de l'image courante et des images apprises en ce lieu.

– Nous apprenons un lieu par les structures locales saillantes (ou objets) détectées en ce lieu, chacune décrite par un ensemble d'images, pour lesquelles on mémorise les points d'intérêt extraits sur l'objet. Nous reconnaîtrons ce lieu en appariant un ou plusieurs objets de l'image courante, avec ceux extraits en ce lieu, la reconnaissance étant fondée sur les appariement entre points d'intérêt.

Dans les deux sections suivantes, nous précisons les caractéristiques de notre approche.

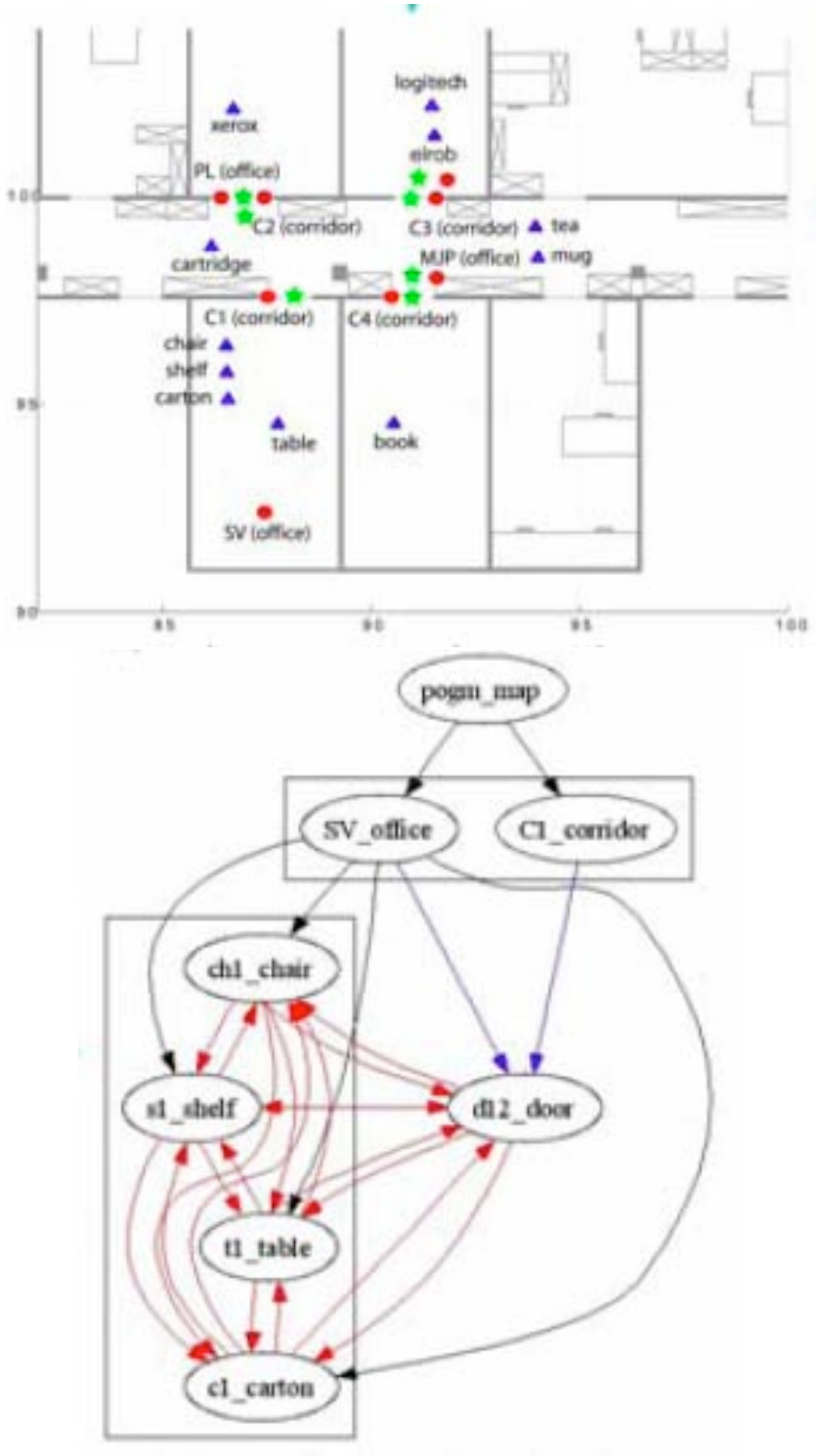

FIG. 3.9 – *Exemple d'un Graphe d'agencement d'Objets (travaux de S.Vasudevan à EPFL)*

3.3 Attention visuelle pour l'exploration autonome

Pour permettre à notre robot de faire de l'exploration visuelle autonome, nous devons lui donner les moyens de sélectionner des zones particulières dites *régions saillantes* dans son champ visuel. Pour cela, nous nous sommes intéressé au phénomène de l'attention en vision.

Traditionnellement, l'attention visuelle est décrite comme un processus automatique qui sélectionne des zones continues du champ visuel. Par exemple, la *théorie du filtrage* [Broadbent 58] supposait que le processus de sélection s'effectuait par filtrage des informations en fonction des capacités de perception et de traitements. La *théorie du point lumineux* [Posner 80] prétend que l'attention est comme un point lumineux qui met en évidence la zone à traiter en se déplaçant d'un point à un autre suivant le mode opératoire *décrochage-déplacement-accrochage*. La *theorie de la loupe* [Eriksen 86] propose que l'attention se porte non plus sur un point, mais sur une région de taille variable.

Les travaux sur l'attention visuelle sont nombreux: ils sont très inspirés par les connaissances disponibles sur la Vision Humaine. Les travaux les plus connus dans la communauté Vision Artificielle sont ceux de L.Itti [Itti 98, Itti 01]. Nous résumons ci-après, les principales fonctions spécifiques aux mécanismes de l'attention visuelle, puis discutons de l'intérêt de ces mécanismes.

3.3.1 Mécanismes de l'attention visuelle

Principales fonctions de l'attention [Tsotsos 95]:

- Sélection d'une région d'intérêt dans le champ visuel: l'image est analysée selon plusieurs échelles et plusieurs attributs, pour construire une carte de saillance (ou *Saliency Map*).
- Sélection de la dimensionnalité d'un attribut et de ses valeurs d'intérêt: pour chaque attribut (couleur, intensité, orientation du gradient...), il faut choisir des échelles pertinentes et un espace de discrétisation.
- Contrôle du flux d'information traité par le système visuel: les cartes de saillance sont construites sur des séquences d'images, acquises pendant le mouvement de l'observateur. Il convient de gérer le flux d'images, de suivre les régions déjà détectées, de détecter celles qui sont apparues avec le nouveau point de vue ...
- Décalage d'une région d'intérêt à la prochaine avec le temps: dans notre approche, un mécanisme de focalisation est appliqué successivement sur chaque région d'intérêt. Il convient alors de gérer en parallèle un processus (dit pré-attentif) de détection

des régions saillantes, et un processus (dit attentif) de focalisation. Un mécanisme d'inhibition permet d'éviter de détecter la région en cours d'analyse par focalisation.

Propriétés de la région d'intérêt

- *les frontières:* il a été mis en évidence qu'il existait une frontière relativement nette entre ce que nous nommerons la région d'intérêt et son voisinage. Selon la "théorie du point lumineux", cette frontière se formerait entre les objets et ne les couperait que rarement. Cependant, la "théorie de la loupe" considère que l'intensité de l'information visuelle est forte au centre de la région d'intérêt et diminue graduellement à l'intérieur d'une frontière de taille et de forme variable.
- *la taille variable de la zone attentionnelle:* dans le modèle à deux processus pré-attentif et attentif de la vision, l'attention peut être régler de façon plus ou moins précise en une distribution uniforme de l'information visuelle. Une approche lâche est appropriée à un traitement en parallèle des objets qui "surgiraient" du processus pré-attentif alors qu'une approche précise conviendrait plus à un traitement séquentiel des objets construits par une conjonction d'attributs. Dans le modèle "loupe", la distribution de l'attention est adaptée par le processus pré-attentif tandis que le processus attentif se charge de détailler l'apparence, éventuellement de reconstruire la région d'intérêt.
- *l'intensité variable de la zone attentionnelle:* la "théorie du point lumineux" considère la région d'intérêt comme une région d'intensité uniforme et déplace l'attention d'une région uniforme à une autre. Le modèle "loupe" assume lui que l'intensité dans la région d'intérêt peut varier mais reste globalement constante.

3.3.2 Intérêts des mécanismes de l'attention

Ces mécanismes attentionnels ont beaucoup été étudiés dans les années 90, avec la vague des travaux sur *Purposive and Active Vision* (Aloimonos, Ballard, Brown, Bajcsy...), entrepris en réaction à la prédominance des approches géométriques de la Vision (Faugeras, Hartley...).

Plusieurs avantages sont classiquement associés à ces approches [LaBerge 95]:

- *de meilleurs jugements perceptuels pour planifier des actions:* l'attention permet d'accroitre la qualité du jugement perceptuel en ne sélectionnant que la partie pertinence de l'information directement à la source du processus cognitif. La sélection hiérarchisée de l'information entrante (par degré décroissant de pertinence) permet également une meilleure planification des tâches à effectuer par le processus cogni-

tif. En robotique, ces tâches peuvent correspondre à des déplacements du Robot: sélection du meilleur point de vue, déplacement du capteur en ce point...

- *rapidité de traitement:* en ne sélectionnant que la partie pertinente de l'information, l'attention réduit et hiérarchise ainsi la quantité d'informations à traiter et accélère le processus cognitif.
- *concentration du processus de pensée:* avec l'attention, une perception ou une action peut être maintenue sur une période étendue.
- *contrôle de l'ordre de traitement de l'information:* grâce à l'attention, les différents détails d'une cible sont sélectionnés et traités en fonction des tâches à accomplir par le système robotique.

Notons également que ce découpage en deux processus perceptuels, pré-attentif pour la détection, attentif pour la focalisation, introduit également deux niveaux de mémorisation: mémorisation *court terme*, avec fonction d'oubli des régions non analysées après un délai court, et mémorisation *long terme* des caractéristiques des régions analysées.

3.4 Une approche active et autonome de type "augmenté"

Nous proposons donc une approche active d'exploration visuelle permettant de dégager de l'environnemment des structures locales - ou proto-objets - susceptibles de porter une information sémantique de type "objet". Comme dans les travaux de J.B.Hayet, ces objets pourraient être exploités pour "augmenter" la carte utilisée pour la localisation, soit une carte métrique construite par une approche SLAM, soit une carte qualitative de type graphe d'agencement des lieux et objets détectés dans l'environnement.

Notre système reprend le modèle dual pré-attentif/attentif, comme résumé dans la figure 3.10: ici, les deux chaines de traitements s'exécutent de manière asynchrones et partagent la même base de données de proto-objets.

Dans le premier processus, des proto-objets sont extraits des images omnidirectionnelles (ou à défaut grand angle) à l'aide d'un modèle attentionnel ascendant. Un module de suivi alimente, au fur et à mesure des déplacements du robot, un module d'ajustement de faisceaux appliqué séparément pour chaque proto-objet suivi: ce module exploite une estimée du déplacement du capteur, estimée obtenue dans nos travaux par intégration des données acquises par des capteurs proprioceptifs (odométrie, gyromètre); de ce fait, sous peine de dérive trop importante, et donc d'erreurs dans l'ajustement, le suivi ne se fait que sur des courtes distances. Dans une approche hiérarchique de construction des

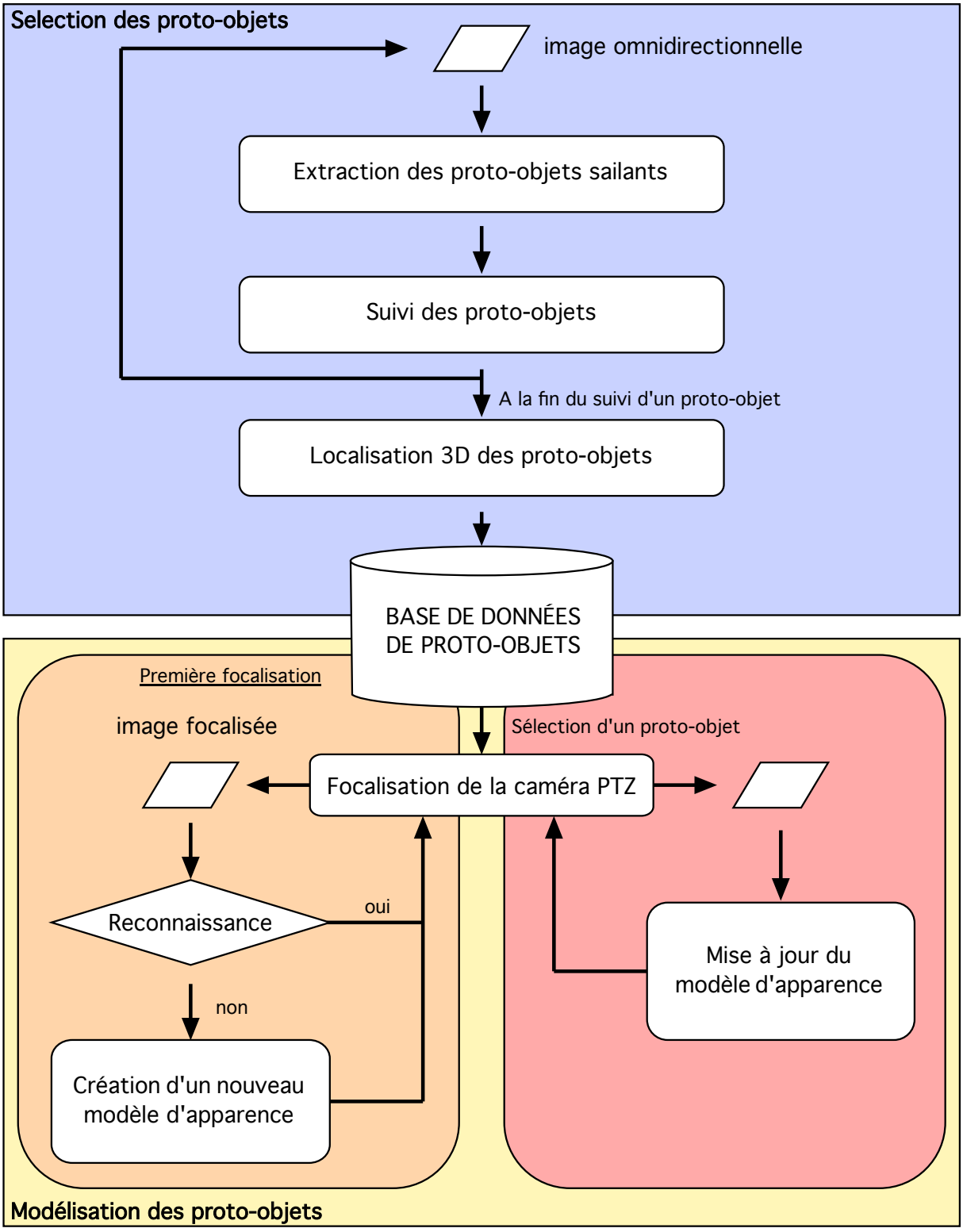

FIG. 3.10 – *Processus pré-attentif et attentif de notre système*

représentations par notre Robot cognitif, ces estimées des déplacements pourraient venir d'un module de localisation exploitant une représentation de plus bas niveau dans la hiérarchie.

Chaque proto-objet est alors sauvegardé sous forme d'un blob 3D: position relative au Robot, et volume approximatif; dans nos travaux, ce volume est approché par une sphère. Chaque blob est ensuite traité par le processus attentionnel: une caméra active PTZ (pan/tilt/zoom) est commandée pour focaliser l'image successsivement sur un proto-objet; les images ainsi obtenues servent à créer un modèle d'apparence.

Il existe plusieurs politiques possibles pour la construction des modèles de proto-objets:

- au fur et à mesure des déplacements du robot, la caméra reste focalisée sur un proto-objet tant que celui-ci est dans son champ de vision: le robot "erre" dans l'environnement pendant cette phase d'exploration, soit en faisant des mouvements aléatoires dans l'espace libre, soit en étant guidé par l'opérateur.
- ou bien les déplacements du robot sont contraints pour pouvoir effectuer une modélisation "complète" du proto-objet. Un planificateur de type *Next Best View* permet de calculer la meilleure position du capteur pour compléter le modèle.

Cependant, pour des raisons techniques et de temps, notre système n'a pu être que partiellement intégré sur un démonstrateur. Ces considérations comportementales feront donc l'objet de travaux ultérieurs.

Si le système présenté peut se comporter de manière complètement autonome en ce qui concerne la localisation et la modélisation des proto-objets, il est cependant incapable de donner un sens à ces données, et notamment de statuer sur la pertinence de ce proto-objet. Un opérateur humain doit alors intervenir pour intégrer les informations sémantiques (notamment les nominations des objets et éventuellement certaines caractéristiques telles que leur mobilité, leur possible déplacement par l'Homme...). Par ce biais, il procède également à un nettoyage de la base de données en supprimant les proto-objets n'ayant aucune valeur sémantique.

3.5 Conclusion

Nous avons dans ce chapitre décrit les deux projets pour lesquels ce travail sur la compréhension de l'environnement a été réalisé. Nous avons d'abord été "mobilisé" pour le développement du robot guide de musée RACKHAM. Cela nous a permis d'appréhender plusieurs approches déjà existantes dans le groupe RIA du LAAS-CNRS, pour représenter

l'environnment dans lequel un robot doit se déplacer:

- carte stochastique construite avec des segments-laser ou des amers visuels, nécessaire pour la localisation et pour la recherche de l'espace navigable;
- carte topologique utilisée pour la planification des chemins (ou routes);
- carte sémantique indispensable dans un lieu public, pour établir le dialogue Homme-Robot.

Notre contribution dans ce projet a été essentiellement technique sur le robot RACKHAM, décrit en annexe A; nous avons développé le module POM -*POsition Manager*- intégré dans la couche fonctionnelle du robot, en charge d'une part de la fusion des diverses estimations de la position du robot, et d'autre part des transformations de coordonnées entre les différents repères existants sur le robot (repères caméra, platine, robot et monde).

Nous avons ensuite participé au projet européen COGNIRON (Janvier 2004-Février 2008), plus spécialement au lot de travail dévolu à l'apprentissage des représentations spatiales par un robot compagnon de l'homme. KTH s'est intéressé aux approches intéractives de type SLAM à l'aide d'un capteur laser; UVA a porté son effort sur la construction d'un modèle topologique de l'environnement, à partir d'une base d'images panoramiques acquises par une caméra omnidirectionnelle; les contributions de UKA et IPA ont plus porté sur la construction de modèle d'objets 3D. Enfin, EPFL et LAAS ont choisi des approches exploitant des objets pour décrire l'environnement:

- S.Vasudevan de EPFL définit a priori hors ligne les modèles d'objets (portes, bouteilles...) que le robot est susceptible de trouver dans l'environnement. Durant son exploration, le robot construit un graphe d'agencement des objets qu'il reconnaît dans les images acquises lors de ses déplacements.
- nous avons au LAAS, proposé une approche plus prospective: lors de l'exploration, le robot détecte d'abord des régions saillantes, appelées *proto-objets* car elles sont supposées contenir un objet intéressant; puis il modélise en ligne l'apparence de ces objets selon différents aspects (ou points de vue).

Nous développerons la détection et la modélisation des proto-objets respectivement dans les chapitres suivants 4 et 5.

Chapitre 4

La découverte de l'environnement

Ce chapitre est donc consacré au processus pré-attentif, qui a pour but de générer une liste de régions d'intérêt détectées dans une séquence d'images acquises à bord du robot, soit par une caméra omnidirectionnelle, fixe sur le robot, soit par une caméra large champ.

Dans la configuration actuelle du démonstrateur exploité au LAAS-CNRS, nous avons surtout travaillé à partir d'images omnidirectionnelles; nous verrons que la sélection de régions saillantes en ce cas n'est pas un problème simple. Pour cette raison, nous avons aussi évalué notre méthode avec des images que nous avons acquises avec un simple appareil photo équipé d'un objectif large champ. Il est possible alors d'obtenir de meilleurs résultats, en évitant les problèmes fréquents de saturation du capteur dans les zones proches des portes et fenêtres. Nous verrons en conclusion, comment mettre en oeuvre une telle solution sur le robot.

Auparavant, nous allons après une courte présentation des modèles computationnels de l'attention visuelle, décrire comment nous construisons des cartes de saillance, comment nous y détectons des régions d'intérêt et comment ces régions sont suivies et grossièrement localisées pour être conservées dans une mémoire à courte durée.

4.1 Modèles computationnels de l'attention visuelle

Il existe aujourd'hui dans la littérature plusieurs modèles computationnels plus ou moins directement inspirés des théories psycho-physiques de l'attention visuelle.

[Tsotsos 95] ont présenté un modèle adaptatif sélectif de la vision attentionnelle. Le processus visuel est représenté par une pyramide neuronale et la sélection spatiale s'effectue par inhibition des connexions non pertinentes de la pyramide. La sélection attentionnelle est effectuée à l'aide d'un réseau WTA[1] descendant.

1. WTA="Winner-Take-All". Ce modèle de réseau simule les mécanismes de compétition existant entre

Le modèle attentionnel basé sur l'utilisation d'une carte de saillance fut proposé par [Koch 85] et implémenté par [Itti 98]. Ce modèle biologiquement plausible repose sur la célèbre "Théorie de l'intégration d'attributs" proposée par [Treisman 80] pour la vision attentive humaine. Selon cette théorie, l'image source est tout d'abord décomposée en un ensemble de cartes d'attributs topographiques. Les différents lieux ainsi mis en évidence, sont alors mis en compétition afin que seuls ceux qui se détachent significativement de leur voisinage subsistent et forment ainsi une carte des lieux saillants dans l'image source.

[Deco 00] adapte la résolution spatiale de l'image à l'aide d'un signal de contrôle attentionnel descendant. Avec une utilisation intensive de boucles de retour et des connexions étendues, [Hamker 05] modélise les interactions entre différentes régions du cerveau impliquées dans les processus de la vision attentionnelle, donnant ainsi un modèle respectant les données physiologiques et comportementales disponibles sur le sujet dans la littérature. Extension du modèle de compétition biaisée de [Duncan 97], [Sun 03] ont développé et implémenté un framework commun pour l'attention visuelle sur objets ou sur lieux par l'utilisation des groupements perceptuels.

Ces méthodes attentionnelles de la vision ont été exploitées par plusieurs auteurs pour des applications robotiques: citons en milieu extérieur, plusieurs travaux sur la recherche active d'amers, en exploitant en particulier, la couleur [Todt 04] ou sur l'attention visuelle pour les robots qui agissent dans des milieux humains [Frintrop 06].

Notre système pré-attentif s'appuie sur le modèle de saillance de [Itti 98] résumé dans la figure 4.1. Celui-ci est utilisé dans de nombreux travaux en vision cognitive et s'intégre de plus en plus en robotique. De plus, ne nous intéressant pas au parcours visuel de l'attention, nous avons pu nous affranchir du réseau de neurones WTA et utiliser des approches plus traditionnnelles de traitement d'images.

4.2 Détection des régions d'intérêt saillantes

4.2.1 Prétraitement

Pour pouvoir calculer la carte de saillance, il faut d'abord extraire de l'image source les différentes composantes de bas niveau du modèle: luminance, oppositions de couleurs et orientations (se référer à [Itti 98] pour l'explication neurophysique des choix des attributs de bas niveaux et la bibliographie correspondante)

neurones ou populations de neurones. Après convergence, seul le neurone ayant la plus grande activité reste actif et inhibe tous les autres

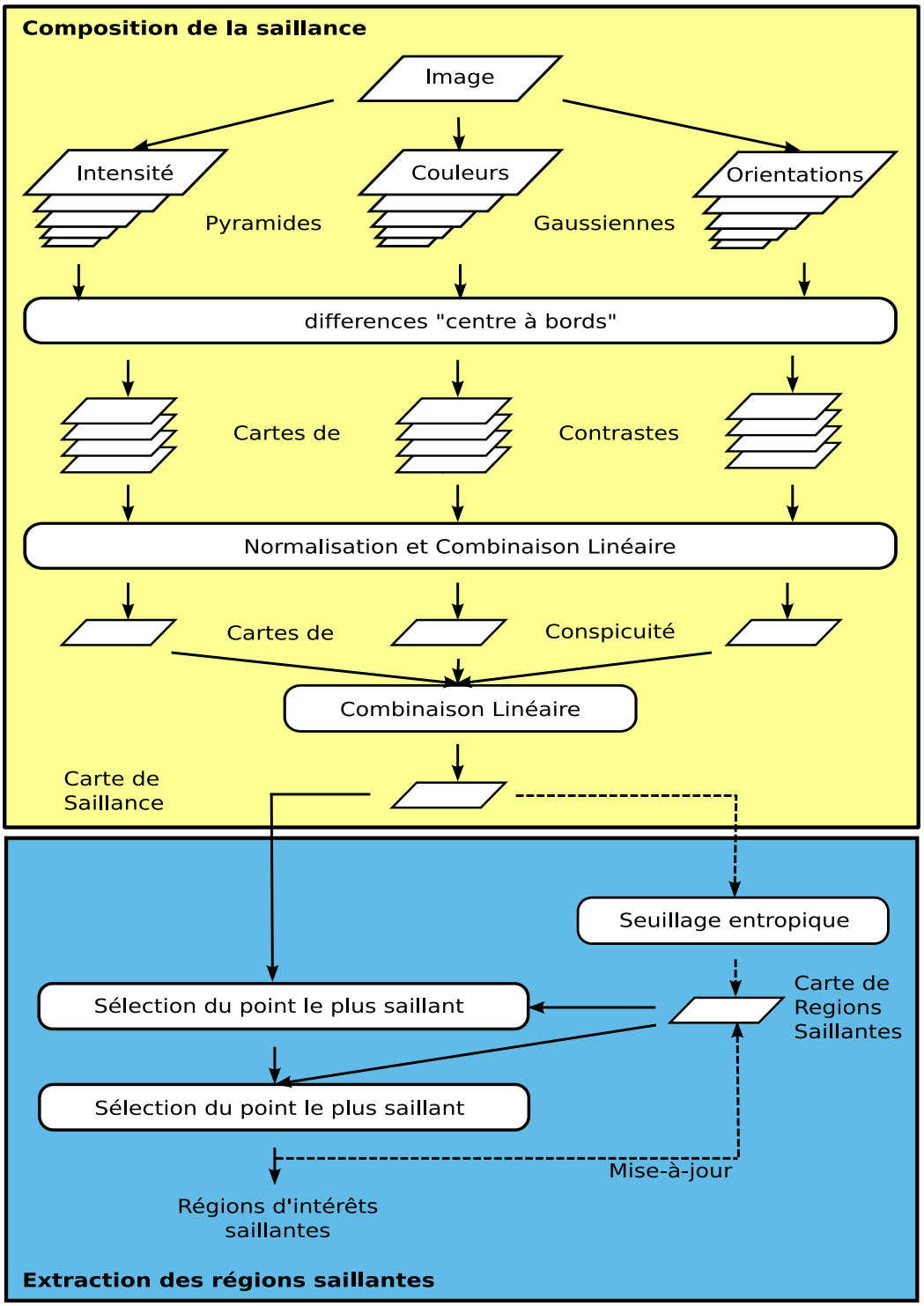

FIG. 4.1 – *Modèle ascendant de la saillance*

En considérant r, v et b les canaux couleurs de l'image source, l'image de luminance est calculée comme:

$$\mathcal{M}_I = \frac{r+v+b}{3} \tag{4.1}$$

$\mathcal{M}_I$ est alors utilisé pour générer une pyramide gaussienne dyadique en appliquant un filtre gaussien séparable linéaire et en divisant sa taille par un facteur 2. Ce procédé est appliqué récursivement pour obtenir tous les niveaux $\mathcal{M}_I(\sigma)$ de la pyramide, avec $\sigma = [0,\ldots,8]$. La résolution du niveau σ est égale à $1/2^\sigma$ fois la résolution de l'image originale (cf. [Burt 83] pour de plus de détails).

Ensuite, les canaux couleurs $\hat{r}$, $\hat{v}$, $\hat{b}$ normalisés par la luminance sont calculés afin de réduire l'influence de la luminance sur la perception de la teinte. Cependant, les variations de teintes étant difficilement détectables pour de faibles valeurs de luminance (et par conséquent elles ne peuvent être saillantes), le calcul n'est effectué que pour les pixels de $\mathcal{M}_I$ dont l'intensité est supérieure au dixième du maximum de luminance $\mathcal{M}_I^{\max}$.

$$\hat{r} = \begin{cases} r/\mathcal{M}_I & \text{si} \quad \mathcal{M}_I > \mathcal{M}_I^{\max}/10 \\ 0 & \text{sinon} \end{cases} \tag{4.2a}$$

$$\hat{v} = \begin{cases} v/\mathcal{M}_I & \text{si} \quad \mathcal{M}_I > \mathcal{M}_I^{\max}/10 \\ 0 & \text{sinon} \end{cases} \tag{4.2b}$$

$$\hat{b} = \begin{cases} b/\mathcal{M}_I & \text{si} \quad \mathcal{M}_I > \mathcal{M}_I^{\max}/10 \\ 0 & \text{sinon} \end{cases} \tag{4.2c}$$

Les composantes d'opposition de couleurs Rouge-Vert $\mathcal{M}_{RV}$ et Bleu-jaune $\mathcal{M}_{BJ}$ sont alors calculées à partir des quatre canaux couleurs affinés R, V, B et J:

$$R = \max(0, \hat{r} - \frac{\hat{v}+\hat{b}}{2}) \tag{4.3a}$$

$$V = \max(0, \hat{v} - \frac{\hat{r}+\hat{b}}{2}) \tag{4.3b}$$

$$B = \max(0, \hat{b} - \frac{\hat{v}+\hat{r}}{2}) \tag{4.3c}$$

$$J = \max(0, \frac{\hat{r}+\hat{v}}{2} - \frac{|\hat{r}-\hat{v}|}{2} - \hat{b}) \tag{4.3d}$$

$$\mathcal{M}_{RV} = R - V \tag{4.4a}$$

$$\mathcal{M}_{BJ} = B - J \tag{4.4b}$$

Comme pour la luminance, nous générons deux pyramides gaussiennes $\mathcal{M}_{RV}(\sigma)$ et $\mathcal{M}_{BJ}(\sigma)$.

Enfin, les quatre pyramides d'orientation sont obtenues par convolution de chaque niveau de $\mathcal{M}_I(\sigma)$ avec un filtre de Gabor d'angle donné $\theta = \{0°,45°,90°,135°\}$.

4.2.2 Composition des cartes de contrastes

Les cartes de contrastes sont obtenues par l'opérateur "différence centre-contour", simulant ainsi les différences de sensibilité du système visuel humain entre le centre et les bords de la rétine: en effet, les neurones photosensibles de nos yeux se montrent les plus sensibles dans un petite région du champ visuel (le centre), alors que les stimuli présents dans une plus large région antagoniste concentrique au centre (le contour) inhibent la réponse neuronale. Une telle architecture, sensible aux discontinuités spatiales locales, est particulièrement adaptée à la détection des lieux saillants du champ visuel, ceux qui se distinguent le plus de leur voisinage.

L'opérateur "différence centre-contour" $\ominus$ est obtenu par différence entre niveau fin et niveau grossier des pyramides dyadiques issues du prétraitement. Dans notre système, le centre est un pixel au niveau $c \in \{2,3,4\}$ et la région contournante est consitituée des pixels aux niveaux $s = c+\delta$, $\delta \in \{3,4\}$. Seuls les niveaux 2 à 8 de la pyramide sont utilisés car les niveaux 0 et 1 sont trop précis, pas assez lisses et les niveaux au delà de 8 ne sont plus significatifs.

Ainsi, nous obtenons:

$$\mathcal{F}_{l,c,s} = \mathcal{N}(|\mathcal{M}_l(c) \ominus \mathcal{M}_l(s)|) \quad \forall l \in L = L_I \cup L_C \cup L_O \tag{4.5}$$

avec

$$L_I = \{I\}, \quad L_C = \{RG,BY\}, \quad L_O = \{0°,45°,90°,135°\}$$

$\mathcal{N}(\cdot)$ est un opérateur de normalisation non linéaire itératif, ayant pour rôle de simuler la compétition locale entre les lieux saillants voisins (cf. [Itti 01]). A chaque itération, la carte est convoluée par une différence de gaussienne, suivie d'une rectification.

Ensuite, ces cartes sont combinées à une échelle spécifique (ici $\sigma = 4$) par addition multi-échelle; puis elles sont normalisées à nouveau.

$$\bar{\mathcal{F}}_l = \mathcal{N}\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{F}_{l,c,s}\right) \quad \forall l \in L \tag{4.6}$$

Les cartes de conspicuité pour la luminance, la couleur et l'orientation, sont issues de la normalisation des cartes obtenues par sommation des sous-cartes calculées aux différentes

échelles (dans le cas de la luminance, la carte de conspicuité est égale à sa sous-carte calculé dans eq.4.6). La combinaison finale des cartes de conspicuité forme notre carte de saillance.

$$\mathcal{C}_I = \bar{\mathcal{F}}_I, \quad \mathcal{C}_C = \mathcal{N}\left(\sum_{l \in L_C} \bar{\mathcal{F}}_l\right), \quad \mathcal{C}_O = \mathcal{N}\left(\sum_{l \in L_O} \bar{\mathcal{F}}_l\right) \tag{4.7}$$

$$\mathcal{S} = \frac{1}{3} \sum_{k \in \{I,C,O\}} C_k \tag{4.8}$$

La figure 4.2 présente un exemple de carte de saillance et de cartes de conspicuité obtenu par notre système.

4.2.3 Extraction des régions saillantes

La carte de saillance obtenue au paragraphe précédant nous fournit donc pour chaque pixel de l'image une mesure de son intérêt, sa saillance. Dans [Itti 98], c'est un réseau de neurones de type "Winner-take-all" qui se charge de sélectionner itérativement les lieux saillants. A chaque pas, le pixel de valeur maximale définit le point saillant courant, c'est-à-dire celui où se porte l'attention. Comme pour la vision humaine, cette attention porte également sur la région circulaire de rayon r fixe centrée sur ce point. Cette région est alors utilisée comme inhibition au pas suivant, permettant ainsi de focaliser l'attention sur le prochain point saillant.

Cette définition ponctuelle des lieux saillants est suffisante lorsque l'on cherche à modéliser le chemin suivi par le mécanisme d'attention visuelle. Mais dans notre cas, où nous cherchons à mettre en évidence des proto-objets (dans le meilleur des cas, des objets) de l'environnement, il est évidemment impossible de réduire un lieu saillant à un point, ni même à une région circulaire de rayon fixé.

Dans un premier temps, nous cherchons à classifier chaque pixel comme appartenant à un objet ou au fond. Pour cela, nous utilisons une méthode de seuillage entropique définie par [kapur 85]. L'idée est de réaliser une partition binaire *objets-fond* telle que la somme de leur entropie soit maximale. Selon la théorie de l'information, l'entropie est une mesure de quantité d'information du système. Soit un ensemble fini $S = s_1, s_2, \ldots, s_k$ et

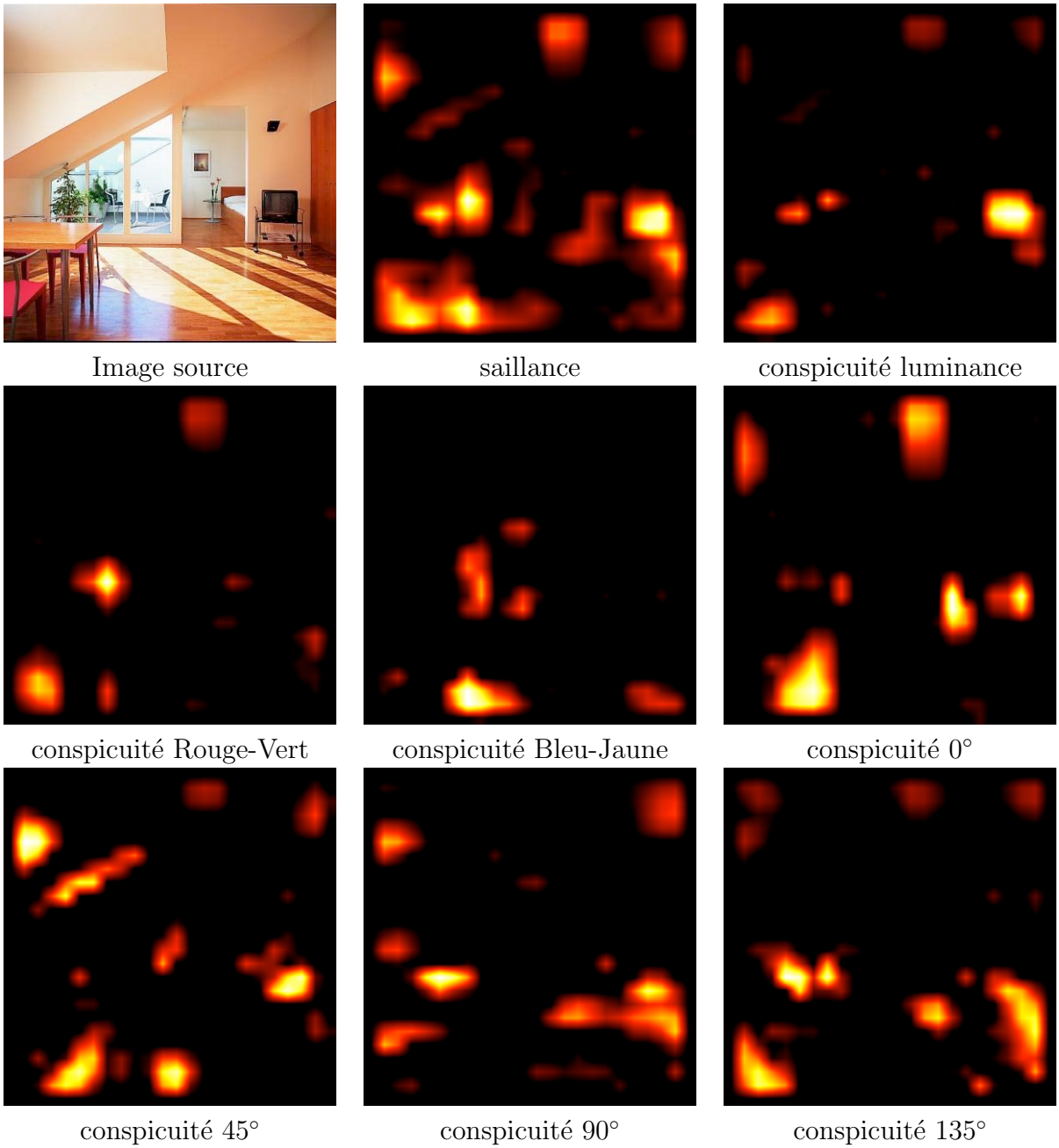

FIG. 4.2 – *Exemple de cartes de conspicuité et de saillance*

p_i la probabilité d'occurrence de chaque élément s_i, l'entropie est définie par:

$$H = -\sum_{i=1}^{k} p_i * log(p_i) \tag{4.9a}$$

$$\text{où} \quad \sum_{i=1}^{k} p_i = 1 \tag{4.9b}$$

Ici, S représente les niveaux de gris et leur probabilité est donnée par l'histogramme normalisé de l'image. En reprenant les mêmes notations, le seuil optimal est donc obtenu ainsi:

$$s_{optimal} = \max_t(H_{objet} + H_{fond}) \tag{4.10a}$$

$$H_{objet} = -\sum_{i=0}^{t} \frac{p_i}{P_t} \log \frac{p_i}{P_t} \tag{4.10b}$$

$$H_{fond} = -\sum_{i=t+1}^{k} \frac{p_i}{1-P_t} \log \frac{p_i}{1-P_t} \tag{4.10c}$$

$$\text{où} \quad P_t = \sum_{i=0}^{t} p_i \tag{4.10d}$$

La prise en compte des distributions de probabilité P_t de l'objet et $(1 - P_t)$ du fond se fait lors du calcul de l'entropie de la partition. D'après [Sezgin 04], cette méthode de classification binaire est l'une des plus performante sur les images de niveaux de gris et rivalise même avec des méthodes plus sophistiquées comme la classification bayésienne ou les chaines de Markov.

La binarisation ainsi obtenue nous donne une carte des proto-objets de l'image, qui est utilisée comme masque lors de leur extraction par sélection itérative du point le plus saillant dans la carte de saillance et croissance de régions à partir de ce point dans le masque binaire. Celui-ci est automatiquement mis-à-jour par l'algorithme de croissance de régions, empéchant ainsi tout point des objets déjà extraits d'être sélectionnné comme prochain point le plus saillant. Enfin, afin de limiter l'extraction aux quelques objets les plus saillants, l'itération se termine lorsque la saillance du point sélectionné est inférieure au seuil $S_t = \overline{S} + k(\hat{S} - \overline{S})$, $k \in [0,1]$ où $\overline{S}$ est la saillance moyenne e $\hat{S}$ le maximum de la saillance.

La figure 4.3 montre les résultats de l'extraction de régions sur des images génériques.

Sur des images panoramiques prise dans la grande salle robotique du LAAS à la

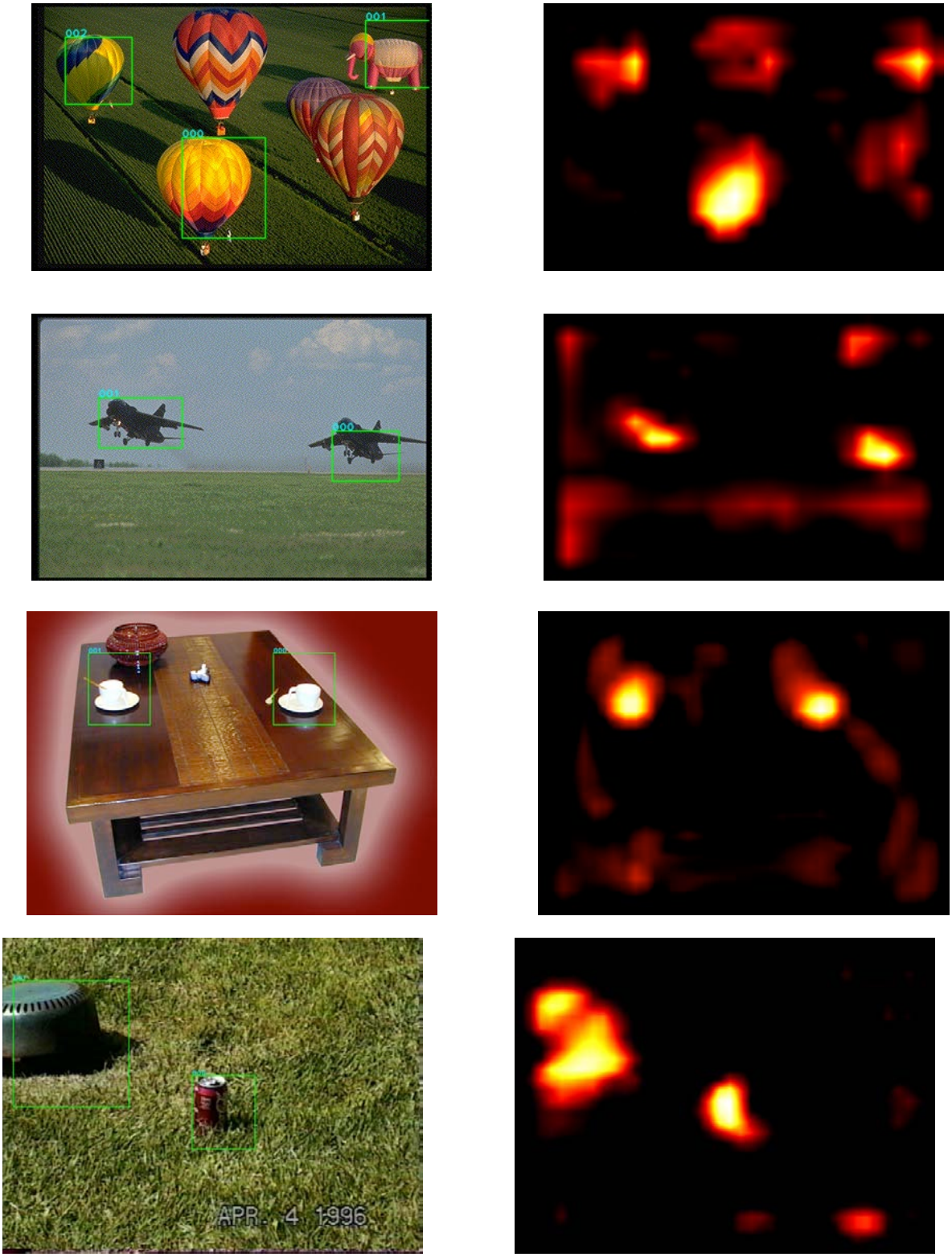

FIG. 4.3 – A gauche*: l'image source avec les proto-objets.* A droite*: la carte de saillance*

figure 4.4, le résultat est dans une certaine mesure conforme à nos attentes, certaines régions étant effectivement extraites sur des objets ou parties d'objets. Néanmoins, nous noterons la sensibilité particulière de notre système aux fortes intensités lumineuses.

4.3 Intégration à la carte de l'environnement

Comme présenté au chapitre 3, dans la section 3.4, nous avons décidé d'augmenter la carte de navigation du robot avec les proto-objets, situés dans l'espace 3D du robot, donc caractérisés par leur position et leur occupation 3D grossières. Ne possédant pas directement d'informations stéréoscopiques (utilisation d'une seule caméra omnidirectionnelle ou large champ), nous devons utiliser les informations de déplacements du robot (nous rappelons que nous considérons la position du robot connu à chaque instant, issue d'un module tiers). Nous avons donc décidé de mettre en oeuvre une fonction de suivi des régions saillantes, puis d'effectuer pour chaque proto-objet, un ajustement de faisceaux à partir de toutes les images dans lesquelles la région saillante correspondante a été suivie.

4.3.1 Suivi des régions saillantes

L'extraction des régions saillantes vue dans la section précédente nous fournit pour chaque région, les informations suivantes:

1. la position (x,y) de la région
2. la taille (w,h) de la boite englobante
3. l'aire a de la région (c'est-à-dire le nombre de pixels de la boite englobante appartenant effectivement à la région)
4. l'histogramme normalisé teinte-saturation de la région

Du fait des multiples normalisations lors de la construction de la carte de saillance, les régions extraites sont très peu discriminantes du point de vue de leur apparence; il est donc impossible de réaliser l'association de données nécessaires à la reconstruction 3D avec des méthodes d'appariement classiques sur la carte de saillance. Nous nous sommes donc orientés vers des méthodes de suivi de régions, notamment le suivi couleur CAMSHIFT disponible dans la bibliothèque OpenCV: le traitement du suivi était alors effectué sur l'image couleur source, initialisée par les régions saillantes que nous avions extraites. Pour notre application, cette méthode de suivi s'est rapidement avérée peu stable lorsque la région initiale était de couleur peu homogène.

Néanmoins, dans des conditions d'illuminations satisfaisantes, le calcul de la saillance est relativement stable en translation dans l'image: un point saillant dans l'image I_t aura à

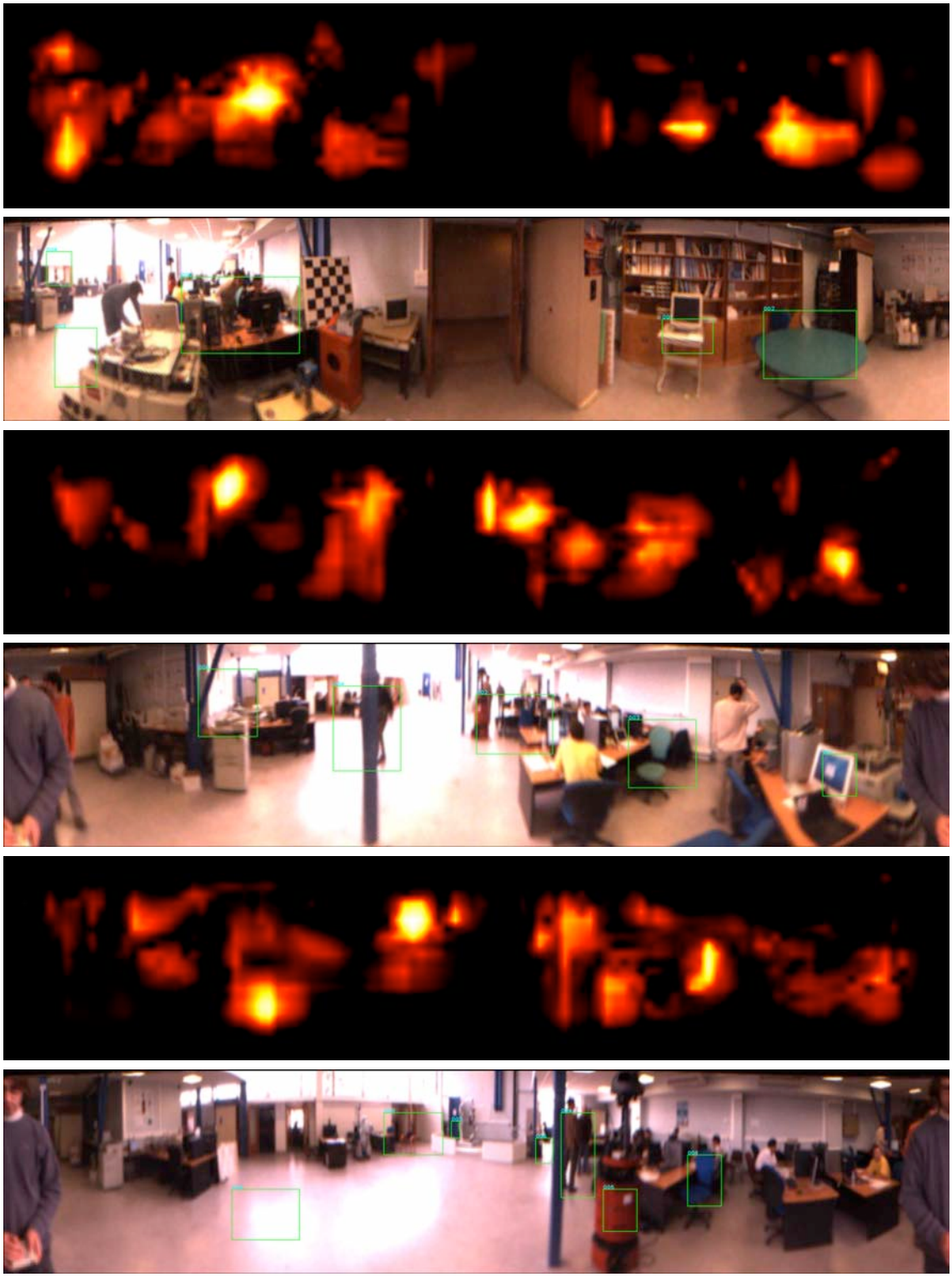

FIG. 4.4 – *Extraction de proto-objets sur images panoramiques*

peu près le même degré de saillance dans l'image I_{t+1}, au déplacement près. Suite à cette observation, nous avons décidé d'implémenter notre propre suivi de régions connexes, basé sur un simple filtre de Kalman et une mesure d'intérêt pour résoudre le problème de l'association de données.

Modèle de mouvement et filtrage de Kalman

Notre moteur de suivi utilise donc un filtre de Kalman pour prédire et minimiser les erreurs de trajectoires. Nous ne développerons ici que les équations fondamentales du filtrage de Kalman pour le suivi et nous invitons les lecteurs à étudier l'article de Welch et Bischop[Welch 01] pour avoir une explication plus poussée.

Nous utilisons un modèle de mouvement de premier ordre en position et taille défini comme suit:

$$s_k = As_{k-1} + w_{k-1} \tag{4.11a}$$

$$z_k = Hs_k + v_k \tag{4.11b}$$

avec

- **observation:** $z_k = \begin{bmatrix} x & y & w & h \end{bmatrix}^T$
- **vecteur d'état:** $s_k = \begin{bmatrix} x & y & w & h & \dot{x} & \dot{y} & \dot{w} & \dot{h} \end{bmatrix}^T$
- **matrice d'observation:** $H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$
- **matrice de transition:** $A = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$
- **bruit gaussien du système:** $w \sim \mathcal{N}(0,Q)$ avec $Q = 10^{-5}I$, $I \in \mathbb{R}^8$
- **bruit gaussien de mesure:** $v \sim \mathcal{N}(0,R)$ avec $R = 10^{-1}I$, $I \in \mathbb{R}^4$

En considérant P la covariance de l'erreur et K le gain du filtre, les deux étapes du

filtrage sont les suivantes :

1. **Prédiction:** calcul des estimations *a priori* de l'état et de la covariance de l'erreur du système

$$\hat{s}_k^- = A\hat{s}_{k-1} \tag{4.12a}$$
$$P_k^- = AP_{k-1}A^T + Q \tag{4.12b}$$

2. **Correction:** calcul des estimations *a posteriori* de l'état et de la covariance de l'erreur du système à partir d'une nouvelle observation

$$K_k = P_k^- H^T (HP_k^- H^T + R)^{-1} \tag{4.13a}$$
$$\hat{s}_k = \hat{s}_k^- + K_k(z_k - H\hat{s}_k^-) \tag{4.13b}$$
$$P_k = (I - K_k H)P_k^- \tag{4.13c}$$

Mise en oeuvre du moteur de suivi

Dans notre moteur de suivi, nous considérons deux types de données : les trajectoires Tr et les candidats C. Chaque entité saillante suivie est représentée par une trajectoire Tr obtenue par filtre de Kalman et par la liste des régions associées à cette entité depuis le début du suivi. Notre modèle de mouvement du premier ordre nous oblige à avoir une association de deux régions à deux instants successifs afin d'avoir une estimée de la vitesse de l'entité suivie pour pouvoir initialiser correctement le filtre de Kalman. Ainsi, à chaque itération, nous effectuons d'abord la mise en association des nouvelles régions saillantes avec les trajectoires actuellement actives dans le moteur de suivi: les positions et tailles prédites par le filtrage de Kalman servent de critères pour une recherche de la région saillante la plus ressemblante dans l'image courante. Cette région n'est cependant ajoutée à la trajectoire et utilisée pour la mise à jour de cette trajectoire par filtre de Kalman, uniquement si sa confiance (cf. paragraphe 4.3.1) est supérieure à un seuil préalablement fixé. Les régions saillantes restantes sont ensuite associées aux régions candidates extraites à l'itération précédente, associations validées là-aussi par la mesure de similarité et donnant lieu à l'initialisation de nouvelles trajectoires. A la fin de l'itération, les trajectoires non mises à jour sont sauvegardées et les régions saillantes qui n'ont été associées ni à une trajectoire ni à un candidat de l'itération précédente, forment les nouveaux candidats.

Le fonctionnement du suivi est résumé dans l'algorithme 1.

Algorithme 1 : Itération du moteur de suivi

Données : regions saillantes $ListSR^t$, trajectoires $ListTR^t$, candidats $ListCA^t$, seuil de similitude sth

Résultat : trajectoires $ListTR^{t+1}$, candidats $ListCA^{t+1}$

$ListTR^{t+1} = \varnothing$, $ListCA^{t+1} = \varnothing$

// les trajectoires en cours sont mise à jour en premier

pour tous les $Tr \in ListTR^t$ **faire**

  si $SR^t \neq \varnothing$ **alors**

    $\overline{sr} = \underset{sr \in ListSR^t}{argmax}$ `Similitude` $(sr, Tr.kalman\_prediction)$

    si `Similitude` $(\overline{sr}, Tr.kalman_prediction) > sth$ **alors**

      mise à jour de Tr avec $\overline{sr}$

      ajout de Tr à $ListTR^{t+1}$

      suppression de $\overline{sr}$ de $ListSR^t$

    fin

  fin

fin

// recherche des nouvelles trajectoires à partir des candidats

pour tous les $C \in ListCA^t$ **faire**

  si $ListSR^t \neq \varnothing$ **alors**

    $\overline{sr} = \underset{sr \in ListSR^t}{argmax}$ `Similitude` (sr, C)

    si `Similitude` $(\overline{sr}, C) > sth$ **alors**

      initialisation de Tr avec $\overline{sr}$ et C

      ajout de Tr à $ListTR^{t+1}$

      suppression de $\overline{sr}$ de $ListSR^t$

    fin

  fin

fin

// les regions saillantes restantes forment les candidats pour l'itération suivantes

$ListCA^{t+1} = ListSR^t$

Mesure d'intérêt

Comme nous l'avons dit précédemment, l'association de données s'effectue d'abord par une recherche du plus proche voisin puis par l'attribution d'un degré d'intérêt à cette association.

Nous utilisons pour cela une mesure d'intérêt inspirée de [Dabis 96]. Cette mesure s'appuie sur la covariance d'erreur produite par le filtre de Kalman, pour prédire l'erreur sur les observations. En reprenant les notations du paragraphe 4.3.1, l'erreur $\mathcal{E}(z - H\hat{s}^-)(z - H\hat{s}^-)^T$ devrait être égale à HP^-H^T. Cette matrice, calculée lors du cycle de prédiction du filtrage de Kalman, nous donne ainsi une estimée du voisinage de chaque région prédite, dans lequel nous espérons trouver notre observation. La mesure est donc définie comme suit:

$$I(sr,Tr) = I_{pos}(\Delta x,\Delta y)I_{size}(\Delta w,\Delta h) \tag{4.14}$$

avec

$$I_{pos}(\Delta x,\Delta y) = e^{\alpha_{pos}\sqrt{\frac{\Delta x^2}{2*\sigma_x^2}+\frac{\Delta y^2}{2*\sigma_y^2}}}$$

$$I_{size}(\Delta w,\Delta h) = e^{\alpha_{size}\sqrt{\frac{\Delta w^2}{2*\sigma_w^2}+\frac{\Delta h^2}{2*\sigma_h^2}}}$$

où Δx, Δy, Δw et Δh représentent les différences sur la position et la taille entre la prédiction donnée par le filtre de Kalman de la trajectoire Tr et la région saillante sr. Les différents termes σ_i sont des déviations standards des prédictions antérieures du filtre de Kalman issues de la matrice HP^-H^T. Enfin, Les paramètres α_{pos} et α_{size} réglent la sensibilité de la mesure en position et taille, c'est-à-dire l'erreur maximale acceptable entre la prédiction et l'observation pour valider l'association.

Cette mesure d'intérêt assure qu'une mauvaise association aura un très faible score et sera donc rejetée. De plus, plus la trajectoire sera grande, plus la mesure sera pertinente dans la sélection de la bonne association.

Dans le cas d'une association avec un candidat C, ne disposant pas de P^-, les deux premiers termes sont remplacés par:

$$I_{pos}(\Delta x,\Delta y) = e^{-KP(\Delta x^2+\Delta y^2)} \quad \text{avec} \quad KP = -\frac{\log(0.1)}{(0.02*W)^2}$$

$$I_{size}(\Delta w,\Delta h) = e^{-KS(\Delta w^2+\Delta h^2)} \quad \text{avec} \quad KS = -\frac{\log(0.1)}{0.5^2}$$

avec W la largeur de l'image source.

la figure 4.6 montre le résultat de suivi de régions sur une séquence d'images panora-

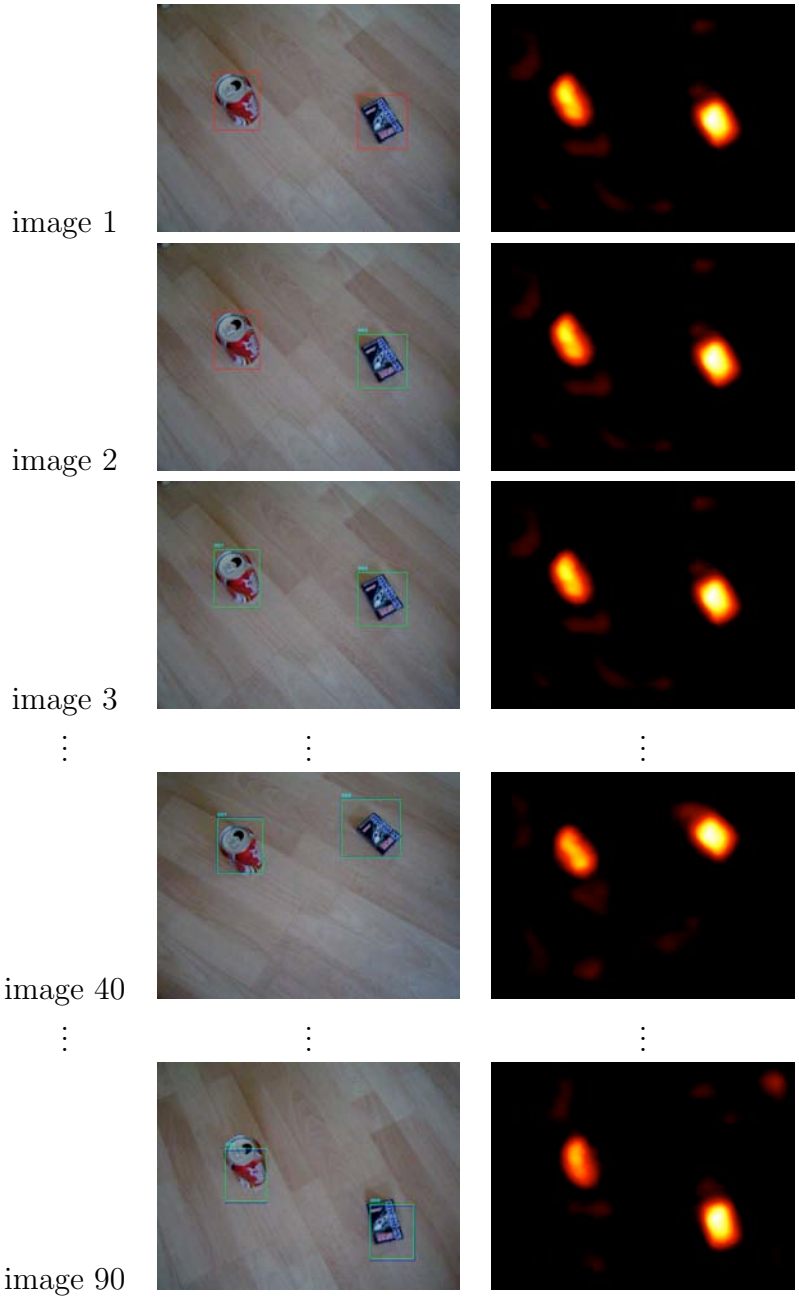

FIG. 4.5 – *Exemple de suivi.* rouge*: candidats.* bleu*:prédiction de l'objet suivi.* vert*: objet suivi avec son numéro d'identification*

miques.

Si ce modèle simple de suivi par filtrage de Kalman est satisfaisant au regard de nos tests, il serait intéressant par la suite de profiter des informations de vitesse et d'accélération disponible sur notre robot au niveau de la matrice de transition A.

image 12

image 14

image 16

image 18

FIG. 4.6 – *Suivi sur images panoramiques*

4.3.2 Ajustement de faisceau

Afin de focaliser le processus attentif sur une région d'intérêt, il convient de localiser cette région dans l'espace (pour piloter l'orientation d'une caméra PTZ) et d'en déterminer la taille (pour piloter la focale de cette même caméra). Nous n'avons pas besoin ni d'une localisation, ni d'une taille très précises, car il s'agit simplement de contrôler un processus visuel, et non d'agir sur l'environnement (saisie ou accostage). Les imprécisions sur la position et la taille 3D d'une région d'intérêt, seront prises en compte quand nous choisirons la focale de la caméra PTZ.

Nous disposons donc de N vues de notre région d'intérêt, N pouvant être grand dans la mesure où cette région est suivie dans la séquence d'images: pour chaque vue, nous disposons d'une position (u,v) de la région et de son contour: ce contour sera approximée par une ellipse 2D, définie par trois paramètres en sus de sa position dans l'image, (θ,a,b) pour l'orientation du grand axe, et la taille maximale sur les deux axes. Par ailleurs pour chaque vue, nous disposons de la position de la caméra dans l'environnement.

Avant de décrire rapidement, la méthode d'ajustement des faisceaux que nous avons mise en oeuvre, il convient de répondre à deux questions:

- *Dans quel repère doit s'effectuer cette localisation?*

  Idéalement il conviendrait de disposer déjà des fonctions de localisation du robot dans l'environnement. Mais cela nécessiterait de découpler l'exploration en plusieurs phases, un premier tour pour apprendre une carte métrique de primitives sensorielles (segments-laser, amers visuels...) exploitées uniquement pour la localisation, un deuxième tour pour construire notre représentation de type Graphe d'Objets. A moins que ces explorations soient totalement autonomes, un tel découplage est exclus car il prendrait trop de temps à l'Homme utilisateur de notre robot compagnon. En conséquence, nous proposons de localiser dans un repère local, posé par le Robot à l'entrée du lieu topologique en cours d'exploration: première position du robot en ce lieu, ou position de la porte d'accès à ce lieu (milieu de la porte par exemple): ce repère pourrait être posé

  - en mode autonome, quand le Robot détecte un franchissement de porte;
  - en mode intéractif, quand l'Homme indique un changement de lieu.

- *Comment modéliser la taille 3D de notre région d'intérêt?*

  Typiquement le volume d'une telle région 3D devrait être représenté par un ellipsoïde, donc avec six paramètres en sus de la position du centre de la région: trois paramètres d'orientation, et trois rayons. Nous avons commencé à poser les équations pour traiter ce problème, mais nous n'avons pas réussi à estimer ces six paramètres à partir de N vues de la région: nous ne sommes pas sûrs que cela soit observable. Aussi nous avons finalement simplifié le problème, en modélisant le volume d'une région d'intérêt par une simple sphère, donc avec un seul paramètre en sus de la position, le rayon.

Nous rappelons les équations utilisées pour traiter de ce problème, avant de donner quelques résultats.

L'ajustement de faisceaux

L'ajustement de faisceaux est une approche géométrique pour la localisation d'objets dans l'espace. « Etant donné une caméra en déplacement (de déplacement connu) et un objet dans l'espace, alors le croisement des rayons visuels donne la position de l'objet dans l'espace ». L'ajustement de faisceaux nécessite de connaître le modèle de la caméra et la relation entre la projection de l'objet dans l'image et ses coordonnées dans l'espace.

Le modèle de caméra que nous avons utilisé est le modèle Pinhole. Ce modèle est schématisé à la figure 4.7.

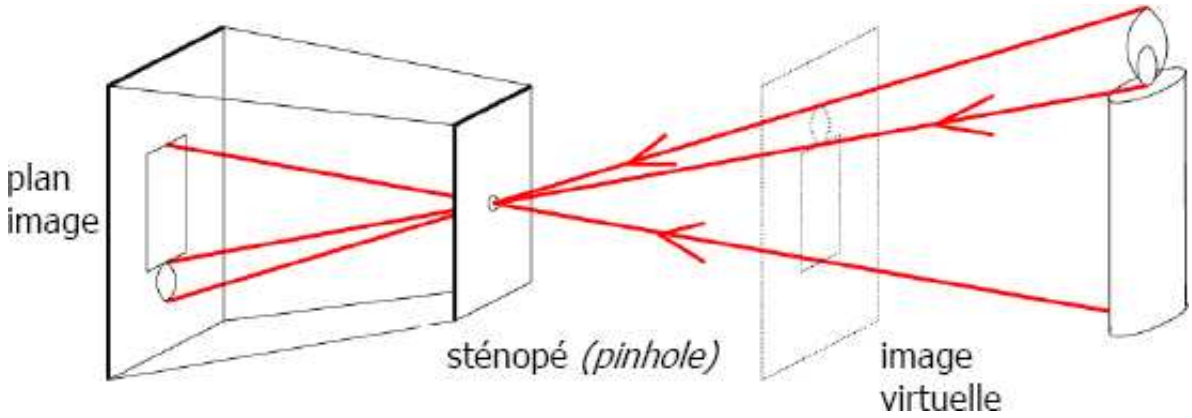

FIG. 4.7 – *Le modèle sténopé*

Le processus de projection génère une image 2D à partir d'objets 3D. Le processus est le suivant:
point 3D ∈ repère monde → point 3D dans repère caméra → point image dans le plan image *rightarrow* pixel dans la matrice.

En utilisant les transformations mathématiques classiques, la relation qui s'établit entre les coordonnées dans l'espace 3D et sa projection en pixel est :

$$\begin{pmatrix} su \\ sv \\ s \end{pmatrix} = IE \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{4.15}$$

avec $(X,Y,Z)^T$ les coordonnées d'un point M de le repère du monde et $(u,v)^T$ les coordonnées, respectivement, en ligne et en colonne du pixel dans lequel il se projette.

I $= \begin{pmatrix} k_u f & 0 & u_0 & 0 \\ 0 & k_v f & v_0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ et est appelée *matrice intrinsèque* puisqu'elle ne tient compte que des paramètres intrinsèques de la caméra.
E $= \begin{pmatrix} R & T \\ 0 & 1 \end{pmatrix}$ et est appelée *matrice extrinsèque*. Elle représente la matrice de passage entre le repère caméra et le repère du monde.

Si nous considérons que nous pouvons avoir différentes images d'un même point dans l'espace comme le montre la figure 4.8 alors nous obtenons un système d'équation qui servira à la détermination de ses coordonnées. Pour cela, réecrivons l'équation 4.15 pour chaque position i de la caméra sous cette forme :

$$\begin{pmatrix} su_i \\ sv_i \\ s \end{pmatrix} = M_i \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{4.16}$$

avec $M_i = IE_i = \begin{pmatrix} l_1^i \\ l_2^i \\ l_3^i \end{pmatrix} \in \Re_{3\times 4}$ alors nous pouvons dire que :

$$\begin{pmatrix} su_i \\ sv_i \\ s \end{pmatrix} = \begin{pmatrix} l_1^i \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} & l_2^i \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} & l_3^i \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \end{pmatrix}^T \tag{4.17}$$

ce qui revient à dire que

$$\begin{cases} (l_{11}^i - u_i l_{31}^i)x + (l_{12}^i - u_i l_{32}^i)y + (l_{13}^i - u_i l_{33}^i)z = u_i l_{34}^i - l_{14}^i \\ (l_{21}^i - v_i l_{31}^i)x + (l_{22}^i - v_i l_{32}^i)y + (l_{23}^i - v_i l_{33}^i)z = v_i l_{34}^i - l_{24}^i \end{cases} \tag{4.18}$$

avec l_{mn}^i la n^{ieme} composante de la ligne m.

Ainsi nous obtenons un système à double équation pour chaque position i de la caméra. Donc pour N vues, nous obtenons un système à 2N équations. La résolution du système $AX = B$ avec $A \in \Re_{2N\times 3}$, $B \in \Re_{2N\times 1}$ et $X \in \Re_{3\times 1}$ peut se faire avec n'importe quel algorithme approprié (voir [Triggs 99] pour les détails), nous avons choisi d'utiliser

celui des moindres carrés.
En utilisant ce procédé, nous pouvons déterminer la position d'un point dans l'espace mais il nous faudrait aussi pouvoir l'inscrire dans un « conteneur » afin de déterminer la taille de la région d'intérêt, taille nécessaire pour piloter la caméra exploitée pour la focalisation. Notons que cette taille pourrait être aussi exploitée pour déterminer une zone de manoeuvrabilité pour le robot autour de l'objet.

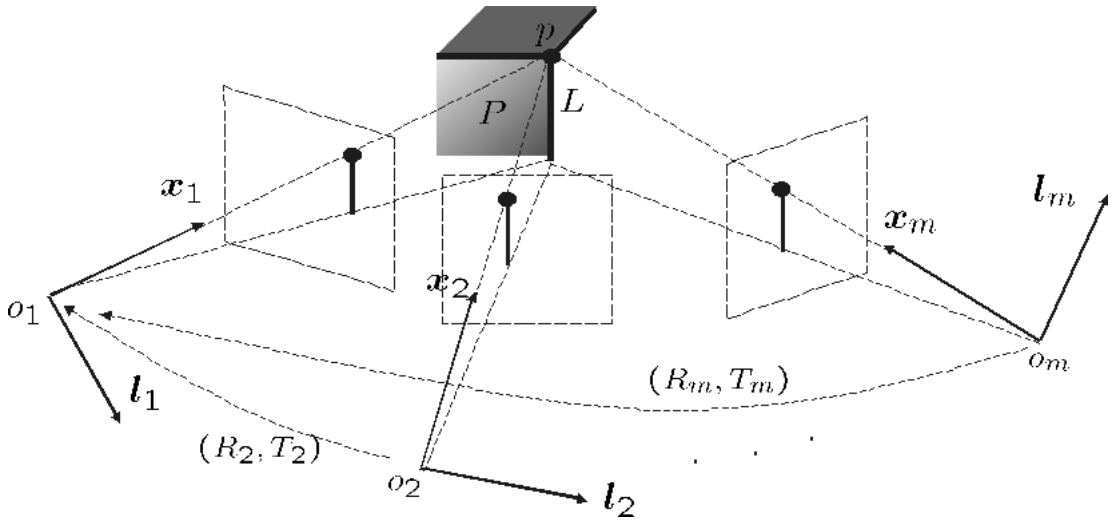

FIG. 4.8 – *Images multiples d'un même point dans l'espace*

Dans [Chaumette 90], Chaumette montre que la projection d'une sphère est une ellipse. Nous utilisons ce résultat pour dessiner les frontières de notre zone de manoeuvrabilité. Le centre de la sphère peut être calculé en utilisant l'ajustement de faisceaux. Le rayon de la sphère est déterminé par moyennage des rayons déterminés à partir des paramètres (θ,a,b) de l'ellipse projection de cette sphère en chacune des vues.

Bundler : l'ajustement des faisceaux étendus aux régions

La concrétisation de ces résultats est faite dans un module développé dans l'environnement JAFAR *bundleR* pour *bundle on region*. JAFAR, repose sur une architecture modulaire de telle sorte que pour chaque besoin cité plus haut, nous diposons d'un module spécifique.

module camera : permet d'obtenir les caractéristiques intrinsèques de la caméra ;

module datareader : se charge de la lecture et de la sauvegarde des données : images et localisation du robot ;

module geom : contient toutes classes et fonctions nécessaires au traitement géométrique ;

module jmath : c'est l'outil mathématique de JAFAR. Il permet un portage dans JAFAR des bibliothèques *boost* et *lapack* ;

module image : est responsable de la manipulation des fichiers images ;

module sas : ce composant est un outil de traitement des images par saillance pour en extraire les régions d'intérêt.

Dans la première implémentation, nous nous sommes attachés à valider le calcul de l'ajustement de faisceau uniquement sur le centre de la région. Les résultats obtenus sont intéressants mais restent très approximatifs, ceux-ci étant fortement dépendant de la qualité de la détection et du suivi.

4.4 Conclusion

Dans notre modèle de Vision Attentionnelle à deux processus, le premier, appelé processus pré-attentif, est donc chargé de détecter et d'extraire des régions d'intérêt depuis la séquence d'images acquises lors des déplacements effectués par le robot en phase d'exploration de l'environnement. Nous avons dans ce chapitre, décrit les méthodes développées pour mettre en oeuvre un tel processus.

Pour l'instant, ces méthodes ont été évaluées hors ligne, sur des séquences acquises depuis le robot ou par nous-même avec une caméra déplacée à la main. Vue l'architecture générale de la couche fonctionnelle sur les robots intégrés par le pôle Robotique et IA du LAAS-CNRS, et vu notre travail sur l'environnement de développement JAFAR (voir une description en Annexe B), l'intégration de ce processus pré-attentif sur le robot sera très rapide, et nous comptons avoir des résultats sur le robot à court terme.

Vu les résultats obtenus sur les images panoramiques, il conviendra de rajouter une couche de sélection des zones de l'image dans lesquelles rechercher des régions d'intérêt, cela pour éviter de sortir des régions dans des zones très saturées. Au fur et à mesure des déplacements les zones inhibées pour ce traitement vont également se déplacer dans les images, permettant au processus d'analyser tout l'espace à explorer dans le lieu courant, à partir de quelques images. L'analyse de l'image ne se faisant pas sur toute l'image, les contraintes de performance seront moins critiques.

Une autre solution consisterait à exploiter plutôt une caméra perspective équipée d'un objectif grand angle (typiquement, 120deg): mais comment explorer tout l'espace en ce cas? Plusieurs solutions existent:

– cette caméra pourrait être fixe, mais la stratégie de déplacement du robot durant la

phase d'exploration serait plus complexe pour couvrir tout l'espace du lieu courant avec le champ visuel de cette caméra.

- cette caméra pourrait aussi être orientable, au moins en azimut, afin d'éviter de rendre trop complexe les mouvements du robot. Rappelons que l'Homme pourrait guider le Robot pendant cette phase d'exploration: donc la plateforme resterait asservie aux mouvements de l'Homme, tandis que cette caméra serait contrôlée de manière autonome pour explorer tout l'espace.
- enfin, nous pourrions monter cette caméra large champ, sur la même tête orientable en site et azimut, que la caméra de focalisation exploitée pour la modélisation des objets. Il conviendrait alors de mettre en oeuvre une stratégie plus complexe de contrôle de cette tête orientable pour éviter la recherche de régions d'intérêt dans des zones saturées tout en continuant de focaliser sur une région en cours de modélisation.

En ce dernier cas, nous nous rapprocherions du modèle de la Vision Humaine: vision fovéale, avec résolution importante au centre et grossière en périphérie. Ne renoncerait-t'on pas trop vite aux possibilités de la technologie? Le developpement dans les années 90 des caméras fovéales (par exemple, le capteur développé par l'IMEC en Belgique) n'a pas à notre connaissance, connu de grand succès, à l'inverse de l'introduction à la même époque des caméras panoramiques, qui sont de plus en plus utilisées dans la communauté Robotique.

De ce fait, pour des travaux futurs, nous préconisons que le processus pré-attentif exploite une caméra panoramique, ou qu'il ait le contrôle total d'une caméra perspective large champ orientable en azimut, afin de découpler processus pré-attentif et attentif.

Chapitre 5

La modélisation des proto-objets

Comme nous l'avons vu dans les chapitres 2 et 3, pour un maximum d'interactivité avec l'homme, le robot doit pouvoir interpréter son environnement en terme d'objets, ce qui nécessite de pouvoir reconnaître un grand nombre d'objets ou de catégories d'objets. Dans la pratique, il est inconcevable qu'un robot puisse connaître l'ensemble des objets qui forment son environnement a priori et doit donc être capable de les découvrir par lui même et/ou de les apprendre.

Dans ce chapitre, nous nous intéresserons aux capacités d'apprentissage de notre robot. Suite à un bref état de l'art, nous proposerons un modèle d'objet basé sur des indices visuels locaux invariants et capable d'encoder les différents points de vues d'un objet 3D de manière compacte. Après avoir détaillé les méthodes d'extraction des indices, nous présenterons un algorithme de construction en ligne et de reconnaissance des modèles.

5.1 Un modèle d'apparence éparse et multi-vues

5.1.1 Etat de l'art

La littérature dans le domaine de la reconnaissance d'objets en robotique est vaste et propose des systèmes adaptés à tous les types de capteurs potentiellement disponible sur un robot: laser 2D, laser 3D ou banc stéréoscopique, ultra-sons, capteurs haptiques, caméras pour ne citer que les principales. Nous ne traiterons ici que des approches par apparence et plus particulièrement celles par apparences éparses, qui en quelques années se sont imposé imposés dans les systèmes de reconnaissance visuelle d'objets. Suivant les préceptes de [Biederman 87], un objet est représenté par un ensemble de composantes locales caractéristiques, liées entre elle par des considérations géométriques plus ou moins rigide. Ces modèles sont souvent plus robustes aux occultations et aux déformations que

les approches par caractérisation globale du type PCA, moments ou histogrammes.

Le développement de tels modèles a notamment profité des recherches sur les détecteurs de point d'intérêts invariants: [Mikolajczyk 02] utilise par exemple une version multi-résolution du détecteur de Harris pour trouver les points d'intérêt dans l'image et utilise l'opérateur du Laplacien pour sélectionner l'échelle dans l'espace des résolutions. [Lowe 04] détecte les points et sélectionne l'échelle simultanément par recherche d'extrama locaux dans une pyramide de différences de gaussiennes. [Matas 02] a introduit les MSER - *Maximally Stable Extremal Region* - extraites par un algorithme de segmentation de type ligne de partage des eaux. Enfin, [Kadir 01] extrait des régions d'intérêt par analyse de leur entropie. Associés à des descripteurs locaux invariants comme SIFT ou steerable filters (comparaison disponible dans [Mikolajczyk 03]), ces points d'intérêt sont invariants, stables et ont une forte répétabilité vis à vis du changement de point de vue, des conditions d'illuminations et du bruit ambiant.

A partir des détecteurs et descripteurs locaux, de nombreux algorithmes ont été développés pour reconnaître soit des instances soit des classes d'objets. Dans le cadre de la reconnaissance d'instances, les descripteurs locaux ne sont pas suffisamment discrimimant pour pouvoir valider la présence d'un objet uniquement sur la base des appariements par apparences. [Schmid 99] utilisent un modèle probabiliste tenant compte et de l'apparence et de la cohérence spatiale des appariements pour dégager d'un ensemble d'hypothèse d'appariements celle qui correspond le plus à l'image. Suivant un modèle probabiliste similaire, [Moreels 04] propose une solution pour s'affranchir du calcul de l'ensemble des hypothèse possibles. En utilisant un algorithme de recherche de type A* sur l'ensemble des appariements possibles, ils construisent directement la meilleure hypothèse pour la description de l'image. [Lowe 99, Lowe 04] génère un ensemble d'hypothèses entre la base de modèles et l'image en utilisant une transformée générale de Hough et les valident par une vérification probabiliste du taux de faux appariement possible étant donnée une hypothèse et la transformation affine associée. [Lowe 01] contient une extension à la reconnaissance d'objet 3D par regroupements de point de vues. Basé sur [Lowe 99], [Murphy-Chutorian 05] propose une réduction de la taille de la base de modèles en utilisant un vocabulaire de descripteurs. Enfin, [Rothganger 06] et [Brown 05] reconstruisent une modèle 3D complet de l'objet à partir des appariements de descripteurs locaux et en utilisant des contraintes spatiales sur un ensemble de point de vues.

Il y a ensuite les méthodes permettant de faire de la catégorisation d'objets: ici, il est question de trouver un modèle discriminant une classe d'objet générique (voiture, visage, etc) et non un objet particulier. Pour cela, [Fergus 03] utilisent un modèle de constellations flexibles d'indices visuels. L'apparence et les positions relatives du modèle sont estimé

par des probabilités jointes, apprise en utilisant l'algorithme Expectation-Maximisation. Leur approche permet de classifier un ensemble d'images non labelisées sans supervision. [Agarwal 04] construisent d'abord un vocabulaire de descripteurs à partir d'un ensemble d'images représentatives d'une classe d'objets. A partir d'un ensemble d'images décrites par le vocabulaire et labelisées comme appartenant la classe, ils entraînent un classifier qui sera utilise ensuite pour la reconnaissance. [Torralba] utilisent également un vocabulaire de descripteurs mais cette fois, les classes d'objets sont déterminés à partir d'un seul et même vocabulaire. Une modification de l'algorithme de boosting leur permette d'apprendre en même temps le vocabulaire et les classes d'objets.

Nous nous sommes attachés ici à présenter un panel des principales approches en modélisation d'objets par descripteurs locaux, que ce soit pour la reconnaissance pure d'objet ou la catégorisation. La catégorisation nécessitant une grande base d'images pour pouvoir générer un modèle discriminatif, notre système se base sur un modèle de reconnaissance d'instance construit séquentiellement.

5.1.2 Description du modèle et notation

Utilisant un système robotique mobile et actif, il nous est donc possible de "voir" les proto-objets sous plusieurs angles. Les approches par indices visuels locaux, bien que relativement invariantes au changement de point de vue, ne suffisent pas à représenter complètement un objet 3D. Nous nous sommes donc orientés vers des approches multi-vues basées sur les indices visuels et les descripteurs de haut-niveaux. Cherchant un modèle compacte, notre approche s'appuie sur [Lowe 01]: elle consiste à intégrer en un seul modèle d'apparence l'ensemble des images d'apprentissage d'un même objet en groupant les images prises d'un point de vue similaire dans une seule et même vue-clé.

Un modèle $\mathcal{M}$ contient donc un ensemble d'indices visuels $\{f_i\}$ et un ensemble de vues-clé $\{V_j\}$. Chaque V_j est lié au sous-ensemble des f_i détectable depuis cette vue: le lien l_{ij} indique la position, l'orientation et la taille de l'indice f_i dans le référentiel 2D lié à la vue V_j. Le descripteur associé à un indice f étant plus ou moins invariant aux transformations affines et caractéristique d'une partie de l'objet, un indice peut être associé à plusieurs vues-clé. Le modèle contient également l'ensemble des informations statistiques et probabilistes nécessaire à sa reconnaissance, soit :

1. *La probabilité d'observer cet indice visuel dans une image correspondant à une vue-clé du modèle:* elle est estimé à partir du nombre de fois que cet indice a été identifié dans les images d'apprentissage.
2. *Etant donné l'observation de cet indice, la probabilité qu'il soit à cet emplacement:* elle est caractérisée par une distribution de probabilité sur la position, taille et

échelle de cet indice dans le référentiel lié à la vue-clé.

3. *Etant donné l'observation de cet indice, la probabilité qu'il ait cette apparence:* elle aussi est caractérisée par une distribution de probabilité sur les valeurs du vecteur descripteur.

Ayant convenu d'un modèle pour la représentation des proto-objets, nous allons maintenant nous intéresser à l'extraction des indices visuels qui serviront à encoder les apparences locales du proto-objets.

5.2 Extraction des indices visuels

parmi des différentes méthodes d'extractions d'attributs visuels, nous utilisons conjointement le "scale saliency" de [Kadir 01] et les différences de gaussiennes de [Lowe 99].

5.2.1 Scale saliency

Le principe du détecteur proposé par Kadir et Brady est d'utiliser la théorie de l'information et la mesure locale d'entropie pour sélectionner à la fois la position et la taille des régions d'intérêts: En effet, l'entropie d'un signal selon Shannon permet de caractériser la complexité d'un signal. Ainsi, une région de l'image uniforme, donc peu intéressante, aura une faible entropie alors qu'une région fortement texturée aura une forte entropie. En utilisant une fenêtre de taille variable et en la déplaçant sur l'ensemble de l'image, la mesure locale d'entropie permet de détecter les régions d'intérêts de l'image.

Etant donné un point x, son voisinage local R_x et une fonction descriptive D prenant ses valeurs dans $[d_1,\ldots,d_r]$ (par exemple, dans le cas d'une image en niveaux de gris 8 bits, D évolue dans $[0,\ldots,255]$), l'entropie locale est donnée par :

$$\mathcal{H}_D(x,R_x) = -\sum_i p_D(x,R_x,d_i)\log_2 p_D(x,R_x,d_i) \tag{5.1}$$

où $p_D(x,R_x,d_i)$ représente la probabilité qu'a la fonction descriptive D de prendre la valeur d_i sur le voisinage R_x.

En partant de cette équation, la mesure de saillance proposée par Kadir et Brady est une fonction de la position x et de l'échelle s définie comme suit:

$$\mathcal{Y}_D(x,s) = \mathcal{H}_D(x,s)\mathcal{W}_D(x,s) \tag{5.2}$$

où $\mathcal{H}_D(x,s)$ est donné par l'équation 5.1 et s représente le rayon du voisinage circulaire utilisé pour le calcul de la fonction de probabilité p_D. Le poids $\mathcal{W}_D(x,s)$ caractérise l'entropie

entre deux échelles:

$$\mathcal{W}_D(x,s) = \frac{s^2}{2s-1} \sum_i |p_D(x,s,d_i) - p_D(x,s-1,d_i)| \tag{5.3}$$

La détection des régions d'intérêts se fait alors par recherche de maxima locaux dans un espace à 3 dimensions (position,échelle). L'algorithme de détection comporte deux phases: réduction de l'espace de saillance en cherchant pour chaque position l'échelle caractéristique (algorithme 2) et regroupement des régions d'intérêt (algorithme 3)

Algorithme 2 : Calcul de scale saliceny sur une image

Données : image source I, intervalle d'échelles $[s_min,s_max]$

pour tous les *pixel x de l'image I* **faire**
  pour tous les *échelle $s \in [s_min,s_max]$* **faire**
    estimer la fonction de probabilité locale $P_D(s)$ (à l'aide d'histogrammes)
    calculer l'entropie $\mathcal{H}_D(s)$ de $P_D(s)$
    calculer la saillance inter-échelle $\mathcal{W}_D(s)$ entre $P_D(s)$ et $P_D(s-1)$
  fin
  appliquer un filtre de lissage sur $\mathcal{W}_D(s)$
  pour tous les *échelle $\overline{s}$ où $\mathcal{H}_D(s)$ atteint un maximum local* **faire**
    $\mathcal{Y}_D(x,\overline{s}) = \mathcal{H}_D(\overline{s})\mathcal{W}_D(\overline{s})$
  fin
fin

Algorithme 3 : Regroupement des régions d'intérêt

Données : liste $IR = r_i = (x,y,s,y,w)$ de régions d'intérêts
Résultat : nouvelle liste de régions $\overline{IR}$

trier IR suivant les valeurs de saillance y décroissante
appliquer un seuil sur les valeurs de saillance inter-échelle w et de saillance y (optionnelle)
tant que $IR \neq \varnothing$ **faire**
  prendre le premier élément $\overline{r}$ de IR, càd la région la plus saillante
  pour tous les $r \in IR$, $r \neq \overline{r}$ **faire**
    si $\|(x,y)_{\overline{r}} - (x,y)_r\|_2 \leq s_{\overline{r}}$ **alors** supprimer r de IR
  fin
  ajouter $\overline{r}$ à $\overline{IR}$
  supprimer $\overline{r}$ de IR
fin
retourner $\overline{IR}$

5.2.2 Différences de Gaussiennes

L'utilisation des Différences de Gaussiennes pour la détection de points d'intérêts a été introduite par D. Lowe dans [Lowe 99, Lowe 04]. Lowe a montré que la recherche d'extrema locaux, en position comme en échelle, dans une pyramide de différences de Gaussiennes permet de sélectionner des points robustes à une majorité de transformations projectives. Ces travaux font suite à ceux de Lindeberg [Lindeberg 94], dans lesquels il montre que sous un ensemble de contraintes acceptables, le seul noyau permettant de définir un espace multi-résolution est le noyau gaussien. Une résolution d'un tel espace $L(\mathbf{x},\sigma)$ est obtenue par la convolution d'une gausienne $G(\mathbf{x},\sigma)$ et d'une image $I(\mathbf{x})$:

$$L(\mathbf{x},\sigma) = G(\mathbf{x},\sigma) * I(\mathbf{x}) \tag{5.4}$$

$*$ représente l'opérateur de convolution et la gaussienne 2D est donnée par:

$$G(\mathbf{x},\sigma) = G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{5.5}$$

Dans le cadre de l'espace multi-résolution de Lowe, une résolution est définie par la convolution de $I(\mathbf{x})$ par une différence de gaussienne $D(\mathbf{x},\sigma)$, calculée comme la différence de deux résolutions proches séparées par un facteur constant k:

$$\begin{aligned} D(\mathbf{x},\sigma) &= (G(\mathbf{x},k\sigma) - G(\mathbf{x},\sigma)) * I(\mathbf{x}) \\ &= L(\mathbf{x},k\sigma) - L(\mathbf{x},\sigma) \end{aligned} \tag{5.6}$$

La différence de gaussiennes se révèle être une très bonne approximation du laplacien de gaussienne normalisé par la résolution $\sigma^2\nabla^2 G$ utilisé par Linderberg. [Mikolajczyk 02] a montré que les extrema de $\sigma^2\nabla^2 G$ produisaient les points d'intérêts les plus stables en comparaison à un ensemble de détecteurs existants. Pour sélectionner les points, chaque pixel est comparé à ses 8 voisins dans sa résolution et à ses 9 voisins dans les résolutions adjacentes et sont retenus ceux dont la valeur est inférieure ou supérieure à l'ensemble de ses voisins.

5.2.3 Descripteur SIFT

Le descripteur SIFT[1] a été introduit par D. Lowe dans [Lowe 99], puis amélioré dans [Lowe 04].

1. Scale Invariant Feature Transform

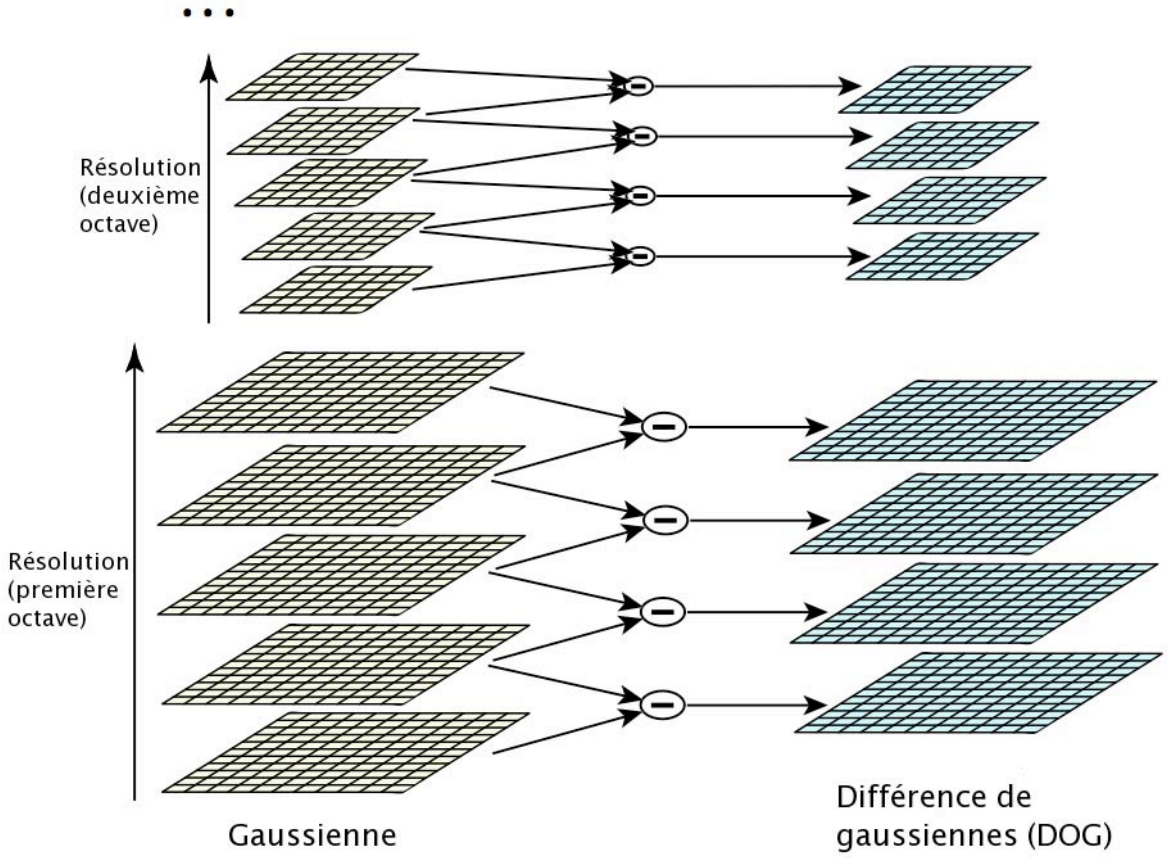

FIG. 5.1 – *Pyramide de différences de gaussiennes*

Région support pour le calcul du descripteur

Lowe calcule directement le descripteur de façon invariante en échelle à partir de la pyramide de différence de gaussiennes. Pour pouvoir calculer le descripteur indépendamment de la méthode de détection des points, nous utilisons une région support normalisée de taille fixe comme dans [Mikolajczyk 05]. Chaque région circulaire, caractérisée par l'échelle du point d'intérêt, est redimensionnée à une taille fixe, ici 41 pixels de diamètre, afin d'avoir une invariance d'échelle. Si la région mesurée est plus grande que la région normalisée, elle est d'abord lissée avant d'être redimensionnée avec un noyau gaussien paramétré avec $\sigma = largeur_{\text{région mesurée}}/largeur_{\text{région normalisée}}$.

Orientation principale

Afin de pouvoir obtenir un descripteur indépendant en rotation, une orientation principale est calculée pour chaque point d'intérêt. Pour chaque pixel $L_{x,y}$ de la région nor-

malisée, sont calculées la magnitude du gradient, m, et son orientation, θ:

$$m = \sqrt{(L_{x+1,y} - L_{x-1,y})^2 + (L_{x,y+1} - L_{x,y-1})^2} \tag{5.7a}$$

$$\theta = \tan^{-1}(\frac{L_{x,y+1} - L_{x,y-1}}{L_{x+1,y} - L_{x-1,y}}) \tag{5.7b}$$

Un histogramme des orientations du gradient est alors calculé à partir des pixels de la région support, chaque contribution étant pondérée par la magnitude de son gradient et par une gaussienne avec σ égale au rayon de la région support. Cet histogramme comporte 36 cellules couvrant les 360° d'orientations possibles.

Les maxima locaux de l'histogramme correspondent aux orientations dominantes de la région d'intérêt. L'orientation principale est donnée par le plus grand maximum local et tout maximum local de valeur supérieure à 80% du maximum principal représente une orientation secondaire. Enfin chaque maximum local est interpolé par une parabole pour affiner la valeur de l'orientation et le point d'intérêt est dupliqué pour chaque orientation secondaire.

Représentation du descripteur

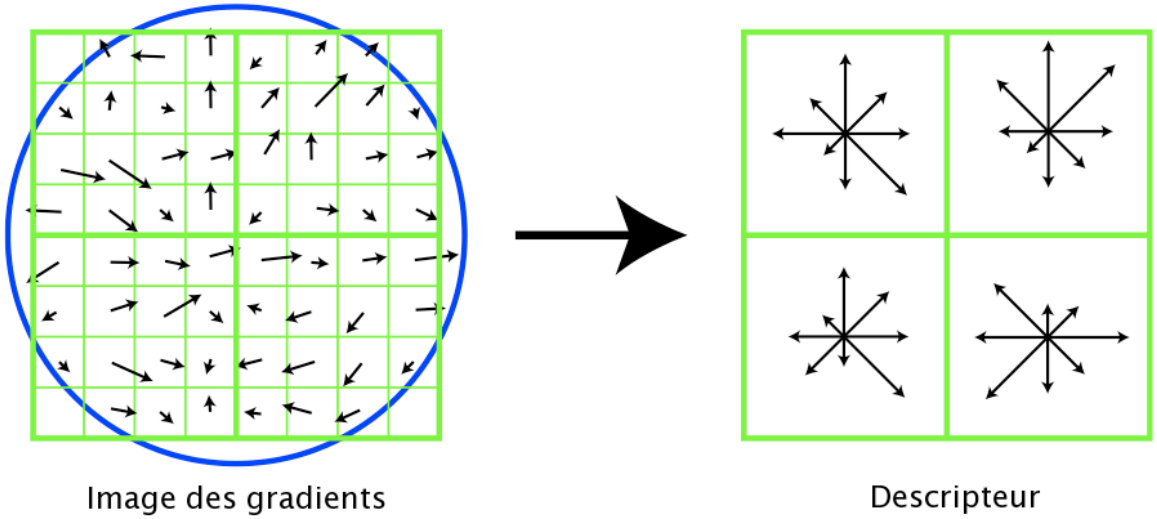

FIG. 5.2 – *descripteur SIFT*

Le descripteur SIFT est lui aussi basé sur les orientations du gradient (cf. figure 5.2). Un ensemble d'histogrammes d'orientations est calculé sur une grille nxn représentant la région support. Chaque pixel d'une cellule de la grille vote pour l'histogramme d'orienta-

tion associé à la cellule à hauteur de la magnitude de son gradient. Une fonction gaussienne avec σ égale à la moitié de la taille de la grille est utilisée pour pondérer chaque vote: les pixels sur les bords de la région support sont plus sensibles aux petites variations de positions de la région que ceux du centre. Ainsi, ils contribuent moins que ceux du centre, rendant le descripteur plus robuste. Enfin, le vecteur obtenu est normalisé pour réduire l'influence de l'illumination. Après une première normalisation à l'unité, le vecteur est seuillé à une valeur maximale de 0.2, puis renormalisé à l'unité. Pour nos expérimentations, nous avons gardé les valeurs de descripteurs proposées par Lowe, c'est-à-dire une grille 4x4 et des histogrammes d'orientations de 8 cellules, donnant au total un descripteur de taille 128.

Les images 5.3 montrent quelques résultats d'extraction d'indices visuels. Nous noterons que l'utilisation conjointe des deux détecteurs permet d'obtenir une description plus complète, l'un étant un détecteur de points et l'autre de régions.

5.3 Construction séquentiel du modèle

Comme nous l'avons vu au paragraphe 5.1.2, notre modèle se présente sous la forme d'un ensemble d'indices visuels regroupés en vues-clé. Le modèle est construit séquentiellement à partir d'une série d'images obtenues par focalisation de la caméra active sur un proto-objet. Chaque image donne lieu à l'extraction d'un ensemble d'indices visuels $\{f_k^I\}$.

5.3.1 Initialisation du modèle associé au proto-objet

Afin d'éviter de créer plusieurs modèles d'un même objet (dont il existerait plusieurs instances dans différents endroits de l'environnement), le traitement de la première image de la séquence donne lieu à une phase de reconnaissance (cf. section 5.4) sur la base de modèles existante. Si le modèle existe déjà, alors il est associé au proto-objet courant et sera utilisé pour la suite du traitement de la séquence. Si non, un nouveau modèle est créé, la première image devenant la première vue-clé.

Une fois la première image de la séquence traitée et le modèle courant initialisé, le processus suit alors quatre étapes, adapté de [Lowe 01].

5.3.2 Mise en correspondance

Cette première étape consiste à mettre en correspondance les indices visuels du modèle avec ceux de l'image courante. Chaque f^I est associé à son plus proche voisin du point

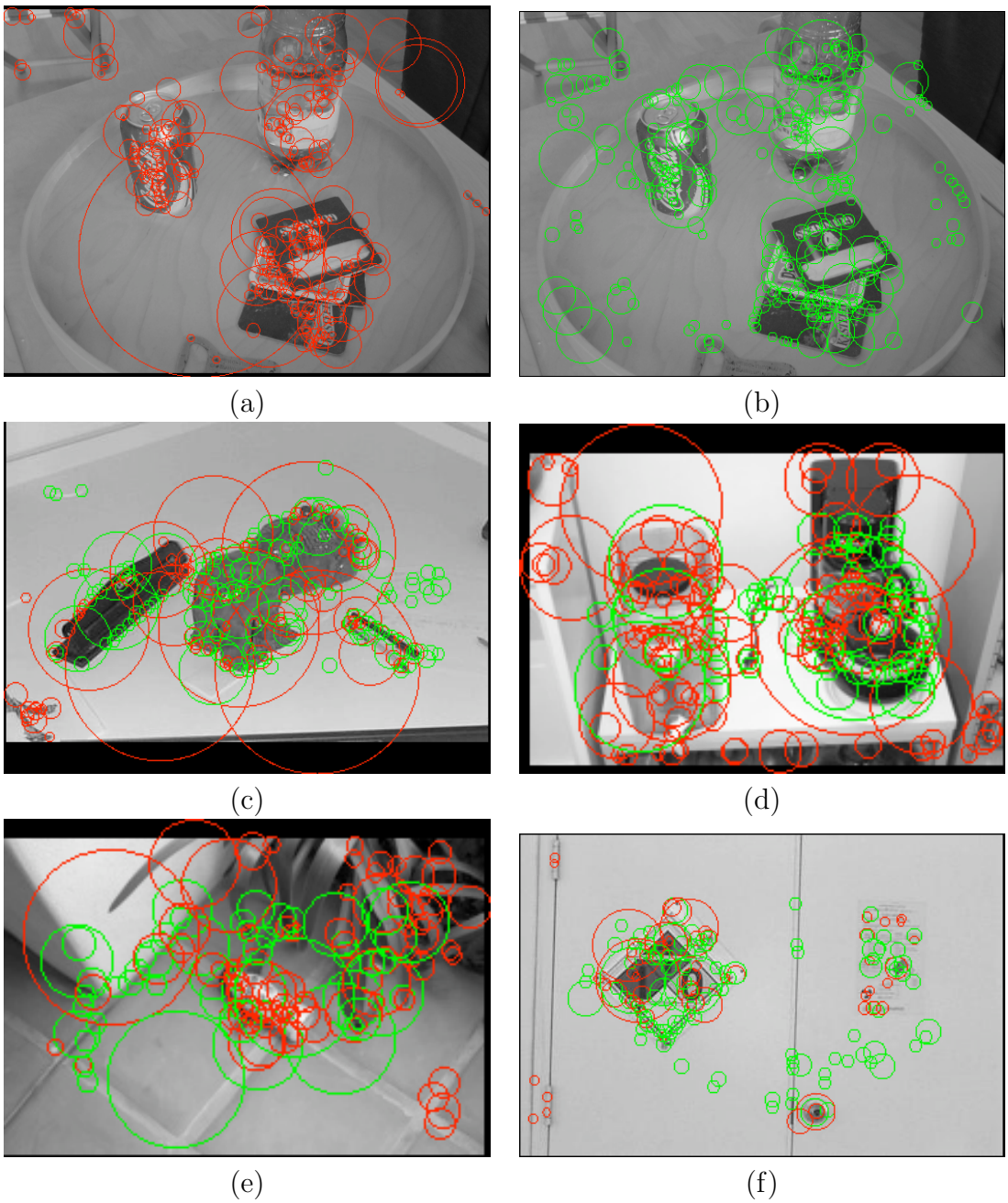

FIG. 5.3 – *(a) Différence de Gaussiennes - (b) Scale Saliency - (c)(d)(e) et (f)* en rouge: *Différence de Gaussiennes.* en vert: *Scale saliency.* NOTA: *les bandes noires ne font pas partie des images originales. Elles sont rajoutées lors de la sauvegarde par notre visualisateur d'image lorsque certains points d'intérêt détectés par les Différences de Gaussiennes sortent de l'image.*

de vue de l'apparence parmi les indices du modèle, noté $f^{\mathcal{M}}$. Le plus proche voisin est déterminé grâce à la distance euclidienne sur le descripteur SIFT. Nous noterons $E = \{e_1,e_2,\dots\}$ l'ensemble des appariements, où $e_k = i$ indiquant que f_k^I est apparié à $f_i^{\mathcal{M}}$.

Bien entendu, un certain nombre de faux appariements peuvent avoir lieu, certains indices de l'image appartenant au fond ou n'ayant pas été détectés dans les images d'apprentissage. Une solution pour éliminer les faux appariements serait d'utiliser un seuil global sur la distance euclidienne; néanmoins cette méthode est peu efficace, certains descripteurs étant plus discriminants que d'autres. [Lowe 99] a proposé une méthode de validation des appariements à partir du ratio entre la distance au plus proche voisin et la distance au second plus proche voisin. Nous éliminons donc tout appariement dont le ratio est supérieur à 0.8.

Enfin, pour accélérer la recherche du plus proche voisin sur un espace de grandes dimensions (ici, la dimension du vecteur SIFT), celle-ci est implémentée à l'aide d'un kd-tree. (Un kd-tree est une structure de données en arbre binaire qui partitionne récursivement un espace de dimension k à la moyenne de la dimension ayant la plus grande variance. La complexité de la recherche du plus proche voisin est alors $O(\log N)$)

5.3.3 Recherche de la vue-clé la plus proche

Une fois la mise en correspondance par apparence effectuée, il nous faut maintenant trouver la vue-clé qui correspond le plus au point de l'image actuellement traitée et éliminer les appariements inconsistants avec cette vue-clé. Pour cela, nous utilisons une Transformation Généralisée de Hough (TGH) [Ballard 81, Grimson 90] sur l'espace discret des similitudes 2D (translation 2D, rotation, échelle) et des vues-clé. Chaque appariement e_k fournit les paramètres d'une similitude pour chaque vue-clé associée à $f_{e_k}^{\mathcal{M}}$ et vote donc pour autant d'entrées dans l'accumulateur de Hough. Pour éviter les effets de bords dus à la discretisation de l'espace de transformation, chaque appariement vote pour les 2 entrées les plus proches de chaque dimension de la similitude. L'espace de transformation est discretisé avec un pas de 1/8 de la taille de l'image en translation, $\pi/8$ *rad* en rotation et une octave en échelle.

Le résultat de la transformation nous fournit alors une hypothèse h_j sur la vue-clé V_j la plus proche de l'image courante, une estimée grossière de la similitude T_j entre la vue-clé et l'image courante ainsi que le sous-ensemble de appariements associé $\{e_k\}_{V_j,T_j}$.

5.3.4 Vérification géométrique

Avant de pouvoir intégrer les informations de la nouvelle image, nous effectuons une vérification géométrique de la consistance de l'appariement obtenu précédemment. A ce stade, à cause des approximations et autres erreurs imputées à l'utilisation de la TGH, l'ensemble d'appariements peut contenir des paires incompatibles. Nous cherchons donc une évaluation plus précise de la transformation T_j par résolution aux moindre d'un système linéaire, si le nombre de appariements d'indices visuels est suffisant.

La similitude liant un point du modèle [x,y] à un point de l'image [u,v] est donné par l'équation suivante:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} s\ \cos\theta & -s\ \sin\theta \\ s\ \sin\theta & s\ \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{5.8}$$

avec t_x,t_y la translation de l'image, θ la rotation de l'image et s l'échelle de l'image. En définissant $m = s\ \cos\theta$ et $n = s\ \sin\theta$, nous obtenons:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} m & -n \\ n & m \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{5.9}$$

En extrayant les paramètres de la similitude, nous obtenons un système linéaire de type $\mathbf{Ax} = \mathbf{b}$ à partir de l'ensemble des appariements, chaque appariement donnant à deux équations comme indiqué dans 5.10:

$$\begin{bmatrix} x_1 & -y_1 & 1 & 0 \\ y_1 & x_1 & 0 & 1 \\ x_2 & -y_2 & 1 & 0 \\ y_2 & x_2 & 0 & 1 \\ \vdots & & & \end{bmatrix} \begin{bmatrix} m \\ n \\ t_x \\ t_y \end{bmatrix} = \begin{bmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ \vdots \end{bmatrix} \tag{5.10}$$

La résolution aux moindres carrés de ce système est donnée par

$$\mathbf{x} = [\mathbf{A^T A}]^{-1}\mathbf{A^T b} \tag{5.11}$$

qui minimise la somme des distances au carré entre la projection d'un point du modèle et son point-image associé. Nous pouvons maintenant éliminer les appariements qui s'accordent le moins avec la similitude trouvée, en ne gardant que celles dont l'erreur de projection est inférieure à $\sqrt{2}\mathcal{E}$, $\mathcal{E}$ étant l'erreur de projection moyenne. De plus, nous rajoutons les appariements exclus à l'étape précédente qui s'accordent avec la similitude

trouvée. Si le nombre de appariements ainsi obtenus est suffisant, nous recalculons la similitude.

5.3.5 Mise à jour du modèle

Maintenant que nous avons un ensemble d'appariements consistants et une évaluation de la transformation, nous pouvons mettre à jour notre modèle $\mathcal{M}$. La décision de mise-à-jour est prise suivant la qualité de la mise en correspondance et l'erreur moyenne e de la transformation entre la vue et l'image. Nous considérons 3 cas:

1. Si le nombre de appariements à l'issue d'une des 3 étapes précédentes est insuffisant (dans notre cas, inférieur à 10), nous considérons que le point de vue de l'image courante est nouveau. Nous intégrons donc les indices visuels au modèle et les lions à une nouvelle vue-clé.
2. Si l'erreur moyenne $\mathcal{E}$ est supérieur à un seuil T, alors nous considérons qu'il existe bien une correspondance entre la vue-clé choisie et l'image courante mais que la différence de point de vue est trop importante pour pouvoir intégrer les nouvelles informations. Une nouvelle vue-clé est donc créée; néanmoins, les indices visuels f_k^I appairés ne sont pas ajoutés au modèle, seul un lien entre l'indice $f_{e_k}^{\mathcal{M}}$ correspondant et la nouvelle-vue clé étant instancié.
3. Enfin, si l'erreur moyenne $\mathcal{E}$ est acceptable, alors les informations de l'image courante sont fusionnées avec la vue-clé sélectionnée. Chaque f_k^I est tout d'abord projeté dans le referentiel de la vue-clé V_j. Si f_k^I a été apparié, alors sa projection sert à mettre à jour le lien $l_{e_k j}$. Sinon f_k^I est intégré au modèle et lié à la vue-clé.

Nous obtenons ainsi un modèle compact de l'objet, intégrant les information des ses différents points de vues, comme en témoigne la figure 5.4.

5.4 Système de reconnaissance

Notre système de reconnaissance s'appuie sur un système probabiliste de génération d'hypothèses et de vérification inspiré de [Schmid 99, Pope 00, Moreels 04].

5.4.1 Principe du système

Pour reconnaître un objet, nous cherchons à déterminer si une des vues-clé de l'objet appartient à l'image de test. Soit $\mathcal{M}$ le modèle à reconnaître, h une hypothèse qu'une des vues-clé est dans l'image test et $\mathcal{O}_h$ les indices visuels observés dans l'image test associée

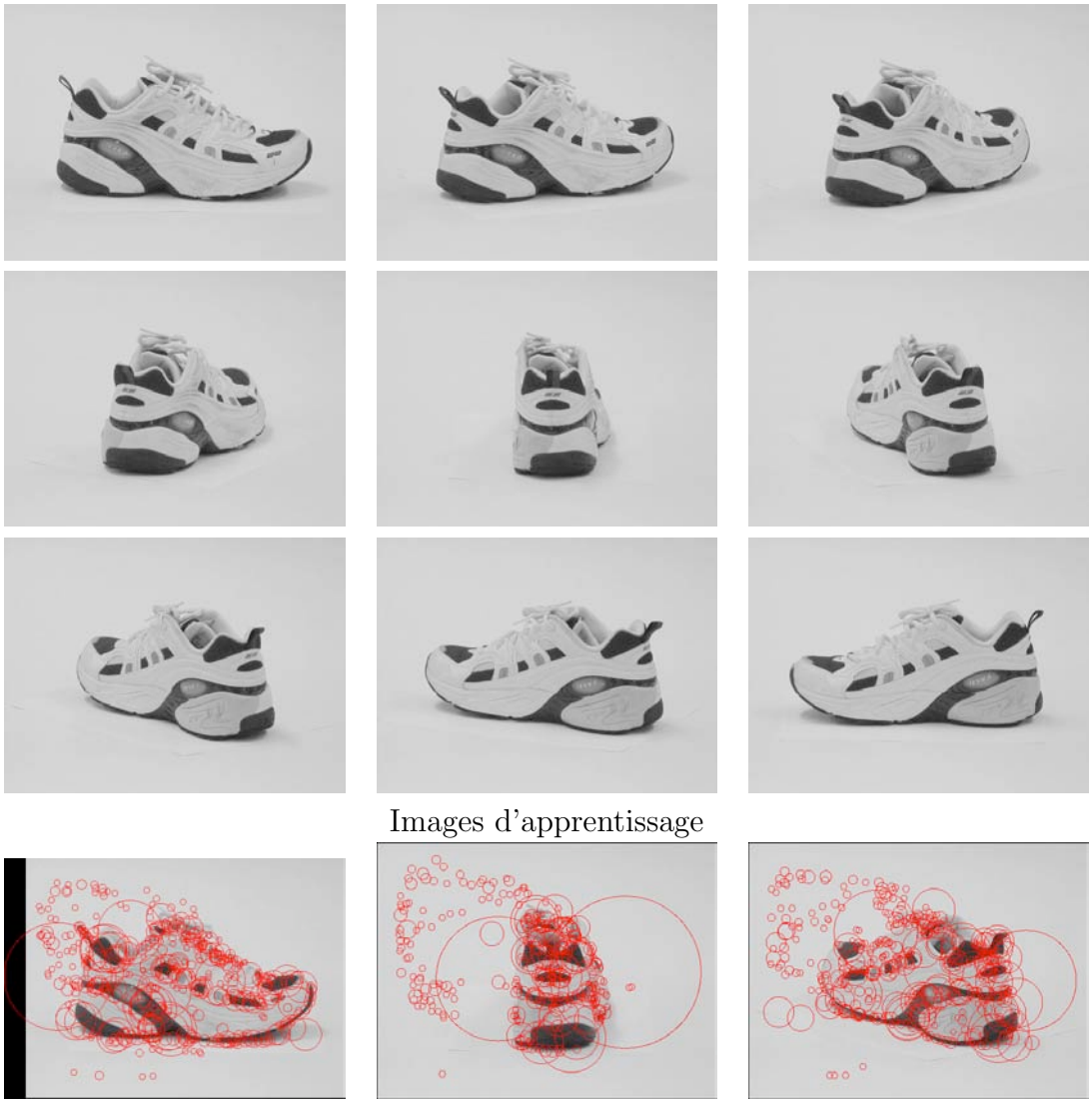

FIG. 5.4 – *Résultat de notre apprentissage d'un modèle de chaussure*

à h. Nous définissons la confiance de l'hypothèse comme $v(h) = p(h|\mathcal{O}_h,\mathcal{M})$. A partir de la règle de Bayes, nous pouvons l'écrire comme:

$$v(h) = p(h|\mathcal{O}_h,\mathcal{M}) = \frac{p(\mathcal{O}_h|h,\mathcal{M})p(h|\mathcal{M})}{p(\mathcal{O}_h|\mathcal{M})} \tag{5.12}$$

L'existence de la vue-clé la plus probable est donc celle associée à l'hypothèse h qui maximise cette confiance. Ainsi nous avons:

$$\hat{h} = \underset{h\in\mathcal{H}}{argmax} \left(\frac{p(\mathcal{O}_h|h,\mathcal{M})p(h|\mathcal{M})}{p(\mathcal{O}_h|\mathcal{M})} \right) \tag{5.13}$$

où $\mathcal{H}$ représentent un ensemble d'hypothèses.

Afin de simplifier l'écriture et le calcul de la confiance d'une hypothèse, nous supposerons que:

- Les probabilités a priori sont uniformes. Nous supprimons donc les termes $p(h|\mathcal{M})$ et $p(\mathcal{O}_h|\mathcal{M})$ de l'équation 5.12
- Etant exprimés dans le même référentiel lié à la vue-clé sous-jacente à l'hypothèse, les indices sont considérés comme indépendants
- Apparence et position d'un indice sont indépendantes.

Nous obtenons donc que la probabilité que la détection des indices du modèle se produirait avec la position et l'apparence spécifiée par $\mathcal{O}_h$ s'ecrit:

$$p(\mathcal{O}_h|h,\mathcal{M}) = \prod_{f \in \mathcal{O}_h} p_{app}(f|h,\mathcal{M}) p_{pos}(f|h,\mathcal{M}) \tag{5.14}$$

L'estimation des densités de probabilités p_app et p_pos sera discutée au paragraphe 5.4.2 Néanmoins, trouver la meilleure des hypothèses ne suffit pas à déterminer si l'objet est présent ou non. La décision est donnée par la règle:

$$R = \frac{p(\mathcal{O}_{\hat{h}}|\hat{h},\mathcal{M})}{p(\mathcal{O}_{\hat{h}}|\hat{h},Bg)} \tag{5.15}$$

Le dénominateur correspond à la probabilité d'avoir l'hypothèse $\hat{h}$ considérant le fond de l'image comme objet. En supposant que chaque point de l'image est susceptible de donné lieu à un indice visuel considéré comme le fond de manière équiprobable, l'estimation de cette probabilité devient:

$$p(\mathcal{O}_{\hat{h}}|\hat{h},Bg) = \left[\frac{1}{A}.\frac{1}{2\pi}\right]^n \prod_{f \in \mathcal{O}_{\hat{h}}} p_{Bg}(f|\hat{h},Bg) \tag{5.16}$$

avec A la taille en pixel de la projection de la vue-clé et n le nombre d'indices de $\mathcal{O}_{\hat{h}}$.

La reconnaissance utilise les mêmes fonctions pour la construction du modèle et se déroule comme suit:

1. Mise en correspondance des indices visuels.
2. Génération d'un ensemble d'hypothèses $\mathcal{H}$ par TGH. Les hypothèses correspondent aux maxima locaux de l'accumulateur du TGH contenant plus de 5 appariements.
3. Affinement des hypothèses par vérification géométrique. Tout hypothèse comptant moins de 5 appariements est alors invalidée.
4. Evaluation de la meilleure hypothèse $\hat{h}$

5.4.2 Densité de probabilité d'apparence et de position

Pour l'estimation des densités de probabilité, nous avons utilisé la même approche que [Moreels 04].

Apparence d'un indice associé à l'objet

La densité d'apparence décrit à quel point l'indice de l'image ressemble à l'indice du modèle: elle est modélisée par une gaussienne de covariance Σ_{app} comme suit:

$$p_{app}(f|h,\mathcal{M}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_{app}|} e^{-\frac{1}{2}(f_{app}-f_{app}^{\mathcal{M}})^T \Sigma_{app}^{-1}(f_{app}-f_{app}^{\mathcal{M}})} \tag{5.17}$$

avec f^M l'indice du modèle associé à f par l'hypothèse h et d la taille du descripteur. La matrice de covariance Σ_{app} est estimée à partir des différences d'apparences entre les indices du modèle appariés par plus proche voisin.

Position d'un indice associé à l'objet

Les positions des indices associés à l'objet sont supposées être cohérentes avec les positions des indices du modèle associés: la transformation T_h associée à l'hypothèse h permet d'obtenir la position de chaque indice du modèle dans le référentiel de l'image. Ainsi nous obtenons :

$$p_{pos}(f|h,\mathcal{M}) = G_{loc}(f|T_h(f^{\mathcal{M}}))G_{\theta}(f|T_h(f^{\mathcal{M}}))G_s(f|T_h(f^{\mathcal{M}})) \tag{5.18}$$

où G_{loc}, G_θ et G_s sont des densités de probabilité gaussiennes pour respectivement la position, l'orientation et l'échelle. Les paramètres de covariance de ces gaussiennes sont pour l'instant fixé manuellement à 20 pixels pour la position, une demie-octave en taille sur une échelle logarithmique et et $\frac{\pi}{3}$ en orientation.

Apparence d'un indice associé au fond

Lors de la reconnaissance, tout ce que n'est pas l'objet dans l'image forme le fond. En général, celui-ci est composé d'une multitude d'autres objets que celui qui nous intéresse. L'apparence d'un indice associé au fond ressemble donc à n'importe quel indice extrait de n'importe quel objet. Nous modélisons donc cette densité d'apparence par une densité

gaussienne de moyenne μ_{Bg} et de matrice de covariance Σ_{Bg}:

$$p_{Bg}(f|h,Bg) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_{Bg}|} e^{-\frac{1}{2}(f_{app}-\mu_{Bg})^T \Sigma_{Bg}^{-1}(f_{app}-\mu_{Bg})} \tag{5.19}$$

Pour l'estimation des paramètres μ_{Bg} et Σ_{Bg}, nous avons utiliser 200 images issues de la série "background" de la base d'images du Caltech Computational Vision group (disponible sur http://www.vision.caltech.edu/archive.html). Pour chaque image, les deux détecteurs fournissaient entre 400 et 700 indices visuels.

5.4.3 Résulats expérimentaux

Nous avons appliqué notre algorithme de modélisation sur les objets de la base "Objet Recognition Database" du Ponce group, Computer Vision and Robotics, University of Illinois-Urbana-Champaign (disponible sur http://www-cvr.ai.uiuc.edu/ponce_grp/data/): cette base est consitutée d'images d'apprentissages de 8 objets pris sous différents points de vues et de 51 images de tests. Le seuil T pour la décision de mise-à-jour a été fixé à 0.05 fois la plus grande des dimensions des images d'apprentissage. Le tableau 5.1 résume pour chaque objet les données d'apprentissages et les données du modèle obtenu.

	pomme	*camion*	*teddybear*	*venom*	*pavement*	*chaussure*	*sel*	*pot*
	Apprentissage							
♯ images	29	28	20	16	16	16	16	20
♯ indices	5551	7169	25782	9654	38266	8410	8219	12031
	Modèle							
♯ vues	6	3	4	6	9	7	1	3
♯ indices	1093	792	4792	3574	21043	3527	555	1655

TAB. 5.1 – *Résultats de la construction des modèles sur la base "Object Recognition Database"*

En n'utilisant que la génération d'hypothèse par TGH et sélection de la meilleur hypothèse disponible, nous obtenons dans la majorité des cas la bonne location de l'objet, comme le montre les figure 5.5 et 5.6 .

Nous avons quelques résultats sur des modèles appris avec un fond non uniforme. Cependant, la capacité du système à générer des hypothèses valides est fortement diminuer.

Dans l'ensemble, notre système de reconnaissance est moins performants que les solutions existantes présentées dans notre bibliographie, avec un taux de reconnaissance

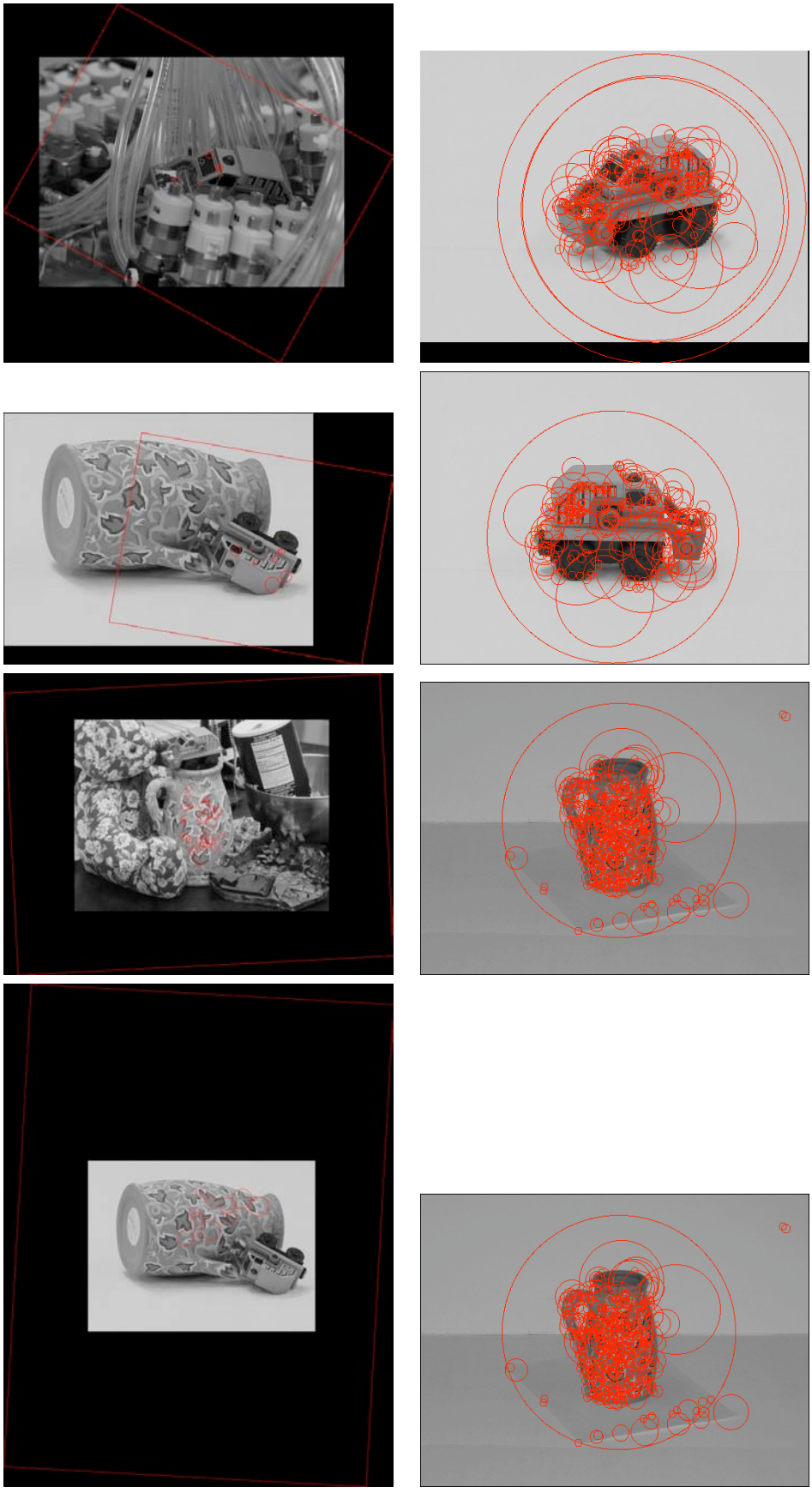

FIG. 5.5 – à droite: *image de test.* à gauche: *vue-clé du modèle apparié*

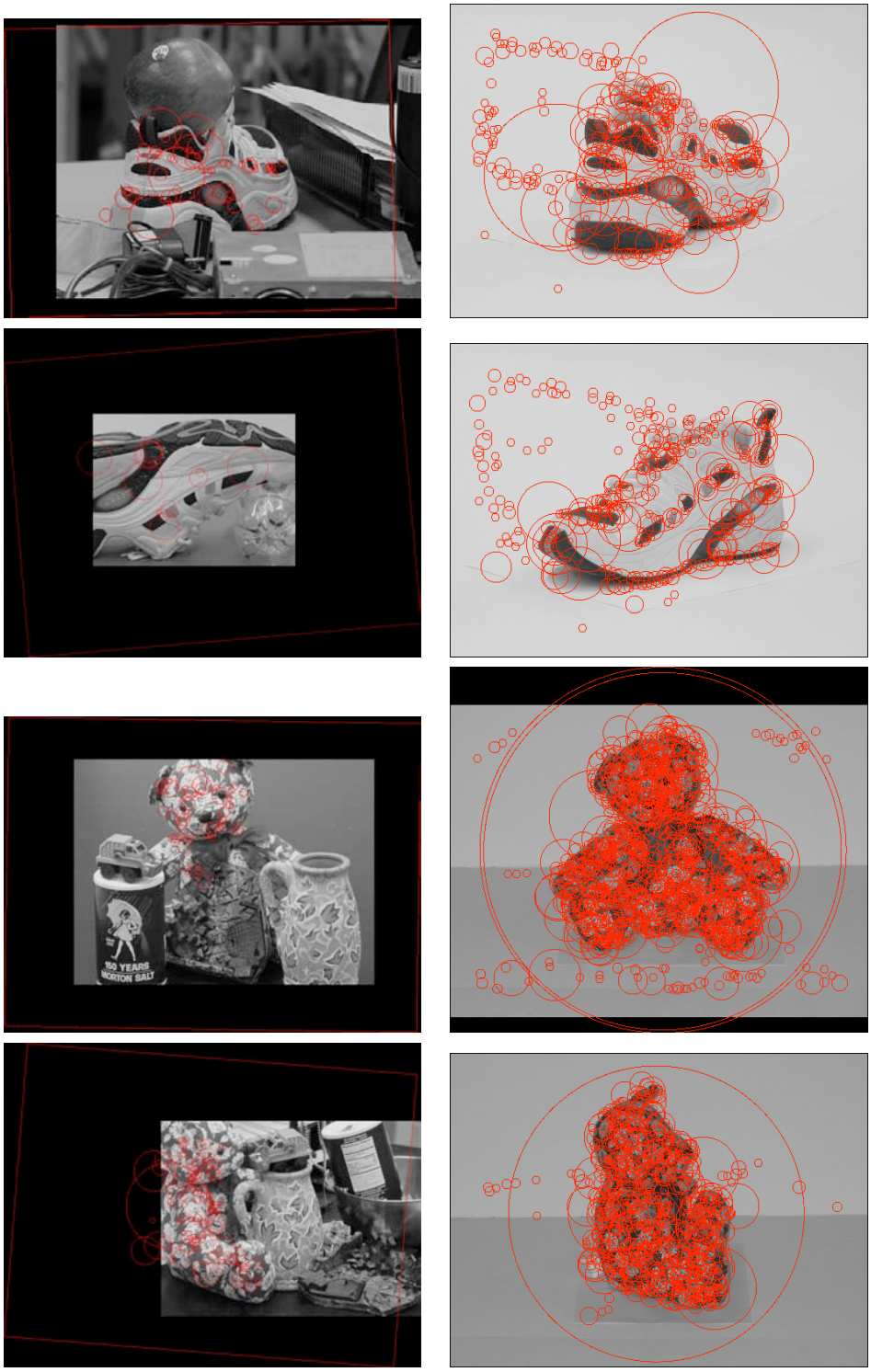

FIG. 5.6 – à droite: *image de test.* à gauche: *vue-clé du modèle apparié*

moyen de seulement 50 %. Nous avons également remarqué une très grande variabilité du taux de reconnaissance en fonction du modèle appris (100 % dans le cas du Teddy Bear, 0 % dans le cas de la pomme).

Notre code n'étant nullement optimisé, nous espérons gagner en performance avec une réécriture complète. Cependant, il nous semble indispensable de revoir également en détail le formalisme de notre modèle probabiliste.

5.5 Conclusion

Nous avons présenté dans ce chapitre du deuxième processus de notre système dédié à la modélisation et à la reconnaissance des proto-objets de l'environnement. Nous avons décrit dans ce chapitre un modèle compacte d'apparences par intégration de vues et partage d'indices locaux et proposer une méthode de construction séquentielle de ce modèle ainsi d'un système de validation probabiliste.

Nous remarquons néanmoins que la qualité de la reconnaissance est grandement perturbé par la présence de descripteurs n'appartenant pas à l'objet. Nous essaierons de filtrer par la suite les indices de fond.

Chapitre 6

Conclusion

Dans ce manuscrit, nous nous sommes intéressés à l'évolution d'un robot compagnon dans un environnement humain et notamment à ses capacités d'extraire de cette environnement des informations de haut niveaux susceptibles d'aider à l'interaction avec un opérateur humain.

Nous avons en premier lieu porté notre réflexion sur la représentation spatiale d'un environnement de type appartement dans lequel doit évoluer un robot personnel. Nous avons proposé une représentation sous la forme d'un graphe des objets que le robot découvre durant la phase d'exploration. De fait, nous n'avons pas construit un tel graphe, car nous avons focalisé nos travaux sur la détection autonome des objets d'intérêt dans l'environnement, puis sur la construction de la représentation de ces objets. Par contre, dans le cadre du projet COGNIRON, S.Vasudevan a plutôt travaillé sur la construction d'un tel graphe, une fois que le robot connaît les modèles des objets qu'il est susceptible de trouver. Il est cependant difficile d'évaluer si une telle représentation permettra d'améliorer les capacités de réflexion et l'autonomie d'actions du système robotique dans ses différentes tâches.

Nous avons ensuite proposé une procédure d'exploration active, en deux phases exécutées de manière asynchrone par le robot, équipé d'un système visuel multi-focal. Le robot exploite sa vision grand champ (ou omnidirectionnelle) pour la détection, le suivi et la localisation grossière de régions d'intérêt saillantes: nous avons exploité l'approche de type *Carte de saillance*, proposée initialement par L.Itti, approche inspirée de ce que l'on connaît de la Vision humaine. Les résultats de nos tests montrent que notre algorithme peut détecter des régions caractérisant un objet dans le sens humain du terme. Néanmoins, il souffre d'un ratio plutôt faible entre le nombre de régions significatives et le nombres total de régions extraites.

Une fois détectée et localisée par le processus pré-attentif, la position et la taille ap-

proximative d'une région d'intérêt sont mémorisées dans une mémoire courte durée. Le processus attentif analyse successivement chacune de ces régions, par focalisation d'un capteur visuel actif sur la zone de l'espace où elle se trouve. Par rapport aux nombreux travaux existants sur l'apprentissage de modèles d'apparence d'objets 3D, modèles exploitables pour la reconnaissance, nous avons souhaité proposer une approche exécutable en ligne par un robot, donc sans utiliser les artifices usuels comme l'objet sur un fond uniforme, monté sur une platine en azimut tournant devant la caméra , ou encore l'objet déplacé par l'opérateur devant la caméra. Mais notre approche est-elle réaliste?

Cette brève discussion sur la portée de nos résultats, nous amène à proposer plusieurs directions possibles pour des travaux futurs.

◇ Nous avons vu que les approches de type *Carte de Saillance* sont très sensibles aux problèmes d'illumination dans l'environnement, et très dépendantes de la paramétrisation. Il serait intéressant de rendre plus robuste et plus stable ce niveau pré-attentif, en ajoutant une couche d'auto-apprentissage pour que le robot évalue la qualité des images acquises, afin d'adapter les valeurs des poids donnés à chaque carte caractéristique (couleur, intensité...). En particulier, l'éclairage entrant par les fenêtres peut biaiser ce processus: comment détecter et inhiber les zones saturées?

Par ailleurs, nous avons dit que l'Homme exploitait pour optimiser cette tâche d'exploration d'un environnement inconnu, une substrat très important de connaissances a priori: par exemple, il reconnaîtra de suite les portes nécessaires pour changer de lieu topologique. Comment prendre en compte des informations contextuelles pendant la phase d'exploration? Des probabilités a priori de découvrir une région d'intérêt dans telle zone de l'image? Par exemple, si le robot connait le concept de *Table*, il pourra focaliser sa recherche de régions saillantes sur la zone de l'image correspondant au dessus de la table ...

◇ Concernant le processus attentif, il conviendrait d'abord, de mieux valider notre approche d'apprentissage du modèle d'apparence d'un objet 3D, par une caméra mobile, asservie en orientation et zoom. Même si la focalisation sur la zone contenant l'objet, facilite la segmentation entre primitives sur l'objet et primitives du fond, nous avons plusieurs fois constaté que des points appartenant au fond sont très souvent indûment intégrés dans le modèle d'un objet. Nous devrons améliorer les tests de validation qui permettent de filtrer ces points.

Un des problèmes-clé en ce domaine, est la catégorisation: le concept d'objet générique doit ressortir de la phase d'exploration. Il existe aujourd'hui de nombreuses méthodes de catégorisation visuelle basées sur des descripteurs locaux invariants identiques à ceux que

nous avons utilisés. Il parait donc envisageable d'utiliser nos modèles pour soit générer les modèles de classes soit pour les généraliser.

◇ Il conviendrait de revenir sur l'ensemble des représentations spatiales nécessaires pour la navigation d'un robot mobile dans un environnement humain tel qu'un appartement. Nous avons vu que le projet COGNIRON propose une représentation hiérarchique, qui part des données capteurs pour atteindre le niveau sémantique. Plusieurs instances d'une telle représentation hiérarchique ont été proposées: (1) modèle topologique construit à partir d'une base d'images omnidirectionnelles, (2) modèle topologique construit à partir d'une carte stochastique de segments-laser, et (3) modèle qualitatif de type *Graphe d'Objets* construit à partir d'un ensemble d'objets connus a priori ou découverts lors de l'exploration.

Dans chaque cas, le niveau sémantique est obtenu grâce au nommage des lieux, et pour la troisième approche, des objets que le robot sait reconnaître.

Notre approche est de type objet. Cette représentation répond-elle aux besoins de la navigation? Dans le cas (1), la navigation est traitée par asservissement visuel pour parcourir une trajectoire définie dans une base d'images; dans le cas (2), c'est une navigation classique, par génération d'une trajectoire dans l'espace libre, puis exécution en se localisant sur les segments-laser de la carte; dans le cas (3), la navigation n'est pas encore traitée. Elle devrait aussi être réalisée par asservissement visuel 2D ou 3D pour parcourir une trajectoire définie dans le graphe des objets. Même si cela a priori ne pose pas de difficultés, il conviendrait néanmoins de valider la pertinence d'une représentation de type Graphe d'Objets pour la navigation. Comment enchaîner les asservissements successifs sur les objets? Comment éviter des obstacles, prendre en compte des occultations...? Surtout comment mettre à jour le graphe si un objet a disparu, ou a été déplacé?

◇ Nous pensons enfin qu'il serait intéressant d'analyser les liens antre l'approche procédurale que nous avons proposée, et les approches plus inspirées des neurosciences, qui exploitent des techniques neuronales. Citons en particulier les travaux de doctorat de N.Dohuu [Huu 05], réalisés également dans le cadre du projet COGNIRON au LAAS-CNRS, en parallèle avec les nôtres.

Annexe A

Présentation du robot-guide Rackham

voici une courte présentation du démonstrateur utilisé dans le cadre de l'exposition temporaire "Mission Biospace" qui s'est tenue à la Cité de l'Espace à Toulouse en 2004-2005.

Le démonstrateur: Rackham (figure A.1) est un robot mobile de type colonne B21R de chez iRobot de 118 cm de haut et 58 cm de diamètre. Il embarque 2 ordinateurs basés sur un mono-processeur et un bi-processeur P3 1GHz, un télémètre laser à balayage SICK, 24 capteurs ultra-sons répartis en deux rangés en haut et en bas du corps, une caméra active SONY EVI-D70, une caméra firewire montée sur une platine PAN/TILT, un écran tactile et une paire de haut-parleurs.

Architecture logicielle: le démonstrateur est régi par une instance de l'architecture logicielle LAAS[1] présentée dans la figure A.2: il s'agit d'un système hiérarchique à deux niveaux comprenant un superviseur procédural écrit avec openPRS[2], contrôlant un ensemble distribué de modules fonctionnels.

Chaque module, généré par le générateur de modules GenoM[2], est un composant logiciel indépendant qui contient un ensemble de fonctions plus ou moins complexes et contraintes au niveau temps d'exécution, permettant la réalisation d'un service nécessaire au fonctionnement du démonstrateur. Nous distinguerons deux niveaux de services:

– bas niveau: les services concernant le contrôles des capteurs (télémètre, caméras, ...)

1. LAAS architecture for Autonomous System
2. L'ensemble des outils nécessaire à la mise en place d'une architecture LAAS (GenoM, OpenPRS, Pocolib,*etc*) sont disponible sur http://softs.laas.fr/openrobots

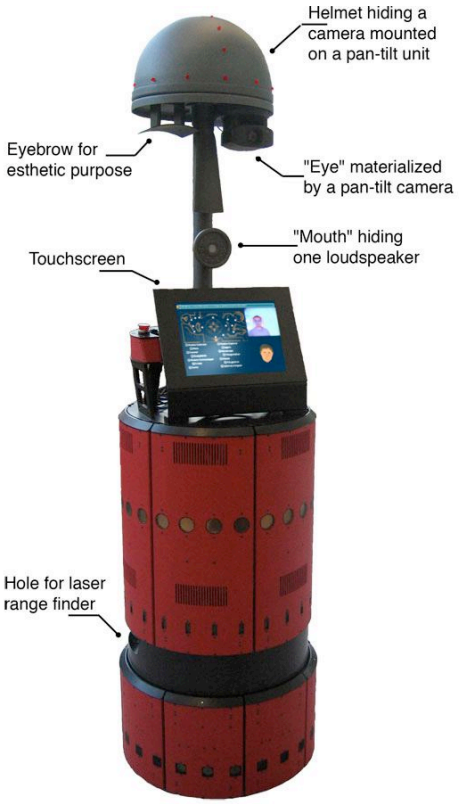

FIG. A.1 – *le démonstrateur Rackham*

et des effecteurs (contrôle des roues, de la platine PAN/TILT, ...)

– haut niveau: les services s'occupant du traitement de l'information comme SEGLOC pour la localisation sur segments laser ou NDD pour l'exécution dynamique de trajectoire.

Les modules sont régis par un comportement standard et contiennent un ensemble d'interfaces entrées/sorties permettant la communication entres modules ou avec le superviseur.

Ce ne sont pas moins de seize modules qui fonctionnent de manière asynchrones et distribués sur les deux ordinateurs qui équipent le robot Rackham. Au travers des ces modules et du superviseurs, Rackham est capable de se localiser et de naviguer dans un environnement 2D connu, de détecter, apprendre et reconnaître des visages grâce à sa caméra active, de dialoguer avec un utilisateur par reconnaissance/synthèse vocale, etc. L'écran tactile permet également à l'utilisateur d'avoir accès à un certain nombre d'informations complémentaires et de donner des ordres au robot.

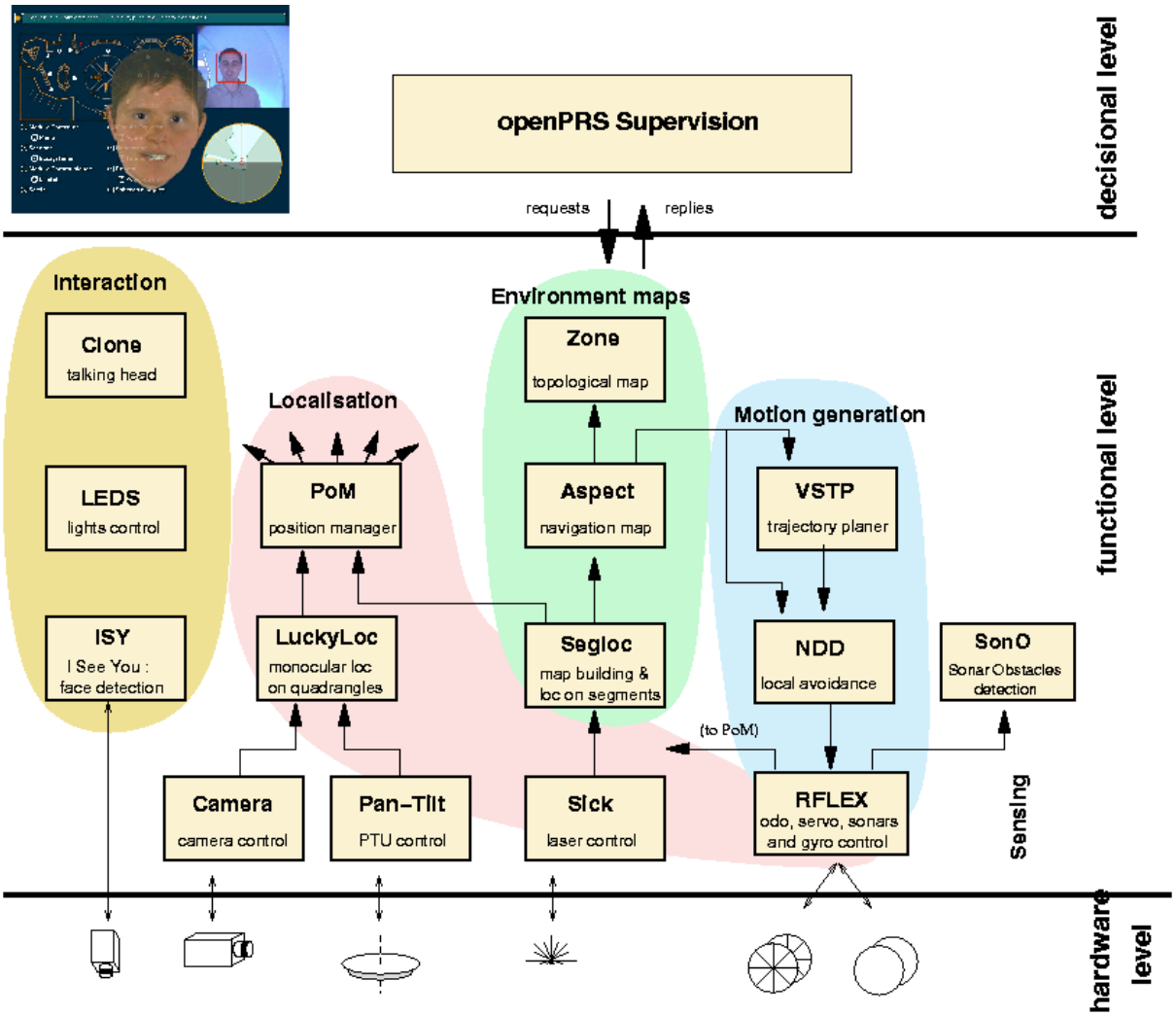

FIG. A.2 – *Architecture logicielle LAAS embarquée dans Rackham*

Annexe B

JAFAR

Je suis l'un des principaux développeurs de *JAFAR* (JAFAR is A Framework for Algorithms [development] in Robotics): il s'agit d'un environnement de développements C/C++ intéractif actuellement utilisé dans les groupes robotiques du LAAS.

Motivations: Ce type d'outil est essentiel dans une équipe de recherche. De nombreuses personnes développent des logiciels pour tester et démontrer leurs algorithmes. *JAFAR* sert de base au travail collaboratif de différents chercheurs dans la production de logiciels pour robots autonomomes. Il s'agit d'un environnement de développement proposant un ensemble de structures de données standards et des algorithmes principalement axés traitements d'images et robotique. Il est fourni avec une documentation complète et chaque développeur peut facilement intégrer de la documentation dans ses propres modules.

JAFAR succède à une longue lignée d'environnements de développement en vision, *VIZIR* (ViZion In Robotics, 1981-1986), issue du projet ARA, et *CALIFE* (1983-2005).

Présentation: Les libraires sont développés en C/C++ et sont automatiquement disponibles dans un shell interactif *ala* Matlab. Actuellement, les langages dynamiques supportés sont Tcl/Tk (`http://www.tcl.tk/`) and Ruby (`http://www.ruby-lang.org`). Cette approche en deux couches, adoptée dans *JAFAR* et reprise de son prédécesseur *CALIFE* permet de créer des bibliothèques d'algorithmes en C/C++ et de tester rapidement et interactivement leurs fonctionnalités. Le lien entre les librairies et le shell interactif est généré par Swig (`http://www.swig.org`). La documentation est quant à elle entièrement générée avec Doxygen (`http://www.doxygen.org`), qui permet une mise en page claire sous plusieurs formats (html, pdf, etc), l'utilisation de liens hypertextes et fournit, dans sa sortie html, une intéressante fonction de recherche.

La figure B.1 montre la structure d'un module *JAFAR* et les différents mécanismes de

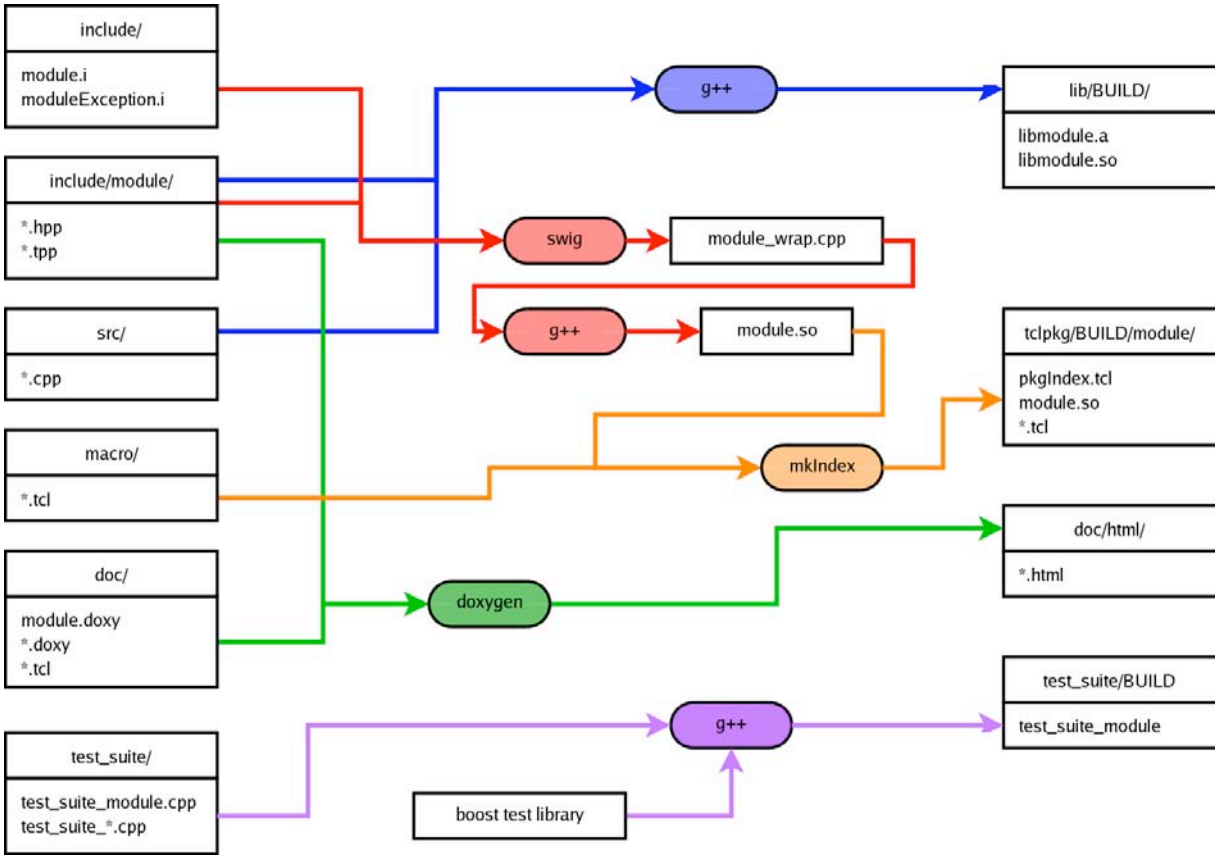

FIG. B.1 – *Structure d'un module* JAFAR. *A gauche: les fichiers éditables par le développeurs. A droite: les différents éléments produits lors de la génération du module.*

génération des librairies.

JAFAR s'appuie sur de nombreux outils libres et très actifs. Il est lui-même disponible sous licence type BSD. Il est intégré aux autres outils du LAAS du projet *OpenRobots* (http://www.openrobots.net).

Après 3 ans de développement et une douzaine d'utilisateurs (doctorant, post-doc, stagiaires,...), une quarantaine de modules sont aujourd'hui disponibles sur notre serveur subversion.

Site de présentation de Jafar : http://www.laas.fr/~tlemaire/jafar/

Bibliographie

[Agarwal 04] S Agarwal, A Awan & D Roth. *Learning to detect object in image via a sparse, part-based representation.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 11, 2004.

[Ballard 81] D.H Ballard. *Generalizing the Hough Transform to detect arbitrary patterns.* Pattern Recognition, vol. 13, no. 2, pages 111–122, 1981.

[Biederman 87] I Biederman. *Recognition-by-components: A theory of human image understanding.* Psycological Review, vol. 94, pages 115–147, 1987.

[Booij 06] O Booij, Z Zivkovic & B Kröse. *From images to rooms.* Dans IEEE International Conference on Intelligent Robots and Systems, 2006.

[Broadbent 58] D.E Broadbent. Perception and communication. Pergamon Press, 1958.

[Brown 05] M Brown & D Lowe. *Unsupervised 3D object recognition and reconstruction in unordered datasets.* Dans International Conference on 3-D Digital Imaging and Modeling, 2005.

[Bulata 96] H Bulata & M Devy. *Incremental Construction of a Landmark-based and Topological Model of Indoor Environments by a Mobile Robot.* Dans Proc. 1996 IEEE International Conference on Robotics and Automation (ICRA'96), Minneapolis (USA), Rapport LAAS N.96190, 1996.

[Burgard 99] W Burgard, A.B Cremers, D Fox, D Haehnel, G Lakemeyer, D Schulz, W Steiner & S Thrun. *Experiences with an interactive museum tour-guide robot.* Artificial Intelligence, vol. 114, pages 3–55, 1999.

[Burt 83] P.J Burt & E.H Adelson. *The laplacian pyramid as a compact image code.* IEEE Transactions on Communications, vol. 31,

no. 4, pages 532–540, 1983.

[Castellanos 01] J.A Castellanos, J Neiro & J.D Tardos. *Multisensory fusion for simultaneous localization and mapping*. Dans IEEE Conference on Robotics and Automation, volume 17, pages 908–914, 2001.

[Chaumette 90] F Chaumette. *La relation vision-commande : théorie et application à des tâches robotiques*. PhD thesis, UniversitÈ de Rennes 1, 1990.

[Chella 97] A Chella, M Frixione & S Gaglio. *A cognitive architecture for artificial vision*. Artificial Intelligence, vol. 89, no. 1-2, pages 73–111, 1997.

[Chella 01] A Chella, M Frixione & S Gaglio. *Conceptual spaces for computer visual representations*. Artificial Intelligence, 2001.

[Choset 01] H Choset & K Nagatami. *Topological simultaneous localization and mapping (SLAM): Toward exact localization without explicit localization*. Dans IEEE Conference on Robotics and Automation, volume 17, 2001.

[Clodic 06] A Clodic, S Fleury, R Alami, R Chatila, G Bailly, L Brethes, M Cottret, P Danes, X Dollat, F Elisei, I Ferrane, M Herrb, G Infantes, C Lemaire, F Lerasle, J Manhes, P Marcoul, P Menezes & V Montreuil. *Rackham: An interactive robot guide*. Dans IEEE Symposium on Robot and Human Interactive Communication, 2006.

[Crowley 98] J.L Crowley, F Wallner & B Schiele. *Position estimation using principal components of range data*. Dans IEEE Conference on Robotics and Automation, 1998.

[Dabis 96] H Dabis, P Palmer & J Kittler. *An interest operator based on perceptual grouping*. Dans Scandinavian conference on Image analysis, pages 211–224, 1996.

[Davidson 02] A.J Davidson & D.W Murray. *Simultaneous localization and mapping using active vision*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, pages 865–880, 2002.

[Deco 00] G Deco & B Schürmann. *A hierarchical neural system with attention top-down enhancement of the spatial resolution for object recognition*. Vision Research, vol. 40, no. 20, pages 2845–2859, 2000.

[Demerci 04] M.F Demerci, A Shokoufandeh, S Dickinson, T Keselman & L Bretzner. *Many-to-many feature matching using spherical coding of directed graph*. Dans European Conference of Computer Vision, 2004.

[Duncan 97] J Duncan. *Integrated mechanisms of selective attention*. Current Opinion in Biology, vol. 7, pages 255–261, 1997.

[Edelman 97] S Edelman. *Computational theories of object recognition*. Cognitive sciences, 1997.

[Edelman 99] S Edelman. Representation and recognition in vision. MIT Press, 1999.

[Elfes 89] A Elfes. *Using occupancy grid for mobile robot perception and navigation*. Computer, vol. 22, no. 6, pages 46–58, 1989.

[Eriksen 86] C.W Eriksen & J.D Saint-James. *Visual attention within and around the field of focal attention: a zoom lens model*. Preception and psychophysics, vol. 40, no. 4, pages 225–240, 1986.

[Fergus 03] R Fergus, P Perona & A Zisserman. *Object Class Recognition by Unsupervised Scale-Invariant Learning*. Dans IEEE Conference on Computer Vision and Pattern Recognition, 2003.

[Fox 98] D Fox, W Burgard & S Thrun. *Active markoc localization for mobile robots*. Robotics and Autonomous Systems, vol. 25, pages 192–207, 1998.

[Frintrop 06] S Frintrop, P Jensfeld & H Christensen. *Pay Attention When Selecting Features*. Dans Proc. 18th Int. Conf. on Pattern Recognition (ICPR'06), Cambridge (UK), 2006.

[Galindo 05] C Galindo, A Saffiotti, S Coradeschi, P Buschka, J.A Fernandez-Madrigal & J Gonzalez. *Multi-hierarchical semantic maps for mobile robotics*. Dans IEEE International Conference on Intelligent Robots and Systems, 2005.

[Gärdenfors 00] P Gärdenfors. Conceptual spaces: the geometry of thought. MIT Press, 2000.

[Gasos 99] J Gasos & A Saffiotti. *integrating fuzzy geometrics maps and topological maps for robot navigation*. Dans International Symposium on Soft Computing, pages 754–760, 1999.

[Gibson 50] J.J Gibson. The perception of visual world. Houghton Mifflin, 1950.

[Gibson 79] J.J Gibson. The ecological approach to visual perception. Houghton Mifflin, 1979.

[Gonzalez-Barbosa 05] J.J Gonzalez-Barbosa. *Vision panoramique pour la robotique mobile: stéréovision et localisation par indexation d'images.* PhD thesis, Université Paul Sabatier - Toulouse III, 2005.

[Grimson 90] E Grimson. Object recognition by computer: the role of geometric constraints. The MIT Press: Cambridge MA, 1990.

[Hamker 05] F Hamker. *The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision.* Computer Vision and Image Understanding, vol. 100, no. 1-2, pages 64–106, 2005.

[Hayet 02] J.B Hayet, C Esteves, M Devy & F Lerasle. *Qualitative modeling of indoor environments from visual landmarks and range data.* Dans Proc. 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems, Lausanne (Suisse), Rapport LAAS N.02149, Juillet 2002.

[Hayet 03] J.B Hayet. *Contribution à la navigation d'un robot mobile sur amers visuels texturés dans un environnement structuré.* PhD thesis, Université Paul Sabatier - Toulouse III, 2003.

[Hoppenot 01] P. Hoppenot & E. Colle. *Localization and control of a rehabilitation mobile robot by closehuman-machine cooperation.* IEEE Trans. on Rehabilitation Engineering, vol. 9, no. 2, pages 181–190, 2001.

[Huang 04] H.M Huang, E Messina, R Wade, R English, B novak & J ALbus. *Autonomy measures for robots.* Dans International Mechanichal Engineering Congress, 2004.

[Huu 05] N Do Huu, W Paquier & R Chatila. *Combining structural descriptions and image-based representations for image object and scene recognition.* Dans International Joint Conference on Artificial Intelligence (IJCAI), Edinburgh, Scotland, 2005.

[Huu 06] N Do Huu & R Chatila. *Learning the sensory-motor loop with neurons : recognition, association, prediction, decision.* Dans FS2HSC - IEEE/RSJ IROS 2006 Workshop - From Sensors to Human Spatial Concepts (FS2HSC 2006), Beijing, China, 2006.

[Itti 98] L Itti, C Koch & E Niebur. *A Model of Saliency-Based Visual Attention for Rapid Scene Analysis.* IEEE Transactions on Pat-

tern Analysis and Machine Intelligence, vol. 20, no. 11, pages 1254–1259, citeseer.ist.psu.edu/itti98model.html.

[Itti 01] L Itti & C Koch. *Feature Combination Strategies for Saliency-Based Visual Attention Systems*. Journal of Electronic Imaging, vol. 10, no. 1, pages 161–169, 2001.

[Jensfelt 01] J Jensfelt & H Christensen. *Pose tracking usgin laser scanning and minimalistic environmental models*. Dans IEEE Conference on Robotics and Automation, 2001.

[Kadir 01] T Kadir & M Brady. *Scale, saliency and image description*. International Journal of Computer Vision, 2001.

[Kamsickas 03] G.M Kamsickas & J.N Ward. *Developing UVGs for the FCS Program*. SPIE, vol. 5083, 2003.

[kapur 85] J.N kapur, P.K Sahoo & A.K.C Wong. *"A new method for gray-level picture thresholding using the entropy of the histogram*. Graphical Models and Image Processing, vol. 29, pages 273–285, 1985.

[Koch 85] C Koch & S Ullman. *Shifts in selective visual attention: towards the underlying neural circuitry"*. Human Neurobiology, vol. 4, pages 219–227, 1985.

[Korthenkamp 94] D Korthenkamp & T Weymouth. *Topological mapping for mobile robot using a combinaison of sonar and vision sensing*. Dans National Conference of Artificial Intelligence (AAAI), 1994.

[Kröse 01] B Kröse, O Vlassis, R Bunschoten & Y Motomura. *A probabilistoic model for appearance-based robot localization*, 2001.

[Kubacki 05a] J Kubacki, B Giesler & C Parlitz. *Active Autonomous Object Modelling for Recognition and Manipulation: Towards a Unified Object Model and Learning Cycle*. Dans Proc. Fachgesprâ°che fÂ¸r Autonome Mobile Systeme, Stuttgart (Germany), Decembre 2005.

[Kubacki 05b] J Kubacki & K Pfeiffer. *Using Range Imaging Sensors with Color Imaging Sensors in Cognitive Robot Companions: A New and Simple Calibration Technique Based on Particle Swarm Optimization*. Dans Proc. First Range Imaging Research Day (RIM Day), Septembre 2005.

[Kuipers 77] B.J Kuipers. *Representing knowledge of large-scale space*. Rapport technique TR-418, MIT-CSAIL, 1977.

[Kuipers 00] B.J Kuipers. *The spatial semantic hierarchy.* Artificial Intelligence, vol. 119, pages 191–233, 2000.

[LaBerge 95] D LaBerge. Attentional processing; the brain's art of mindfulness. Havard University Press, 1995.

[Lampe 06] A Lampe. *Méthodologie d'évaluation du degré d'autonomie d'un robot mobile terrestre.* PhD thesis, Institut National Polytechnique de Toulouse, 2006.

[Leonard 91] J.J Leonard & H.F Durrant-Whyte. *Simultaneous map building and localization for an autonomous mobile robot.* Dans International Workshop on Intelligent Robots and Systems, volume 3, pages 1442–1447, 1991.

[Leonard 92] J.J Leonard & H.F Durrant-Whyte. *Directed sonar sensing for mobile robot navigation*, 1992.

[Leonardis 02] A Leonardis, H Bischof & J Maver. *Multiple eigenspaces.* Pattern Recognition, pages 1–15, 2002.

[Lindeberg 94] T Lindeberg. Scale-space theory in computer vision. Kluwer Academic Publishers, 1994.

[Lowe 99] D Lowe. *Object recognition from local scale-invariant features.* Dans International Conference on Computer Vision, 1999.

[Lowe 01] D Lowe. *Local feature view clustering for object recognition.* Dans Conference of vision and pattern recognition, 2001.

[Lowe 04] D.G Lowe. *Distinctive image features from scale-invariant keypoints.* International Journal of Computer Vision, 2004.

[Marr 78] D Marr & H.K Nishihara. *Representation and recognition of the spatial organization of three-dimensional shapes.* Proc. of Royal Society of London, vol. Serie B (200), pages 269–294, 1978.

[Matas 02] J Matas, O Chum, U Martin & T Pajdla. *Robust wide baseline stereo from maximally stable extremal regions.* Dans British Machine Vision Conference, volume 1, pages 384–393, 2002.

[Mikolajczyk 02] K Mikolajczyk & C Schmid. *An affine invariant interest point detector.* Dans European Conference of Computer Vision, 2002.

[Mikolajczyk 03] K Mikolajczyk & C Schmid. *A performance evaluation of local descriptors.* Dans IEEE Conference on Computer Vision and Pattern Recognition, 2003.

[Mikolajczyk 05] K Mikolajczyk, T Tuytelaars, C Schmid, A Zisserman, J Matas, F Schaffalitzky, T Kadir & L Van Gool. *A comparison of affine region detectors*. International Journal of Computer Vision, vol. 65, no. 7, pages 43–72, 2005.

[Moravec 88] H.P Moravec. *Sensor fusion in certainty grids for mobile robots*. Artificial Intelligence, vol. 9, no. 2, pages 61–74, 1988.

[Moreels 04] P Moreels, M Maire & P Perona. *Recognition by probabilistic hypothesis construction*. Dans European Conference of Computer Vision, 2004.

[Moutarlier 89] P Moutarlier & R Chatila. *Stochastic multisensory data fusion for mobile robot localization and environment modelling*. Dans International Symposium of Robotic Research, 1989.

[Murphy-Chutorian 05] E Murphy-Chutorian & J Triesch. *Shared features for scalable appearance-based object recognition*. Dans IEEE Workshop on Application of Computer Vision, 2005.

[Pope 00] A.R Pope & D.G Lowe. *Probabilistic models of appearance for 3D object recognition*. International Journal of Computer Vision, vol. 40, no. 2, pages 149–167, 2000.

[Posner 80] M.E Posner. *Orienting of attention*. Q. J. Exp. Psychol., vol. 32, pages 3–25, 1980.

[Remazeilles 05] A Remazeilles. *Navigation à partir d'une mémoire d'images*. PhD thesis, IRISA/INRIA, Université de Rennes, 2005.

[Rensink 97] J Rensink, J O'Regan & J Clark. *To see or not to see: The need fit attention to perceive changes in scenes*. Psycological Science, vol. 8, pages 368–373, 1997.

[Rensink 00] R.A Rensink. *The dynamic representation of scenes*. Visual Cognition, vol. 7, pages 17–42, 2000.

[Riesenhuber 99] M Riesenhuber & T Poggio. *Separate visual pathway for perception and action*. Nature, vol. 2, no. 11, pages 1019–1025, 1999.

[Rothganger 06] F Rothganger, S Lazebnik, C Schmid & J Ponce. *3D object modeling and recognition using local affine invariant image descriptors and multi-view spatial constraints*. International Journal of Computer Vision, vol. 66, 2006.

[Schiele 00] B Schiele & J.L Crowley. *Recognition without correspondence using multidimensional receptive field hidtograms*. International Journal of Computer Vision, 2000.

[Schmid 99] C Schmid. *A Structured Probabilistic Model for Recognition*. Dans IEEE Conference on Computer Vision and Pattern Recognition, 1999.

[Sezgin 04] M Sezgin & B Sankur. *Survey over image thresholding techniques and quantitative performance evaluation*. Journal of Electronic Imaging, vol. 13, no. 1, pages 146–165, 2004.

[Stevens 78] A Stevens & P Coupe. *Distortions in judged spatial relations*. Cognitive Psychology, vol. 13, pages 422–437, 1978.

[Sun 03] Y Sun & R Fischer. *Object-based visual attention for computer vision*. Artificial Intelligence, vol. 20, no. 11, pages 77–123, 2003.

[Tapus 05] A Tapus & R Siegwart. *Incremental robot mapping with fingerprint*. Dans IEEE International Conference on Intelligent Robots and Systems, 2005.

[Tarr 98] M.J Tarr & H.H Bülthoff. Object recognition in man, monkey and machine. MIT Press, 1998.

[Thorpe 96] S Thorpe, D Fize & C Marlot. *Spped of porcessing in the human visual system*. Nature, vol. 381, pages 520–522, 1996.

[Thorpe 98] S Thorpe & J Gautrais. *Rank order coding: A new coding scheme for rapid processing in neural network*. Computational Neuroscience: Trends in Research, pages 113–118, 1998.

[Thrun 96] S Thrun & A Bucken. *integrating grid-based and topological maps for mobile robot navigation*. Dans National Conference of Artificial Intelligence (AAAI), 1996.

[Thrun 98] S Thrun. *Learning metric-topological maps for indoor mobile robot navigation*. Artificial Intelligence, vol. 99, pages 21–71, 1998.

[Thrun 00] S Thrun, M Beetz, , M Bennewitz, W Burgard, A.B Cremers, F Dellaert, D Fox, D Haehnel, C Rosenberg, N Roy, J Schulte & D Schulz. *Probabilistic algorithms ans the interactive museum tour-guide robot Minerva*. Journal of Robotics Research, vol. 19, no. 11, 2000.

[Todt 04] E. Todt & C. Torras. *Detecting salient cues through illumination-invariant color ratios*. Robotics and Autonomous Systems, vol. 48, no. 2-3, pages 111–130, Septembre 2004.

[Tolman 48] E.C Tolman. *Cognitive maps in rats and mens*. Psycological Review, 1948.

[Topp 06] E Topp, H Huettenrauch, H Christensen & P Eklundh. *Bringing Together Human and Robotic Environment Representation - A Pilot Study*. Dans Proc. 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'2006), Seoul (Korea), 2006.

[Torralba] A Torralba, K.P Murphy & W.T Freeman. *Shared features for multiclass object detection*. Towards Category-Level Object Recognition. Springer Lecture Notes in Computer Science.

[Treisman 80] A Treisman & G Gelade. *A feature-integration theory of attention*. Cognitive Psychology, vol. 12, no. 1, 1980.

[Triggs 99] B Triggs, P McLauchlan, R Hartley & A Fitzgibbon. *Bundle Adjustment: A Modern Synthesis*. 1999.

[Tsotsos 95] J.K Tsotsos, S.M Culhane, W.Y.K Wai, Y.H Lai, N Davis & N Nuflo. *Modeling visual attention via selective tuning*. Artificial Intelligence, vol. 78, pages 507–545, 1995.

[Turvey 81] M Turvey & C Carello. *Cognition: the view from ecological realism*. Cognition, 1981.

[Tversky 93] B Tversky. *Cognitive maps, cognitive collages and spatial mental models*. Spatial Information Theory, 1993.

[Vasudevan 07] S Vasudevan, S Gachter, V Nguyena & R Siegwart. *Cognitive maps for mobile robots : an objetc based approach*. Robotics and Autonomous Systems, vol. 55, pages 359–371, 2007.

[Welch 01] G Welch & G Bischop. *An introduction to the kalman filter*. Rapport technique TR 95-041, University of North Carolina at Chapel Hill, Department of Computer Science, 2001.

[Zivkovic 05] Z Zivkovic, B Bakker & B Kröse. *Hierarchical map building using visual landmarks and geometric constraints*. Dans EEE/RSJ International Conference on Intelligent Robots and Systems, 2005.

Résumé

Dans le cadre de la robotique mobile de service, la compréhension de l'environnement joue un rôle primordiale: c'est en effet une des clés vers l'autonomie et l'interaction de haut niveau avec l'utilisateur. Pour cela, le robot doit acquérir un ensemble de représentations adaptées aux différentes tâches à effectuer. Des modèles métriques et topologiques sont utilisés pour les déplacements, des modèles 3D métriques d'objets pour la manipulation ou des modèles d'apparences pour la reconnaissance et l'interaction: ces représentations permettront au robot de répondre à des requêtes telles que "Va chercher un verre dans la cuisine". Un robot de service doit notamment être capable d'acquérir de nouvelles connaissances au fur et à mesure qu'il explore un nouvel environnement. C'est dans cette perspective que s'inscrivent les travaux de cette thèse: nous proposons d'apprendre en ligne un modèle d'apparence localisé de structures locales qui pourront être nommées par l'utilisateur; dans l'apprentissage d'un environnement intérieur par exemple, il sera alors possible de caractériser un lieu topologique (ex: la cuisine) par un ensemble de structures locales ou d'objets s'y trouvant (réfrigérateur, cafetière, évier, ...).

Pour découvrir ces structures locales, nous proposons une approche cognitive, exploitant des processus visuels pré-attentif et attentif, mis en oeuvre à partir d'un système sensoriel multi-focal. Le processus pré-attentif a pour rôle la détection de zones d'intérêt, supposés contenir des informations visuelles discriminantes : basé sur le modèle d'attention visuelle de Itti et Koch, il détecte ces zones dans une carte de saillance, construite à partir des images acquises avec une caméra large champ (éventuellement panoramique) ; une zone détectée est ensuite suivie sur quelques images afin d'estimer grossièrement la taille et la position 3D de la structure locale de l'environnement qui lui correspond. Le processus attentif exploite une caméra active PTZ, pour focaliser sur la zone d'intérêt : la caméra est contrôlée pour que l'image reste centrée et focalisée sur la structure tant qu'elle reste dans le champ visuel accessible par le robot. Le but est de caractériser chaque structure locale, par un modèle d'apparence sous la forme de mémoires associatives vues-patches-aspects. De chaque image sont extraits des points d'intérêt, caractérisés par un descripteur d'apparence local. Ce modèle d'apparence est ensuite exploité pour la recon-

naissance et la localisation grossière d'un objet perçu par le robot. Enfin, l'utilisateur peut à tout moment nommer les structures locales significatives et éliminer les autres.

Mots-clé Apprentissage en ligne - Reconnaissance d'objets - Modèle d'apparence - Objet 3D - SIFT - Attention - Exploration - Environnement Intérieur

Summary

Spatial awareness for a companion mobile robot is essential for its autonomy and its high-order interaction with users. It has to learn several task-adapted models of the environment : metric and topologic models are used for localization and navigation, 3D metric models for objets handling and appearance models for recognition and user-robot interactions. With such models, the robot will be able to answer requests in common language like "go to the kitchen and bring me back a glass". Moreover, those models have to be learn online while the robot explore its environment. Thus, our purpose in this thesis is an attentional approach for visual exploration of indoor environment based on the detection and modeling of salient objects which could be named by the user. then, a topological place (ex: the kitchen) could be described by an set of its objects (refrigerator, coffee machine, sink, ...)

In order to discover such objects of interest, we proposed a cognitive approach based on two pre-attentive and attentive processes and a multi-focal visual system. The pre-attentive process is used to highlight potential interesting locations: using the visual attentional system of Itti and Kock, regions of interest (ROI) are selected in a saliency map built with a widefield camera; then each ROI is tracked over some successive images and associated with a coarse 3D position and size in the environment by bundle adjustment. The second attentive process use a PTZ camera to focus on the previously extracted 3D locations and capture some viewpoints of the object. Each focused image is represented by invariant interest features and invariant descriptors and is used to learn a views-features appearance based model of the object. Those models are then used during the recognition phase for scene interpretation. Finally, the detected objects are validated by the user who can name or suppress them.

Keywords: Online learning - 3D Object Recognition - Appearance based model - SIFT - Attention - Exploration - Indoor

Exploration Visuelle d'Environnement Intérieur par Détection et Modélisation d'Objets Saillants

Résumé :
Un robot compagnon doit comprendre le lieu de vie de l'homme pour satisfaire une requête telle que "Va chercher un verre dans la cuisine" avec un haut niveau d'autonomie. Pour cela, le robot doit acquérir un ensemble de représentations adaptées aux différentes tâches à effectuer. Dans cette thèse, nous proposons d'apprendre en ligne un modèle d'apparence de structures locales qui pourront être nommées par l'utilisateur. Cela permettra ensuite de caractériser un lieu topologique (ex: la cuisine) par un ensemble de structures locales ou d'objets s'y trouvant (réfrigérateur, cafetière, evier, ...). Pour découvrir ces structures locales, nous proposons une approche cognitive, exploitant des processus visuels pré-attentif et attentif, mis en oeuvre à partir d'un système sensoriel multi-focal. Le processus pré-attentif a pour rôle la détection de zones d'intérêt, supposées contenir des informations visuelles discriminantes: basé sur le modèle de « saillance » de Itti et Koch, il détecte ces zones dans une carte de saillance, construite à partir d'images acquises avec une caméra large champ; une zone détectée est ensuite suivie sur quelques images afin d'estimer grossièrement la taille et la position 3D de la structure locale de l'environnement qui lui correspond. Le processus attentif se focalise sur la zone d'intérêt : le but est de caractériser chaque structure locale, par un modèle d'apparence sous la forme de mémoires associatives vues-patches-aspects. De chaque image sont extraits des points d'intérêt, caractérisés par un descripteur d'apparence local. Après cette phase d'exploration, l'homme peut annoter le modèle en segmentant les structures locales en objets, en nommant ces objets et en les regroupant dans des zones (cuisine). Ce modèle d'apparence sera ensuite exploité pour la reconnaissance et la localisation grossière des objets et des lieux perçus par le robot.

Mots-clés : apprentissage en ligne, reconnaissance, modèle d'apparence, objet 3D, SIFT, attention

Abstract:
A robot companion has to understand a domotic environment in order to execute requests like «Search a glass in the kitchen» with a high level of autonomy. So the robot must acquire several representations adapted to the tasks to be executed. This thesis proposes an on line learning method of an environment model expressed as a set of local structures described by appearance-based characteristics, and possibly named by a tutor. Such descriptions could be used in order to define a topological area (e.g. the kitchen) by a set of local structures or objects that could be found here (e.g. glasses, fridge, pans).
For the construction of such a representation, it is proposed a cognitive mehod, based on attentive and preattentive visual processes, acquiring images from a multifocal sensor. The preattentive process aims at detect interest regions, that could contain discriminant visual information ; based on the saliency concept proposed initially bu Itti and Koch, interest regions are extracted from a saliency map, built from images acquired by a short lens or panoramic camera (large view field). Such a region is then tracked on several successive images acquired while the robot is moving, so that the size and the 3D position of the corresponding local structure could be coarsely estimated. Then the attentive process exploits attentionam mechanisms in order to be focused successively on each interest region : it aims to characterize each local structure by an appearance-based model defined by an associative memort views-patches-aspects. Salient scaled patches or SIFT features are extracted from every image. After this exploration step is over, the robot tutor could annotated the model, segmenting local structures in objects, naming objects and grouping them in areas (kitchen). Then, the robot exploits this environment model for the recognition and the coarse localization of objects and areas.

Keywords: on line learning, 3D object recognition, appearance-based model, interest points, attention