Table des matières

Introduction

Le sujet de notre étude est l'interprétation, par un système informatique autonome, de séquences temporelles d'images, appelée également interprétation dynamique de scènes. Un système autonome doit être capable d'appréhender et de comprendre le monde qui l'entoure à l'aide de capteurs qui sont, en ce qui nous concerne, des caméras. Les données perçues à travers les séquences d'images doivent être transformées en représentations internes de la scène permettant au système autonome d'expliquer ces données à l'aide de scénarios et de leur attribuer un sens. Ce passage de la perception à l'attribution de sens constitue le processus d'interprétation.

Ce sujet est ambitieux, car il vise à la description abstraite des activités se déroulant dans une scène et s'attache à expliquer les raisons de ces activités. Il est néanmoins réalisable dès que le contexte de la scène est connu ; ce contexte comprend en particulier des descriptions de l'environnement et des comportements pouvant s'y produire. Par ailleurs, la présence du mouvement facilite la détection d'indices pertinents, aidant à l'élaboration de raisonnements symboliques et à la compréhension de la scène. Ce couplage vision et raisonnement abstrait fait du sujet d'interprétation de séquences d'images un sujet de recherche de premier ordre. Cependant c'est un sujet récent, ayant débuté dans les années 80 en Europe (et plus particulièrement en Allemagne, au Royaume Uni et en France) et commençant maintenant en Amérique ainsi qu'en Asie.

L'objectif de nos travaux est d'une part la modélisation du problème d'interprétation de séquences d'images et d'autre part la validation de ce modèle à travers le développement d'un système générique d'interprétation. Nous illustrons nos propos par l'étude des problèmes de surveillance de scènes et plus particulièrement à travers les applications de vidéosurveillance relatives à des activités humaines. Le chapitre suivant décrit précisément ces objectifs et expose le plan de ce mémoire.

D'un point de vue éthique la surveillance est une tâche noble, malgré sa

mauvaise presse auprès du grand public. Dans notre cas les applications envisagées consistent à automatiser le travail rébarbatif d'un opérateur humain devant surveiller un grand nombre d'écrans. Il s'agit alors de lui signaler la détection des événements dignes d'intérêts. De plus, l'objectif premier de la surveillance n'est pas la répression, mais plutôt la sécurité et la protection des individus contre les violences du monde moderne. Par exemple, certains systèmes de surveillance sauvent déjà des vies humaines sur des autoroutes en alertant les secours et en leur permettant de porter rapidement assistance aux accidentés de la route.

Comme tout outil puissant, l'utilisation du système obtenu à des fins néfastes est dangereuse. Cependant il serait dommage de ne plus utiliser le feu sous prétexte qu'il peut brûler.

Chapitre 1

Analyse du problème

Ce chapitre présente le cadre du problème que nous cherchons à résoudre. Nous commençons par décrire nos motivations en termes d'applications et par expliquer l'intérêt de ce problème en tant que sujet de recherche. Ensuite dans la section 1.2, pour permettre au lecteur une meilleure compréhension de nos travaux, nous donnons les définitions des termes utilisés dans la suite du mémoire.

Dans la section 1.3 nous présentons et comparons plusieurs architectures de systèmes d'interprétation, et nous en déduisons les principes généraux du processus d'interprétation de séquences d'images. Dans la section 1.4, nous continuons en exposant nos objectifs et en décrivant le type des résultats escomptés. Enfin, nous concluons en donnant le plan de lecture du mémoire.

1.1 Applications cibles

Cette section décrit nos motivations en termes d'applications. En fonction des différents domaines d'application envisagés, nous décrivons les objectifs que doit atteindre un système d'interprétation et ses caractéristiques. Puis nous présentons des applications en cours ou ayant déjà traité des aspects du problème d'interprétation de séquences d'images. Enfin, nous décrivons plus précisément le type d'application que nous avons choisi pour valider notre approche.

Faire le point sur les domaines d'application est doublement important. Il permet en effet d'illustrer nos propros et de fournir un cadre de validation des modèles proposés. De plus, il permet d'établir l'intérêt de nos recherches et de déterminer la partie réaliste et directement utilisable de notre travail.

1.1.1 Domaines d'application

Nous nous intéressons ici à l'interprétation de séquences temporelles d'images, dont l'objectif est l'étude des comportements des objets mobiles évoluant dans une scène. Il s'agit alors de réaliser un système, appelé système d'interprétation, analysant automatiquement une scène donnée à partir de séquences d'images. Les motivations d'ordre applicatif sont multiples et concernent différents domaines :

- **Surveillance du trafic routier** : les applications de ce domaine sont faciles à mettre en œuvre, comparativement à celles des autres domaines. Les véhicules sont des objets mobiles facilement identifiables et la gamme de comportements est limitée par l'environnement. Ce domaine concerne différents types d'application :

  . Détection d'incidents - le système d'interprétation prévient les situations dangereuses sur un réseau routier, telles que *« la circulation à contresens d'un véhicule »*, détecte les incidents et déclenche des alarmes. Il existe des systèmes commerciaux de détection d'incidents déjà diffusés auprès du grand public. D'autres systèmes plus ambitieux dans la reconnaissance de comportements, sont encore du domaine de la recherche. Par exemple, B. Neumann et son équipe ont développé le système Naos, qui à partir d'une séquence d'images, décrit une scène routière à un auditeur se trouvant dans l'impossibilité de l'observer (Mohnhaupt and Neumann, 1990). L'objectif de Naos est de donner une description en langage naturel des activités se déroulant dans la scène. Naos ne fonctionne qu'avec des images synthétisées. De même, H. Nagel et ses collaborateurs ont développé un système, Epex, dédié à l'interprétation de scènes routières à partir de séquences d'images (Nagel, 1988). Epex utilise des images réelles. Enfin, le projet européen VIEWS avait pour objectif la surveillance de scènes extérieures en temps réel, à partir de l'analyse de séquences d'images (Corrall, 1992). Deux applications pilotes ont été développées au cours de ce projet. La première gère le trafic aérien au sol d'un aéroport, et la seconde détecte les situations d'incidents potentiels dans une scène routière.

  . Carrefours intelligents - le système d'interprétation optimise la circulation et le passage des piétons, en contrôlant les feux de circulation de plusieurs carrefours. Des systèmes de carrefours

intelligents sont déjà opérationnels et commencent à être commercialisés, comme celui développé par S. Sellam et son équipe (Sellam and Boulmakoul, 1994).

. Surveillance de zones spécifiques - le système surveille des zones fréquentées par des véhicules, telles que des péages ou des stations services pour détecter des comportements anormaux de véhicules. H. Nagel et ses collaborateurs travaillent par exemple, sur un système de surveillance de stations services (Nagel, 1991). Certains systèmes de surveillance de péages sont déjà commercialisés.

. Navigation dans une scène routière - le système est embarqué dans un véhicule. Il doit détecter les véhicules environnants et analyser leurs comportements. Ces systèmes de navigation sont souvent limités au suivi des véhicules environnants (Hutber, 1995).

. Surveillance aérienne - le système surveille des zones sensibles, telles qu'une route, un champ de bataille, afin de protéger des objets de valeur (p. ex. un convoi), ou de détecter des comportements particuliers. Ces systèmes de surveillance aérienne sont encore à l'étude.

– **Surveillance d'activités humaines** : dans ces applications un système d'interprétation a pour objectif de surveiller des zones à comportements spécifiques, telles que les métros, les parkings, les supermarchés, les aérogares, les banques ou les zones piétonnières. Le but est de détecter les comportements anormaux des individus évoluant dans ces zones et de prévenir les comportements dangereux. Ces comportements correspondent à des actes de vandalisme, de vente à la sauvette (p. ex. vente de stupéfiants), de vol à l'étalage, d'agression ou de terrorisme. Par exemple, l'analyse de comportement de groupes, comme des bandes, permet de prévenir certains actes d'agression. Ce type d'application est en plein essor et de nombreux systèmes de surveillance sont à l'étude dans le monde industriel et académique. Par exemple, le projet européen Esprit HPCN PASSWORDS vise des applications de surveillance de métros, de parkings et de supermarchés à l'aide d'une seule caméra (Chleq and Thonnat, 1996). De même, le projet Perception a comme objectif la surveillance de parkings, mais à l'aide de plusieurs caméras (Castel et al., 1996).

  Des applications similaires visent l'analyse statistique de comportements, dont l'objectif est de compter le nombre d'individus et de déterminer leur flux de circulation. Ces systèmes sont déjà opérationnels

(Sato et al., 1993). Des applications concernant l'analyse de comportements animaliers, comme celles menées par D. Hogg sur des volailles et troupeaux, sont également à l'étude.

- **Analyse de scènes sportives** : dans ces applications un système d'interprétation analyse les comportements des sportifs. Ces applications diffèrent du cas précédent par la présence d'un environnement plus contraint (p. ex. un terrain de sport) et par le nombre limité de comportements. Les sports concernés sont principalement le football (Choi et al., 1997), et également le tennis, le basket et le football américain (Intille and Bobick, 1995). Ce type de système a pour objectif d'aider les entraîneurs dans l'analyse du jeu et de fournir des statistiques sur les tactiques utilisées. Certains de ces systèmes sont en voie de commercialisation, bien que nécessitant encore des corrections manuelles.

  Une variante de ce type d'application a été abordée par le projet VITRA (VIsual TRAnslator). L'objectif de ce projet est de développer un système expliquant à l'aide de dialogues avec un utilisateur, le contenu d'une séquence d'images. VITRA a en particulier donné naissance au système Soccer. Soccer analyse et commente simultanément en allemand de courtes séquences de football, comme dans un reportage radio (Herzog, 1995).

- **Analyse de gestes** : dans ce type d'application, la caméra est fixe, à proximité et en face de l'individu filmé. De plus, les systèmes d'interprétation ne prennent souvent en compte qu'un seul individu. Ces systèmes ont pour but de comprendre les gestes de l'individu, afin de communiquer avec lui. Par exemple, certains systèmes ont pour objectif de lire sur les lèvres ou de comprendre le langage des signes (c.-à-d. langage des sourds-muets). Ces systèmes commencent à être opérationnels, mais restent encore du domaine de la recherche. Une autre application concerne les kiosques intelligents (p. ex. distributeurs automatiques de billets, bornes d'informations). Le système doit réagir en fonction du comportement de l'utilisateur, et comprendre s'il est satisfait. Ce système permet alors d'évaluer le succès du service proposé. Une application similaire concerne les locaux intelligents. Un local intelligent est un local muni de capteurs permettant de répondre aux attentes d'un utilisateur sans qu'il ait besoin de s'encombrer d'outils spécifiques (p. ex. clavier) (Pentland, 1995). Par exemple, A. Bobick a construit une chambre pour des enfants, leur racontant une histoire à l'aide d'effets spéciaux, et réagissant à leurs comportements. De même

dans (Pentland, 1995), l'auteur a développé un système contrôlant l'habitacle d'un véhicule, afin d'anticiper le comportement du conducteur. Dans (Bobick and Pinharez, 1995), le système d'interprétation contrôle les caméras d'un studio de télévision, et obéit aux ordres du réalisateur. Dans le domaine sportif, d'autres systèmes d'interprétation ont pour objectif d'enseigner des enchaînements de gestes. Par exemple dans (Becker and Pentland, 1997), les auteurs présentent un système d'interprétation qui enseigne le T'ai-chi. Dans (Campbell and Bobick, 1995), il s'agit de reconnaître les pas d'une danseuse de ballet classique. D'autres applications concernent les jeux vidéos et la réalité virtuelle. Par exemple A. Pentland utilise une caméra dominant un écran géant qui affiche ce que doit voir l'utilisateur. Le système fait alors évoluer (et interagir) l'utilisateur dans le monde se déroulant à l'écran, reprenant des scénarios de jeux vidéos, tels que *« Doom »*. Enfin d'autres applications concernent la réparation spécialisée de machines et la réalité augmentée. Par exemple dans (Brand et al., 1997), un ouvrier muni de lunettes répare une imprimante et voit s'afficher devant ses lunettes, le nom des pièces ainsi que leur mode de réparation.

- **Analyse de scènes en robotique** : dans ce type d'application, l'objectif est pour un robot mobile de comprendre son environnement. Par exemple dans le cadre du projet SKIDS (Grandjean, 1991), un robot mobile muni de nombreux capteurs est capable en particulier d'analyser des séquences d'images. Son but est de comprendre et d'expliquer son environnement comprenant des objets mobiles. Des applications plus ambitieuses consistent à faire coopérer plusieurs robots. Par exemple, une coupe du monde de football entre robots *« Robot-Cup »*, est organisée entre différents centres de recherche en robotique.

Cette énumération n'a pas pour but d'être exhaustive, mais de fixer les idées sur les possibilités d'application d'un système d'interprétation. Une première caractéristique de ces applications est le traitement à la volée des images (c.-à-d. traitement continu et en direct). Cependant, un bon nombre de ces applications peut être envisagé avec un traitement en différé des séquences d'images. Par exemple, en analyse de scènes sportives, l'entraîneur n'a pas besoin d'une analyse en direct des performances de ses joueurs. Dans ce type d'application, une correction manuelle de l'analyse des comportements est alors possible, permettant à une entreprise de commercialiser l'analyse d'une séquence d'images, sans que pour autant le système d'interprétation soit totalement autonome.

La problématique de l'interprétation en différé rejoint celle de l'indexation de séquences d'images par leur contenu. En effet, dans des bases de données contenant un grand nombre de séquences vidéo, il est souvent intéressant de rechercher automatiquement une séquence, à partir d'une description des activités se déroulant dans la séquence. Le problème réside alors dans la reconnaissance automatique de scénarios suffisamment précis et informatifs, pour discriminer la séquence de la base de séquences vidéo. En vidéosurveillance par exemple, les opérateurs conservent en général un grand nombre de séquences. Dès qu'ils souhaitent retrouver une séquence particulière, ils sont obligés de visionner manuellement toutes les bandes vidéo.

Dans toutes ces applications, il existe différents degrés de réalisation. Dans le secteur industriel, les systèmes d'interprétation ont comme premier objectif la robustesse. Ces systèmes se limitent alors souvent à une détection des objets mobiles. En ce qui concerne le secteur académique, les systèmes d'interprétation ont pour objectif d'analyser des comportements complexes. Bien qu'étant utilisables avec des conditions moins contraignantes (p. ex. scènes filmées en laboratoire), ces systèmes permettent de délimiter les possibilités attendues de l'interprétation automatique de séquences d'images.

1.1.2 Applications abordées

Dans le cadre de cette thèse, la classe d'applications choisie pour valider notre système d'interprétation est la vidéosurveillance de scènes intérieures et extérieures, partiellement structurées et observées à l'aide d'une caméra monoculaire, couleur et fixe. Dans ce cadre, les objets mobiles sont indifféremment des êtres humains ou des véhicules. Les données d'entrée du système d'interprétation sont des séquences d'images de métro et de parking, prises dans le cadre du projet européen Esprit HPCN PASSWORDS. L'objectif du système est de déclencher une alarme dès qu'un comportement anormal est reconnu. Nous avons choisi ce type d'application pour plusieurs raisons :

- La vidéosurveillance est un domaine nouveau et porteur. De nombreux projets tant industriels qu'académiques, sont en train de naître sur ce thème. Le thème de la vidéosurveillance est en effet intéressant en tant que sujet de recherche, car il permet d'étudier des comportements complexes, mettant en scène plusieurs objets mobiles en même temps et sur de longues séquences d'images (dépassant la dizaine de minutes).

- Ce domaine favorise l'élaboration de systèmes génériques, car les applications de vidéosurveillance sont suffisamment complexes et diverses

pour nécessiter la conception d'un système d'interprétation sophistiqué. Ce système peut alors être facilement modifié, afin de traiter d'autres types d'application mettant en scène des activités moins complexes.

- Ce type d'application nous permet de nous confronter à des conditions réelles d'utilisation. Ces applications nécessitent généralement un traitement en temps réel des séquences d'images, et possèdent souvent de mauvaises conditions d'acquisition d'images. Par exemple, les caméras sont en général de faible résolution et les scènes sont soumises aux évolutions irrégulières de l'éclairage, dues par exemple aux variations des conditions météorologiques.

- Le choix de ce thème nous permet également de bénéficier des travaux accomplis dans le cadre du projet européen Esprit HPCN PASSWORDS (Chleq and Thonnat, 1996). Nous pouvons ainsi récupérer des séquences d'images, données d'entrée du système d'interprétation et bénéficier d'une étude de marché du domaine. Cette étude suggère de considérer des conditions réelles d'utilisation d'un système d'interprétation. Ces conditions consistent à utiliser le parc de caméras déjà installées sur sites (c.-à-d. des caméras fixes monoculaires de faible sensibilité). Cette étude facilite également la sélection des activités à analyser, pouvant correspondre aux attentes réelles d'utilisateurs.

Les conditions d'utilisation du système d'interprétation que l'on se propose d'élaborer, sont ainsi particulièrement contraignantes afin de correspondre à des conditions réelles. Notre objectif est de déterminer ce qu'il est possible de faire sous de telles conditions, et d'établir les limites d'utilisation de ce système d'interprétation. Si les résultats du système ne satisfont pas l'utilisateur potentiel, ce dernier pourra alors de son côté réduire les exigences et restreindre le domaine d'utilisation.

1.2 Définition des termes employés

Dans cette section, nous décrivons les différents termes et notions utilisés dans ce mémoire :

- **Termes employés pour décrire les données utilisées par le système d'interprétation** : la *« séquence d'images »* représente la donnée d'entrée du système. Sa cadence est de quatre à cinq images par

seconde et sa durée est de l'ordre d'une dizaine de minutes. La *« scène »* correspond au volume 3D filmé par la caméra, comprenant l'environnement statique ainsi que les objets mobiles. Comme le montre la figure 1.1, le *« fond de la scène »* correspond à l'environnement statique, c'est-à-dire à la scène sans les objets mobiles. Le *« contexte de la scène »* rassemble l'environnement statique (p. ex. un banc), les informations décrivant l'acquisition de la séquence d'images (p. ex. le type de la caméra), des informations symboliques sur l'état du système (p. ex. la liste des objets déjà reconnus) et des informations sur les desiderata de l'utilisateur. Cette notion de contexte est difficilement formalisable. Elle est plus amplement décrite dans le chapitre 2.

- **Termes techniques correspondant aux données manipulées par le système d'interprétation** : une *« région mobile »* est une portion de l'image, dont l'intensité change au cours de la séquence. Elle peut correspondre soit à du bruit dans l'image, tel qu'un reflet, soit à un objet de la scène, tel qu'un individu. Une région mobile est caractérisée par des données numériques, appelées *« mesures »*, telles que la largeur de la région. Un *« objet mobile »* représente une entité en mouvement dont on suit les déplacements au cours du temps et dont on analyse le comportement. Il peut être constitué d'une ou plusieurs régions mobiles. Un objet mobile peut aussi bien correspondre à un bruit (p. ex. un reflet que le système considère être un individu), à une partie d'un objet de la scène (p. ex. la tête d'un individu dépassant d'un mur), à un objet de la scène (p. ex. un individu) ou à un groupe d'objets de la scène (p. ex. une foule indissociable). Un objet mobile est caractérisé par des *« propriétés »* numériques comme sa largeur, et symboliques comme l'allure de sa trajectoire. Ces propriétés symboliques correspondent à la perception des mouvements des objets de la scène, sur de faibles intervalles de temps. Les *« scénarios »* sont des données abstraites, définies à partir des propriétés des objets mobiles. Un scénario représente la perception du comportement d'un objet de la scène, sur un intervalle de temps important (p. ex. l'individu louvoie entre des véhicules).

- **Termes empruntés au langage courant permettant de décrire des scènes dynamiques** : les termes suivants décrivent les relations sémantiques, intervenant entre un ou plusieurs objets de la scène. Ces termes n'ont pas de définition formelle et sont à considérer comme des nuances qualifiant ces relations sémantiques. Un *« mouvement »* cor-

respond à un déplacement bref, mesurable et explicable par des lois physiques (p. ex. lever un bras). Un *« événement »* correspond à une modification de l'état d'un objet de la scène, à laquelle on peut attribuer un sens (p. ex. toucher la porte). *« Motivé »* signifie ici que l'objet de la scène réalise l'événement dans le but d'atteindre un objectif. Un *« non événement »* correspond de façon symétrique, à l'absence de changement dans l'état d'un objet de la scène, à laquelle on peut attribuer un sens (p. ex. rester assis). Une *« action »* est une séquence d'événements courte et motivée (p. ex. l'individu tourne). Une *« situation »* correspond à un cliché instantané, faisant intervenir simultanément plusieurs objets de la scène (p. ex. le joueur X est démarqué et le joueur Y possède la balle). Un *« comportement »* est une séquence d'événements longue et motivée (p. ex. l'individu tourne autour du véhicule. Une *« activité »* est un terme générique, pouvant combiner plusieurs comportements faisant intervenir différents objets de la scène (p. ex. les ouvriers déchargent le camion).

A B

FIG. 1.1 – *Ces images montrent une allée entre deux rayonnages. L'image A représente le fond de la scène, contenant l'environnement statique (p. ex. les rayonnages et les panneaux suspendus au plafond). L'image B est une image de la séquence, donnée d'entrée du système. Elle représente un instantané de la scène. Au milieu de l'allée, on peut voir une femme aux commandes d'un véhicule en train de nettoyer le sol.*

L'interprétation de séquences d'images ne possède pas une ontologie faisant l'unanimité parmi les chercheurs du domaine. Cet état de fait est dû principalement à deux raisons. Premièrement, le vocabulaire utilisé n'est pas technique, souvent extrait d'épisodes de la vie courante. Deuxièmement, il existe un grand nombre de termes et une grande diversité d'applications uti-

lisant ces termes. Il est alors difficile de trouver des termes ayant un sens précis et convenant à tous les types d'application.

Malgré les ambiguïtés de sens des termes répandus dans le domaine, les termes utilisés dans le reste de ce mémoire se réfèreront aux définitions de cette section.

1.3 L'interprétation de séquences d'images

Cette section présente plusieurs architectures de système d'interprétation puis propose un modèle d'architecture, qui sera discuté ci-après au chapitre 7.

1.3.1 Présentation de différentes architectures

Dans cette sous-section, nous présentons différentes architectures de système d'interprétation que nous considérons comme complets. Un système complet signifie pour nous que le système prend en entrée des séquences d'images réelles et produit comme résultat une analyse sur des comportements non élémentaires. Bon nombre de systèmes ne sont pas alors considérés comme complets, dans la mesure où ces systèmes mettent l'accent sur la détermination du mouvement et d'actions élémentaires et ne traitent pas le problème de reconnaissance de scénarios non élémentaires. Nous avons ainsi répertorié cinq systèmes complets :

- Le projet Vitra a développé le système d'interprétation *« Soccer »* (André et al., 1988), (Herzog et al., 1989). Ce projet a été mené en Allemagne à l'université de Saarlandes, ainsi qu'à ITTB Karlsruhe, à l'université de Karlsruhe et à GmbH Saarbrücken. *« Soccer »* est composé d'une base de contexte comprenant l'environnement statique, et de trois modules principaux associés à deux modules complémentaires. Les trois modules principaux réalisent la chaîne complète de l'interprétation, du traitement des images jusqu'à l'analyse d'activités. Le premier module *« Action »*, calcule la position et la vitesse de chaque objet mobile. Le second module *« Reconnaissance d'événements »*, produit un ensemble de propositions correspondant à des relations spatio-temporelles relatives à des objets mobiles, telles que *« le joueur A fait une passe au joueur B »*. Le troisième module *« Sélection et Génération »*, choisit les événements pertinents et génère des phrases en allemand décrivant la scène. Les deux modules complémentaires ont pour objectif d'améliorer la sélection des événements pertinents. Le premier module complémentaire *« Replay »* (Retz-Schmidt, 1991), reconnaît

les plans et stratégies des objets mobiles et filtre les événements reconnus, données d'entrée du module de sélection et de génération. Le second module complémentaire *« Antlima »* (Schirra and Stopp, 1993), représente l'état mental d'un auditeur virtuel et permet de sélectionner les événements les plus pertinents pour les utilisateurs. Le système *« Soccer »* est ainsi découpé en deux parties. La première, constituée du module *« Action »*, détecte, reconnaît et suit les objets mobiles. La deuxième partie, constituée des quatre autres modules, reconnaît les actions et génère des phrases décrivant la scène. La seconde partie est composée d'un grand nombre de modules, dûs aux nombreux collaborateurs participant au projet et à leur motivation première de générer des phrases en langage naturel.

- H. Nagel et son équipe ont en Allemagne (ITTB Karlsruhe et l'université de Karlsruhe) développé un système complet reconnaissant le comportement de véhicules dans une station service (Nagel, 1991). Ce système est composé de deux parties, une vision et l'autre interprétation. La partie vision est constituée d'un ensemble de méthodes sophistiquées, permettant de détecter et de suivre les véhicules à partir de leur flot optique. La partie interprétation, plus sommaire, analyse les comportements des véhicules, tels que *« se garer pour prendre de l'essence »*. La disproportion entre ces deux parties est due d'une part, à la provenance de cette équipe d'un laboratoire de vision et d'autre part, à la volonté de réaliser d'abord un système de vision performant et ensuite d'aborder le problème d'interprétation.

- Le projet européen VIEWS a été mené par plus de 17 collaborateurs à l'ITTB Fraunhofer (Atlas Elektronik GmbH), à GEC Hirst Research Centre, à GEC Marconi Reseach Centre, à Marconi Command and Control Systems, à l'université de Reading, au Queen Mary and Westfield College et à FTC (Framentec Cognitech). Ce projet a développé un système d'interprétation constitué de deux parties : le composant perceptuel (partie vision) et le composant conceptuel (partie interprétation) (Corrall, 1992). Ce système est décrit sur la figure 1.2. Le composant perceptuel est constitué de trois modules principaux. Le premier module détecte les régions mobiles et sélectionne l'attention du système sur les régions d'intérêts de l'image. Le second module suit les régions mobiles détectées et le troisième module identifie les objets mobiles à l'aide d'une méthode de classification. Le composant perceptuel signale ainsi au composant conceptuel les changements in-

tervenus dans l'état du monde. Le composant conceptuel est constitué de trois niveaux : événement, comportement et dynamique. Les événements sont des changements considérés comme intéressants (c.-à-d. ils sont reconnus comme appartenant à une classe d'événements intéressante pour l'utilisateur). Les comportements sont des séquences intéressantes d'événements. Les dynamiques sont des ensembles intéressants de comportements, faisant intervenir plusieurs objets mobiles en même temps. Chaque niveau possède trois tâches principales : classer une entité (p. ex. un événement), vérifier sa cohérence avec les résultats déjà obtenus et prédire l'entité suivante (p. ex. le prochain événement). Le niveau des dynamiques n'a pas été complètement achevé. Ce système se singularise par son grand nombre de modules et de fonctionnalités, expliqué en partie par le nombre important de participants à ce projet.

- Le projet européen Esprit HPCN PASSWORDS a été mené par plus de 12 collaborateurs sur 3 ans à l'université de Genova (DIBE), au Research Center CRIF, à l'INRIA Sophia Antipolis et à Sepa (Fiat Research Centre). Ce projet a développé un système d'interprétation composé de trois modules (Bogaert et al., 1996). Le premier module détecte les régions mobiles, le second module suit les régions détectées et le troisième module identifie les objets mobiles et analyse leurs comportements. Chaque module a été développé par des équipes partenaires différentes. Le système est donc structuré par l'enchaînement de trois parties : vision bas-niveau, vision intermédiaire, interprétation.

- Le projet Perception a été mené par plus de 12 collaborateurs en France au CERT DERA, au CERT DERI, à l'ONERA DES-SIA et à l'ONERA DES-STD. Ce projet a développé un système comprenant une base de contexte, deux modules principaux et un module complémentaire (Castel et al., 1996). Le premier module principal *« Traitement numérique »* est le module vision. Il détecte, reconnaît et suit les objets mobiles. Le second module *« Traitement symbolique »* est le module interprétation. Il reconnaît les scénarios relatifs aux activités des objets mobiles. Le module complémentaire *« Gestion de la perception »*, gère les ressources du système, comprenant capteurs et traitements de perception. Ce système est ainsi globalement constitué d'une partie vision et d'une partie interprétation.

Ces systèmes ont tous été développés par des équipes nombreuses au cours de recherches menées à long terme (de 3 à plus de 10 ans). Ils ont

tous le même enchaînement d'opérations : détection et suivi de régions mobiles, identification des objets mobiles et analyse de leurs comportements. De plus, ces systèmes se caractérisent en général, par une séparation marquée en deux parties regroupant les modules vision et les modules interprétation, sans qu'il ait de réelles coopérations entre ces deux parties. Les modules vision regroupent les tâches de détection et de suivi de régions mobiles. Les modules d'interprétation contiennent les tâches de reconnaissance de scénarios élémentaires et complexes. Si la partie vision est suffisamment performante, alors elle contient également la tâche d'identification des objets mobiles. Dans le cas contraire, cette tâche appartient à la partie interprétation. La principale différence entre ces systèmes d'interprétation provient de la proportion des modules vision, par rapport aux modules d'interprétation. Cette différence de proportion est due premièrement à la provenance de l'équipe concevant le système (origine vision plus ou moins marquée) et d'autre part aux objectifs que cette équipe s'est fixée.

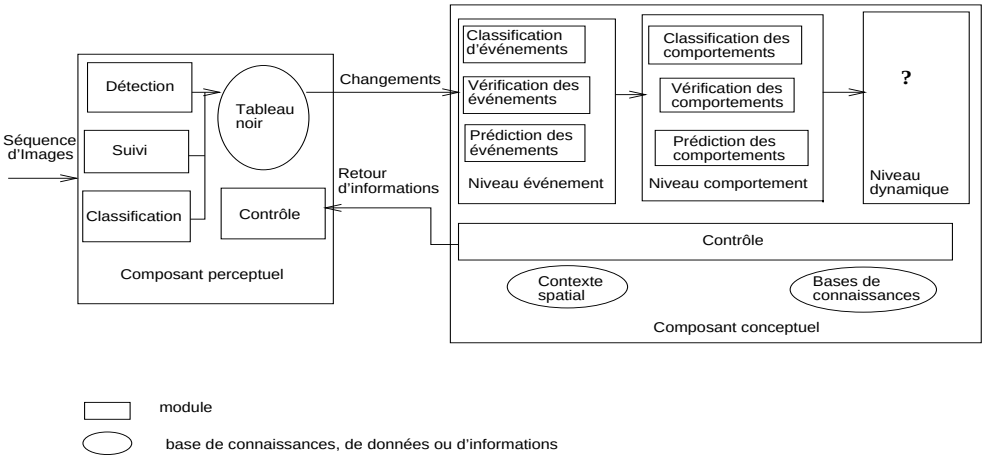

FIG. 1.2 – *Le système d'interprétation du projet VIEWS.*

1.3.2 Approche proposée pour l'interprétation de scènes

De façon comparable aux architectures présentées dans la section précédente, nous proposons un modèle du processus d'interprétation de séquences d'images, basé sur la coopération de trois tâches principales : (1) la détection des régions mobiles à l'aide de traitements d'images bas-niveau, (2) le suivi des régions mobiles à partir des régions détectées et (3) la reconnaissance des scénarios relatifs aux activités des objets mobiles associés aux régions sui-

vies. Ce modèle se caractérise par l'utilisation d'informations contextuelles, contenant en particulier les informations relatives à l'environnement de la scène. Il se caractérise également par la gestion de l'incertitude du processus d'interprétation, facilitant le passage des grandeurs numériques caractérisant les régions mobiles, aux propriétés symboliques servant à reconnaître les scénarios. À partir de ce modèle, nous avons développé un système générique d'interprétation de séquences d'images, constitué d'une base de contexte et de trois modules (un pour chaque tâche principale) : le module de traitement d'images, le module de suivi et le module de reconnaissance des scénarios. La structure de ce système est décrite figure 1.3.

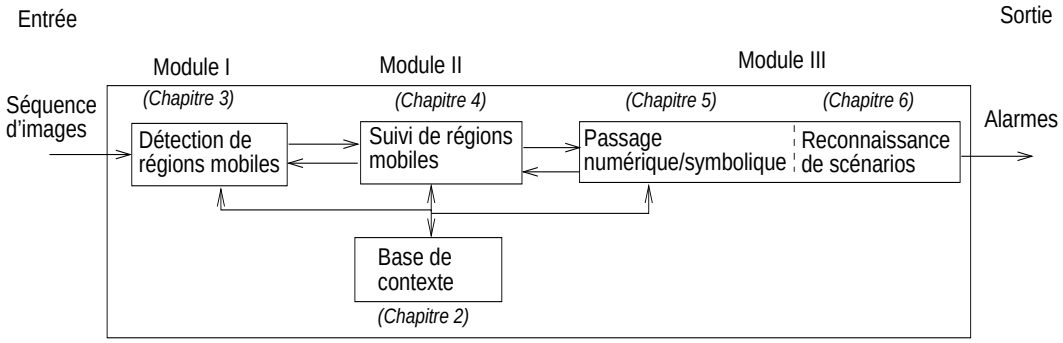

FIG. 1.3 – *Le système d'interprétation de séquences d'images est structuré en une base de contexte et trois modules.*

1.4 Objectifs à atteindre

Le principal objectif de nos travaux est de proposer un modèle générique du processus d'interprétation d'images. Proposer un modèle consiste à étudier différents aspects du processus :

- Il est d'abord nécessaire d'établir les conditions d'utilisation du processus d'interprétation (p. ex. type de caméra, type de scène). Puis il faut déterminer les connaissances *a priori* qui sont indispensables pour que le processus d'interprétation puisse fournir des résultats. Il est également nécessaire de mettre en évidence les parties du processus qui sont dépendantes de l'application. Ces différents points permettent de déterminer le niveau de généricité du processus.

- Il est également important de comprendre la structure du processus d'interprétation et de mettre en évidence les différentes tâches de base

constituant ce processus. Il s'agit alors de préciser les entités manipulées par ces tâches de base, ainsi que leurs caractéristiques.

- Nous nous proposons de déterminer une architecture de l'ensemble, intégrant ces tâches de base. Ce point étudie l'enchaînement des tâches et leurs interactions.

- Enfin, il est nécessaire de préciser les résultats que nous pouvons attendre d'un tel processus. En particulier, ce point revient à énumérer les problèmes restants non résolus et à proposer des perspectives de recherche.

Notre second objectif est de réaliser un système d'interprétation permettant de valider le modèle proposé. Ce système se doit d'être développé en respectant des conditions réelles d'utilisation et se doit d'être suffisamment souple et générique afin d'être extensible et utilisable pour d'autres types d'application.

Par conséquent, nous ne nous intéressons pas ici aux aspects de performance d'un système d'interprétation. Nous souhaitons plutôt déterminer les résultats qu'il est possible d'obtenir en interprétation de séquences d'images et donner les limites de ce processus.

1.5 Plan de lecture

Le mémoire est organisé selon la structure du système d'interprétation, c.-à-d. en suivant le flot de traitement des données. Les chapitres majeurs, contenant les principales contributions, correspondent à chaque composant du système :

Chapitre 2 : le contexte - ce chapitre propose une définition du contexte du processus d'interprétation, une représentation des informations contextuelles, et décrit la construction de la base de contexte utilisée par le système d'interprétation. Ces travaux ont été présentés dans (Brémond and Thonnat, 1996a), (Brémond and Thonnat, 1997b).

Chapitre 4 : le suivi des régions mobiles - ce chapitre propose une méthode générique de suivi, traitant en particulier le cas d'objets non rigides comme par exemple les êtres humains. Il décrit également le module de suivi de régions mobiles. Ces travaux ont été présentés dans (Brémond and Thonnat, 1997e).

Chapitre 5 : le passage du numérique au symbolique - ce chapitre propose une méthode permettant le passage de grandeurs numériques, les mesures calculées sur les régions mobiles, à des propriétés symboliques, les propriétés des objets mobiles utilisées pour la reconnaissance de scénarios. Cette méthode se caractérise par la gestion de l'incertitude relative aux propriétés et aux scénarios. Ce chapitre décrit également le procédé permettant au module de reconnaissance de scénarios d'utiliser les résultats du module de suivi. Ces travaux ont été présentés dans (Brémond and Thonnat, 1996b), (Brémond and Thonnat, 1997c).

Chapitre 6 : la reconnaissance de scénarios - ce chapitre propose un formalisme permettant de décrire des activités humaines et de reconnaître les scénarios associés à ces activités. Ce chapitre décrit également le module de reconnaissance de scénarios. Ces travaux ont été présentés dans (Brémond and Thonnat, 1997a), (Brémond and Thonnat, 1997d).

Le choix de découper ce mémoire en quatre chapitres distincts est motivé par la présence de méthodes appartenant à quatre domaines bien différenciés, ayant peu de parties communes. Dans chacun de ces quatre chapitres, nous commençons par présenter un état de l'art du domaine, puis par proposer le modèle de la méthode étudiée. Nous décrivons ensuite l'implantation de cette méthode dans le système d'interprétation et nous en donnons un exemple d'utilisation. Enfin, nous concluons chaque chapitre en présentant nos contributions et en décrivant les perspectives de recherche. L'état de l'art est ainsi réparti au début de chacun de ces chapitres. De même, les principales contributions concluent ces chapitres.

Les chapitres complémentaires sont :

Chapitre 1 : l'analyse du problème - ce chapitre présente les principes généraux du processus d'interprétation de séquences d'images. Il décrit également nos objectifs et nos motivations pour mener à terme ces travaux.

Chapitre 3 : la détection de mouvement - ce chapitre explique quels ont été nos critères pour choisir un module de détection de régions mobiles. Notre motivation principale est d'obtenir un système d'interprétation opérationnel, prenant en compte les spécifications de nos applications cibles.

Chapitre 7 : le système d'interprétation - ce chapitre propose une architecture pour le système global d'interprétation, présente deux exemples d'utilisation de ce système et décrit ses performances.

Conclusion : cette partie résume nos principales contributions à cette problématique et en expose les perspectives les plus prometteuses.

Chapitre 2

Contexte

La base de contexte comprend l'ensemble des informations contextuelles relatives à l'environnement de la scène et utilisées par le processus d'interprétation de séquences d'images. L'objectif de ce chapitre est de proposer un formalisme permettant de représenter et d'utiliser ces informations contextuelles. Le contexte constitue un ensemble essentiel d'informations pour le système d'interprétation. Il permet d'améliorer le processus global d'interprétation. De plus, ces informations sont souvent indispensables pour résoudre certaines situations délicates. Le contexte étant dépendant de l'application, les problèmes majeurs de son utilisation sont d'abord de définir précisément la notion de contexte et deuxièmement de représenter ces informations afin de faciliter leur acquisition et leur utilisation.

Dans ce chapitre, la section 2.1 présente un état de l'art sur la représentation et l'utilisation du contexte. La section 2.2 propose ensuite une définition du contexte et une méthodologie pour délimiter les informations contextuelles. Puis dans le cas particulier du processus d'interprétation de séquences d'images, la section 2.3 propose une représentation du contexte facilitant son acquisition et son utilisation. Enfin, la section 2.4 présente un logiciel d'acquisition du contexte développé dans le cadre de cette thèse, permettant de construire interactivement une base de contexte.

2.1 État de l'art

Cette section présente différents travaux ayant abordé le problème de modélisation du contexte. Tout d'abord nous montrons que ce problème concerne une grande variété de disciplines. Puis dans le cas de l'interprétation de scène, nous étudions les différents domaines d'utilisation du contexte.

2.1.1 Multi-disciplinairité

L'importance de la notion de contexte est maintenant largement répandue, comme le montre le nombre important d'ateliers scientifiques, de symposiums et de séminaires organisés sur la notion de contexte depuis 1995.

Les **sciences cognitives** ont, depuis longtemps, débattu sur le sens général du contexte dans un processus intellectuel. En **raisonnement symbolique**, différentes formalisations du contexte ont été proposées. Par exemple dans (McCarthy, 1993), l'auteur définit le contexte comme une entité mathématique. Il propose alors différentes règles permettant de rendre un contexte plus générique, ou de munir d'une relation d'ordre les niveaux de contexte. En **linguistique**, le contexte est également une notion centrale permettant de comprendre et de générer des textes en langage naturel. Pour les **systèmes à base de connaissances**, les informations contextuelles facilitent l'adaptation de ces systèmes aux conditions du monde réel. Par exemple dans (Turner, 1995), l'auteur utilise le contexte dans un raisonnement à base de cas. Il a construit un système qui pilote un véhicule sous-marin et qui utilise le contexte pour permettre au véhicule de réagir à des événements imprévisibles. Lorsqu'un événement survient, tel que *« l'arrivée à proximité des côtes »*, le système adapte le pilotage du véhicule en prenant des dispositions, telles que *« réduire la vitesse »*. Le principal problème soulevé par cet auteur est la gestion du contexte par le système. Enfin en **vision par ordinateur**, de nombreux systèmes utilisent le contexte. Par exemple dans (Strat, 1993), l'auteur énumère différents systèmes en reconnaissance d'objets utilisant fortement le contexte. L'auteur est en particulier confronté aux problèmes de représentation du contexte, afin que les systèmes de reconnaissance puissent l'utiliser de manière systématique (Strat and Fischler, 1990).

Pour toutes ces disciplines, des questions similaires se posent : comment définir le contexte, comment le générer et comment l'utiliser? Bien qu'aucune solution n'émerge réellement, la confrontation multi-disciplinaire des problèmes de chacun ne peut qu'aider à résoudre les problèmes fondamentaux de la modélisation et de l'utilisation du contexte. Dans les sous-sections suivantes, on présente dans le cas particulier de l'interprétation de scènes différentes approches permettant d'utiliser le contexte.

2.1.2 Contexte local

En interprétation de scènes, une première approche consiste à utiliser le contexte local aux objets mobiles détectés dans la scène.

On définit alors un contexte propre à chacun des objets détectés et on

utilise ce contexte afin de suivre avec plus de confiance l'objet mobile en question. Par exemple dans (Intille and Bobick, 1995), les auteurs suivent des joueurs de football en s'aidant de contextes locaux à chaque joueur, appelés *« mondes clos »*. Ces contextes correspondent à la portion de l'image entourant le joueur. Ils contiennent la distribution de couleurs de l'environnement à proximité du joueur permettant de prendre en compte les objets statiques environnant, tels que des *« marquages au sol »*, au niveau de l'algorithme de suivi. De même dans (Choi et al., 1997), les auteurs utilisent un histogramme de couleurs sur le joueur et son environnement proche afin d'améliorer le suivi du joueur et de traiter, en particulier, les situations d'occultation par d'autres joueurs. Dans (P. Remagnino and Kittler, 1993), les auteurs utilisent également un contexte local, appelé contexte spatio-temporel, contenant la durée de vie et la position des objets détectés. Ces contextes permettent d'améliorer le contrôle et le pilotage de la caméra mobile, servant à la perception de la scène.

Une seconde utilisation des contextes locaux consiste à améliorer la reconnaissance des actions, accomplies par les objets mobiles de la scène. Par exemple dans (Nagel, 1988), l'auteur définit un contexte comme étant une description générique associant les aspects spatio-temporels et intentionnels de véhicules évoluant dans des scènes routières. L'auteur compte utiliser cette notion de contexte afin de reconnaître des actions, telles que *« le véhicule se gare »*. Cependant, il ne décrit pas comment représenter et utiliser ces contextes.

Les informations contextuelles locales les plus utilisées sont ainsi des informations numériques, renseignant sur l'environnement proche des objets mobiles. Les travaux de H. Nagel ont essayé d'étendre cette notion de contexte local à des informations plus symboliques.

2.1.3 Capteur mobile et multi-capteurs

Les systèmes d'interprétation multi-capteurs ou munis d'un capteur mobile ont généralement besoin d'informations contextuelles pour gérer le plus efficacement possible le ou les capteurs.

Dans (Ghallab et al., 1992), les auteurs s'intéressent à la modélisation des capteurs et des traitements d'acquisition des données perçues. Ils ont construit un robot mobile muni de nombreux capteurs et capable d'analyser des séquences d'images. L'objectif du robot est d'analyser son environnement, comprenant des objets mobiles. Le mot capteur est à prendre au sens large. Il peut aussi bien représenter un capteur physique (p. ex. un laser 3D), qu'un algorithme (p. ex. un extracteur de régions colorées). La description

d'un capteur est donnée par l'ensemble de trois modèles : le modèle structurel, le modèle d'état et le modèle perceptuel. Le modèle structurel définit le type de données, que le capteur produit et peut être récursivement décrit à l'aide de modèles structurels de capteurs de plus bas niveau. Le modèle d'état est un ensemble de variables caractérisant le capteur, comme la position ou la longueur de la focale d'une caméra. Le modèle perceptuel décrit les relations entre les données produites par le capteur et leur signification dans la scène. Par exemple, ce modèle donne la probabilité de détection d'une primitive déterminée par des programmes de traitement d'images, telle qu'un segment 2D. Ces informations contextuelles permettent alors de gérer les ressources et de contrôler les capteurs, de planifier les programmes de traitement d'images et de prendre en compte la géométrie et la fiabilité des capteurs.

De même dans (Clement et al., 1993), les auteurs ont développé un système multi-capteurs afin d'analyser une scène observée par un satellite. Ils adaptent alors la reconnaissance des objets de la scène (p. ex. des ponts), aux capteurs employés. Ces informations contextuelles permettent essentiellement d'adapter les données perçues par les capteurs, au contexte de la scène.

2.1.4 Carte de la scène

En interprétation de scène, le contexte le plus largement utilisé est une carte (c.-à-d. un modèle spatial) de la scène. Nous présentons ci-dessous, différents systèmes utilisant des informations contextuelles, relatives à la structure spatiale de la scène :

- Tout d'abord dans (Neumann, 1984), (Mohnhaupt and Neumann, 1990), les auteurs ont développé le système Naos qui, à partir d'une séquence d'images, décrit une scène routière à un auditeur, qui est dans l'impossibilité de voir la scène. Naos possède une description géométrique de la scène contenant certaines propriétés photométriques (couleur, illumination, ...). Cette description correspond à une carte 2D de la scène, composée d'un ensemble de cellules, qui est structuré sous la forme d'une grille. Cette représentation de l'espace est dite analogique, car ces cellules correspondent explicitement à la structure intrinsèque de la scène. Pour cette raison, Naos raisonne directement sur la carte 2D, plutôt que sur la séquence d'images. Les relations spatio-temporelles sont alors calculées localement au niveau des cellules puis globalement par propagation. Par exemple, la distance entre les véhicules A et B se calcule par l'envoi d'un signal de la part de A à ses cellules voisines, afin

de les interroger sur la présence de B. Naos utilise ainsi le contexte, afin d'améliorer le calcul des relations spatio-temporelles et surtout comme support de raisonnement du processus d'interprétation.

- De même dans (Tsuji and Li, 1993), les auteurs ont construit une carte représentant la vue panoramique d'un environnement extérieur, correspondant à une route bordée de nombreux bâtiments. Pour condenser l'information, ils ne conservent sur la carte que les monuments marquants, appelés points de repère. Les auteurs utilisent ces informations contextuelles, afin qu'un robot naviguant dans la scène puisse se repérer.

- Dans (Sellam and Boulmakoul, 1994), (Cerf and Pintado, 1997), les auteurs utilisent des modèles sophistiqués de la scène, correspondant à un ensemble connexe de carrefours, afin de gérer la circulation de piétons et de véhicules. Ces modèles contiennent des informations topologiques, telles que les zones d'entrée, de jonctions, de stockage et de conflit des objets mobiles. Ils contiennent également des informations sur les signalisations des carrefours, telles que les phases du trafic, l'état et la localisation des feux rouges, des feux pour piétons. Ces informations permettent alors de suivre les objets mobiles et de déterminer les flux de la circulation.

- Dans (Bobick and Davis, 1996), les auteurs ont développé un système qui contrôle des caméras dans un studio de télévision. Ce système se caractérise par un modèle approximatif du monde, qui est mis à jour tout au long du traitement. Ce modèle contient la structure spatiale de la scène, avec la position et les angles de vue des caméras. Il contient également un ensemble de règles (post et pré-conditions) et permettant de piloter des routines de traitement d'images. Un exemple de règle est : SI *« l'objet mobile est dans la région centrale de l'image »* (pré-condition), ALORS *« extraire la région mobile au centre »* (action) ET *« l'objet et la région mobile doivent avoir la même surface »* (post-condition). Les auteurs utilisent ainsi un modèle approximatif du monde (carte 3D grossière) et des informations symboliques permettant l'utilisation de ce modèle.

- Le projet VIEWS a pour objectif la surveillance de scènes extérieures en temps réel, à partir de l'analyse de séquences d'images (Duong et al., 1990a), (Duong et al., 1990b). Cette équipe reprend la représentation analogique de la scène proposée par B. Neumann, en la structurant

dans une hiérarchie arborescente de cellules (Howarth and Buxton, 1992a), (Howarth and Buxton, 1992b). Cette nouvelle structure permet au projet VIEWS de définir des zones à différents niveaux d'abstraction, incluant des informations symboliques relatives à ces zones. Par exemple, la zone *« route »* possède une zone fille *« feu rouge »*, indiquant que l'arrêt d'un véhicule dans cette zone peut être dû à la présence d'un feu rouge. Ces informations symboliques sont ensuite utilisées comme contraintes, pour vérifier la cohérence du système à tous les niveaux du traitement. Dans le projet VIEWS, les informations contextuelles sont ainsi très largement utilisées, jusqu'au niveau plus abstrait de la reconnaissance des comportements.

Ces informations contextuelles sont principalement utilisées afin d'adapter les données perçues à leur localisation dans l'environnement. Elles sont représentées sous la forme d'une carte de la scène, et sont ainsi largement utilisées depuis les niveaux les plus élémentaires (p. ex. le calcul des relations spatiales), jusqu'à des niveaux plus abstraits (p. ex. l'analyse de comportements).

2.1.5 Construction de contexte

En vision par ordinateur, il existe de nombreux systèmes permettant de construire une partie du contexte relativement à une scène statique donnée. Ces systèmes ont en général pour objectif de reconstruire les structures spatiales de la scène. Dans (Milhaud and Médioni, 1994), (Nevatia and Médioni, 1996), les auteurs présentent différents projets de reconstruction d'une scène statique. En particulier, ils proposent de construire le modèle d'un site, tel que *« une usine »*, à partir d'images aériennes. Dans ce but, ils extraient d'une image des segments de droites, qu'ils regroupent afin de construire le modèle des bâtiments.

Ces systèmes montrent que, sous certaines conditions, il est possible d'acquérir les informations spatiales et structurelles d'une scène statique. Cependant, en interprétation de scènes dynamiques, peu de systèmes possèdent une phase de pré-traitement, dédiée à la construction du contexte de la scène. Dans le meilleur des cas, cette construction se limite à l'établissement d'une carte approximative de la scène.

2.2 Modèle de la base de contexte proposé

D'une manière générale, la communauté scientifique reconnaît l'importance du contexte et son influence sur la qualité des résultats du processus d'interprétation. De nombreux ateliers scientifiques ont été organisés dans ce sens. Par contre, la manière d'utiliser le contexte reste un problème, dû en particulier à la difficulté à formaliser cette notion. Dans cette section, nous commençons par proposer une définition du contexte puis nous appliquons cette définition au cas particulier du processus d'interprétation de séquences d'images. Nous obtenons alors deux règles permettant de délimiter plus précisément les informations contextuelles. Enfin, cette section se termine en abordant les problèmes posés par la représentation du contexte.

2.2.1 Définition du contexte

La définition du contexte d'un processus dépend de la nature du processus. Pour H. Nagel (Nagel, 1988), le contexte du processus de reconnaissance d'actions est une structure complexe comprenant des descriptions génériques de l'espace, l'évolution temporelle de ses structures et l'intention supposée de l'action. Pour T. Strat (Strat, 1993), le contexte du processus d'analyse d'images statiques est, dans son sens le plus large, n'importe quelle information qui peut influencer la manière dont la scène est perçue. Plus généralement, un processus utilise trois types d'informations : les connaissances principales, les informations contextuelles et les informations factuelles. Les **connaissances principales** sont toujours valides, elles sont directement connectées aux objectifs du processus et font souvent partie d'un modèle bien défini. Si une connaissance fait défaut, le processus n'est plus alors capable d'inférer des résultats. Les **informations contextuelles** dépendent de l'application, mais elles restent constantes pendant le traitement. Ce sont des informations secondaires, qui peuvent devenir essentielles pour résoudre des situations particulières. Les informations contextuelles ont pour principal objectif d'améliorer le traitement global du processus. Elles constituent les informations supplémentaires dont le processus a besoin pour fonctionner efficacement. Les **informations factuelles** dépendent de l'état d'exécution du processus. Leur durée d'existence est souvent courte. Elles correspondent aux données d'entrée et aux données calculées.

Nous proposons alors de définir les informations contextuelles d'un processus comme les informations vérifiant deux conditions :

(1) leurs valeurs restent constantes pendant l'exécution du processus.
(2) leurs valeurs sont différentes lorsque le processus est utilisé pour une autre application. (2.1)

Cette définition du contexte a deux conséquences principales. Tout d'abord, elle impose de préciser à partir de quel niveau de traitement on considère une information comme factuelle plutôt que contextuelle. Deuxièmement, elle impose de préciser à partir de quel niveau d'abstraction de l'application on considère une information comme contextuelle plutôt qu'appartenant aux connaissances principales. Ces deux points seront discutés ci-dessous dans le cas du processus d'interprétation.

Les difficultés pour formaliser le contexte proviennent ainsi, d'une part de la dépendance de ces informations au domaine d'application, et d'autre part de la délimitation floue des frontières séparant le contexte des autres types d'informations. Pour ces raisons, il n'existe pas dans la littérature de définition formelle du contexte. Cependant, cette définition est nécessaire dès que l'on souhaite rationaliser l'utilisation du contexte.

2.2.2 Contexte et interprétation de séquences d'images

La définition 2.1 indique que, pour définir le contexte du processus d'interprétation de séquences d'images, il est nécessaire de préciser son niveau de granularité de traitement. Comme exposé au chapitre 1, ce processus peut être décomposé en trois tâches principales : détection de régions mobiles, suivi de régions mobiles et reconnaissance de scénarios. De plus, les tâches de suivi de régions mobiles et de reconnaissance de scénarios comportent une partie commune de raisonnement spatial, qui peut être également vue comme une tâche de base. Pour cette raison, nous considérons que le processus global d'interprétation est composé de quatre tâches spécifiques.

Le processus d'interprétation peut alors être considéré à deux niveaux différents de granularité de traitement : au niveau du processus global ou au niveau des quatre tâches. Nous choisissons de définir le contexte au niveau de granularité des tâches car contrairement au processus global, les quatre tâches possèdent des domaines de connaissances bien délimités et bien définis. Nous pouvons alors décrire le contexte du processus global d'interprétation à l'aide de la formule ci-dessous (*« tâche-i »* représente n'importe quelle tâche du processus d'interprétation) :

$$\text{contexte[processus global]} = \bigcup_i \text{contexte[tâche-i]} \tag{2.2}$$

Cette formule fournit alors une règle permettant de déterminer si une information vérifie la première condition de la définition 2.1 : si elle reste constante pendant l'exécution d'au moins une tâche qui l'utilise, alors l'information n'est pas factuelle.

2.2.3 Domaines source du contexte

Pour définir le contexte du processus d'interprétation de séquences d'images, la définition 2.1 indique également qu'il est nécessaire de définir le niveau d'abstraction de l'application. Ce niveau d'abstraction permet de préciser les domaines d'informations dont le processus d'interprétation dépend ; c'est-à-dire les informations qui sont modifiées lorsque l'on change d'application. Par exemple, le processus d'interprétation peut s'appliquer à un niveau général de la vidéo-surveillance (comme la surveillance du réseau global d'un métro), à un niveau de granularité intermédiaire (comme la surveillance d'une station de métro), ou à un niveau plus spécifique (comme la surveillance particulière d'un quai, observé par une caméra à une position donnée). La délimitation des informations contextuelles dépend alors du niveau d'abstraction choisi. Dans ce mémoire, nous ne nous sommes intéressés qu'à un certain nombre d'applications cibles (voir chapitre 1). Cependant comme le suggère (Strat, 1993), nous pouvons inventorier, pour la plupart des applications utilisatrices de contexte, quatre domaines source d'informations contextuelles :

- **Informations sur l'Environnement de la Scène** (IES) : ce domaine comprend les structures spatiales de la scène (représentées par exemple sous la forme d'une carte de la scène), des plans de calibration, les objets statiques (piliers, escalators), les caractéristiques optiques de la scène (réflexions au sol, occultations) et ses caractéristiques comportementales (zone de sortie, chemin).

- **Informations liées à l'Acquisition d'Images** (IAI) : ce domaine comprend les caractéristiques de la caméra (son type, sa longueur focale), les caractéristiques de l'image (taille et type de l'image, date d'acquisition) et les caractéristiques de l'acquisition (orientation et position de la caméra).

- **Informations Évoluant dans le Temps** (IET) : ce domaine rassemble les résultats de précédentes exécutions de tâches d'interprétation. Ces informations peuvent être vues comme les informations accumulées sur le passé (la liste des objets mobiles déjà détectés et

identifiés), et les informations prédites sur le futur (les intentions des objets mobiles).

- **Informations liées aux Requêtes de l'Utilisateur** (IRU): d'une manière interactive, un opérateur humain peut fournir des informations contextuelles pendant tout le déroulement du processus d'interprétation. Par exemple, une requête typique est la demande de surveillance de personnes ayant une caractéristique particulière, telle que la recherche d'un homme ayant une valise.

Le processus d'interprétation dépend alors de tous ces domaines source de l'information contextuelle. Cet état de dépendance fournit une règle permettant de décider si une information vérifie la deuxième condition de la définition 2.1 : si l'information appartient à un de ces domaines source, alors elle n'appartient pas au domaine de connaissances principales du processus d'interprétation.

2.2.4 Base de contexte multi-points de vue

À la lumière de cette définition et de ces règles, plusieurs problèmes se posent quant à la représentation et à l'utilisation du contexte.

Représentation de l'espace

Un premier problème est dû à l'étendue des domaines couverts par le contexte. Le système global d'interprétation est composé de plusieurs modules, tous utilisateurs de la base de contexte. Les informations contextuelles sont ainsi dispersées dans tout le système, rendant difficile une représentation centralisée et commune du contexte. Un autre problème, complémentaire du précédent, est provoqué par le recouvrement partiel des tâches du processus d'interprétation. Certaines informations contextuelles peuvent alors appartenir au contexte de plusieurs tâches. Ce cas survient principalement avec les informations provenant du domaine source IET (Informations Évoluant dans le Temps). Par exemple, le degré d'intérêt d'un objet mobile est calculé et utilisé par la tâche de reconnaissance des scénarios, mais cette information est également utilisée par les tâches de détection et de suivi de régions mobiles, afin de sélectionner sur l'image suivante les régions d'intérêt. Pour cette raison, il est nécessaire que la base de contexte soit partagée par les différentes tâches du processus d'interprétation. Une manière de résoudre ce problème est d'utiliser une représentation centralisée et uniforme pour

toutes les tâches. Nous proposons alors d'utiliser la représentation de l'espace comme support pour représenter le contexte, et ceci pour deux raisons :

- le contexte du processus d'interprétation est composé principalement d'informations provenant du domaine source IES (Informations sur l'Environnement de la Scène);

- le raisonnement spatial est une tâche de base du processus d'interprétation.

L'idée d'utiliser la représentation de l'espace comme support de représentation du contexte, a déjà été utilisée dans différents travaux. Par exemple, dans (Mohnhaupt and Neumann, 1990), les auteurs utilisent une décomposition de l'espace pour améliorer le calcul des propriétés spatiales et pour représenter les trajectoires typiques des objets mobiles. Dans (Howarth and Buxton, 1992a), les auteurs ont repris ces travaux de B. Neumann, et ont ajouté à la décomposition de l'espace différentes caractéristiques sur les comportements des objets mobiles.

La base de contexte peut alors servir de support de raisonnement du processus d'interprétation. Sa structure est accessible par toutes les tâches du processus et permet de centraliser les résultats de chacune des tâches. La base de contexte est dite analogique, terme suggéré par B. Neumann, car sa structure correspond explicitement à la structure intrinsèque de la scène.

Acquisition du contexte

Le problème majeur du contexte est sa phase d'acquisition. La construction de la base de contexte est rendue particulièrement difficile en raison de l'importance du nombre d'informations à générer et de leur diversité. Ce problème devient crucial dès que la taille du système d'interprétation est importante. Comme le contexte est dépendant de l'application, cette phase d'acquisition doit être menée pour chaque nouvelle scène traitée. Dans les applications cibles abordées (se rapporter au chapitre 1), le contexte est généré par des opérateurs humains. Comme le processus d'interprétation est composé de quatre tâches principales, quatre experts sont nécessaires pour construire la base de contexte (un expert par tâche). La phase d'acquisition du contexte est ainsi particulièrement pénible et coûteuse en temps, limitant l'utilisation des informations contextuelles pour bon nombre de systèmes d'interprétation.

Représentation multi-points de vue

Une première solution pour faciliter la phase d'acquisition est d'organiser la base de contexte. Pour cela, nous proposons d'utiliser une représentation multi-points de vue de la base de contexte. Nous utilisons un point de vue pour chacune des tâches du processus d'interprétation. Un point de vue est un filtre permettant de ne voir, comme informations contextuelles, que celles relatives à la tâche associée au filtre. Bien que la base de contexte soit rassemblée en un lieu unique à l'aide de la représentation de l'espace, les points de vue permettent de structurer ainsi l'information.

2.2.5 Réutilisation de bases de contexte

Une seconde solution pour faciliter la phase d'acquisition consiste à réutiliser des bases de contexte déjà acquises, dans le but de développer de nouvelles applications. L'intérêt de réutiliser des bases de contexte est d'autant plus important que le niveau d'abstraction de l'application est élevé, c'est-à-dire que le nombre d'informations contextuelles est important. En ce qui concerne le processus d'interprétation, c'est la quantité d'informations contenues dans le domaine source IES (Informations sur l'Environnement de la Scène), qui rend important l'enjeu de réutiliser le contexte.

Dans ce but, nous proposons d'abstraire les informations contextuelles et de n'utiliser essentiellement que des informations symboliques : plutôt que d'utiliser une description complète du contexte, on n'utilise que le symbole représentant cette description. La description complète est alors prédéfinie dans des bibliothèques, associées au système d'interprétation. Par exemple, le contexte lié à un objet statique est représenté par le nom de l'objet et par sa position, sa description complète étant définie dans une bibliothèque. Une partie du contexte du processus d'interprétation est alors générique (les descriptions) et peut être utilisée par différentes applications. Cependant, malgré cette utilisation d'informations symboliques, la phase d'acquisition reste toujours coûteuse en temps. Pour cette raison, nous avons développé un logiciel graphique d'acquisition du contexte qui est présenté ci-après dans la section 2.4.

Les solutions proposées afin de faciliter la phase d'acquisition sont des solutions essentiellement logicielles. Elles montrent que l'utilisation du contexte est possible, mais qu'elle n'est pas gratuite.

2.3 Représentation et utilisation d'une base de contexte

Cette section a pour objectif d'expliquer comment nous représentons le contexte du processus d'interprétation. Tout d'abord, nous décrivons comment la représentation de l'espace est utilisée comme support de représentation de toute la base de contexte. Ensuite, pour chaque tâche du processus d'interprétation, nous décrivons comment le contexte est représenté et est utilisé.

2.3.1 Décomposition de l'espace

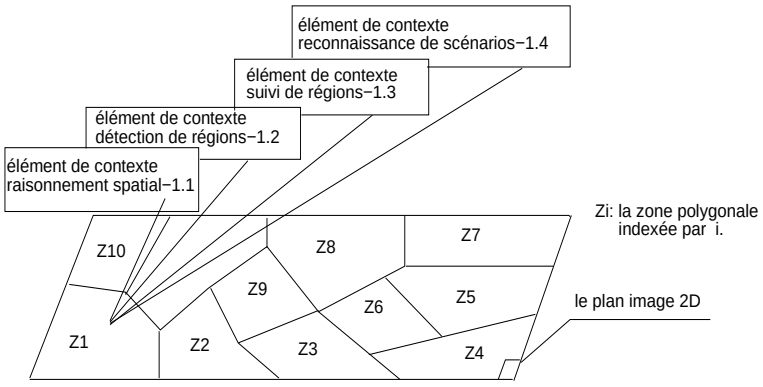

FIG. 2.1 – *Un élément de contexte, défini pour chacune des tâches du processus d'interprétation, est relié à chaque zone polygonale de la décomposition de l'espace.*

Comme l'explique la section précédente, nous proposons de représenter la base de contexte à l'aide de la représentation de l'espace. Dans notre système d'interprétation, l'espace correspond à la projection perspective de la scène 3D sur le plan image 2D. Nous proposons alors de représenter l'espace à travers la décomposition du plan image en une partition de zones délimitées par des polygones. Nous avons choisi des polygones car c'est une structure suffisamment simple pour être implantée et suffisamment flexible pour délimiter des informations générées par un être humain. Cette décomposition de l'espace en zones polygonales est dessinée, lors d'une phase de pré-traitement, par un opérateur humain à l'aide du logiciel d'acquisition du contexte présenté ci-après dans la section 2.4. Comme le montre la figure 2.1, chaque zone est reliée à quatre éléments de contexte ; on appelle **élément de contexte** une portion des informations contextuelles associée

à une zone et à une tâche du processus d'interprétation. La zone polygonale délimite alors l'emplacement où les informations contenues dans les éléments de contexte sont utiles. La figure 2.1 montre la zone polygonale Z1 avec ses quatre éléments de contexte correspondants, notés *« élément de contexte du raisonnement spatial-1.1 »* ... jusqu'à *« élément de contexte de la reconnaissance de scénarios-1.4 »*. Par exemple, la tâche de suivi de régions mobiles utilise *« élément contexte du suivi de régions-1.3 »* pour améliorer le suivi des régions mobiles, détectées dans la zone Z1 pendant l'exécution du processus d'interprétation. La décomposition de l'espace sert ainsi de support pour représenter et centraliser tous les éléments de contexte.

Un élément de contexte est comme son nom l'indique, un sous-ensemble d'informations contextuelles. Il est représenté à l'aide du formalisme de *« frames »*, où chaque attribut correspond à une propriété du contexte. Il existe quatre classes d'élément de contexte, une pour chaque tâche du processus d'interprétation. La structure d'un élément de contexte dépend alors de la tâche et sa valeur de la zone associée. La base de contexte est l'ensemble de tous ces éléments de contexte. Elle vérifie les spécifications requises (décrites dans la section 2.2) facilitant l'acquisition et l'utilisation du contexte :

- La base de contexte est centralisée en un seul lieu, prenant comme support la décomposition de l'espace. La base est alors qualifiée d'analogique.

- Le contexte de chaque tâche du processus d'interprétation est représenté à l'aide d'un formalisme commun, le formalisme de *« frames »*.

- Les informations contextuelles appartenant à un point de vue, associé au contexte d'une tâche du processus d'interprétation, sont facilement filtrées à l'aide de la classe correspondante des éléments de contexte. La base est alors appelée base multi-points de vue.

- La représentation du contexte permet l'utilisation d'informations symboliques, puisque les attributs des éléments de contexte peuvent être définis comme des symboles. Les informations contenues dans la base de contexte sont alors réutilisables.

Cette représentation de la base de contexte facilite ainsi l'utilisation des informations contextuelles.

2.3.2 Contexte dans le raisonnement spatial

Utilité du contexte

Une obligation pour tout système d'interprétation consiste à calculer les relations spatiales intervenant entre les objets mobiles. Les informations contextuelles du domaine source IES (Informations sur l'Environnement de la Scène), et plus particulièrement la décomposition de l'espace en zones polygonales, permettent d'améliorer ces calculs. Effectivement, en indexant les objets mobiles à l'aide de la décomposition de l'espace (on relie chaque objet à la zone qu'il occupe), on facilite le calcul de plusieurs relations spatiales. Par exemple, on peut calculer la densité d'objets mobiles par zone. Comme le suggère (Mohnhaupt and Neumann, 1990), on peut également calculer les objets mobiles proches d'un objet donné, en comptant le nombre d'objets mobiles indexés par la zone associée à l'objet donné. Cette technique permet de réduire le temps de calcul de cette propriété spatiale de base, réduisant ainsi de façon sensible le temps de traitement du système d'interprétation. De plus, la décomposition de l'espace permet de raisonner à différents niveaux d'échelle spatiale. Pour cela, on utilise plusieurs décompositions de l'espace, utilisant une taille de zone polygonale importante pour raisonner à un niveau grossier, ou une taille réduite pour raisonner à un niveau plus fin. Dans (Sol, 1997), l'auteur utilise cette technique, codée à l'aide de la géométrie de Peano, afin de décrire les comportements des objets mobiles à différents niveaux d'échelle spatiale. L'indexation des objets mobiles à l'aide de la décomposition de l'espace permet ainsi d'utiliser le contexte comme support de raisonnement.

La seconde utilisation des informations contextuelles de l'environnement de la scène concerne le calcul des propriétés 3D des objets mobiles. Ces propriétés sont nécessaires afin d'obtenir des propriétés spatiales, rendues imprécises à cause de la profondeur de la scène, ou de la superposition de la perception de plusieurs objets. Les informations contextuelles nécessaires à ces calculs sont les matrices de calibration de la scène, permettant le passage des coordonnées 2D du plan image, aux coordonnées 3D de la scène et inversement (Robert, 1993).

Représentation du contexte

Dans le système d'interprétation proposé nous définissons l'élément de contexte d'une zone donnée, associé à la tâche de raisonnement spatial, à

l'aide de six attributs :

- La liste des objets mobiles indexés par la zone donnée. Cette liste est mise à jour par le module de reconnaissance de scénarios.
- La liste des objets statiques de l'environnement contenus dans la zone, tels qu'un siège, un pilier. Le volume occupé par ces objets est approximé à l'aide de parallélépipèdes 3D, dont les dimensions sont saisies interactivement par un opérateur humain, à l'aide du logiciel graphique présenté dans la section 2.4.
- Les références aux zones polygonales voisines, permettant d'améliorer le calcul des propriétés spatiales 2D des objets mobiles. Cet attribut permet de calculer les objets de la scène, voisins d'un objet mobile donné.
- La matrice de passage des coordonnées 3D de la scène aux coordonnées 2D du plan image, ainsi que le coefficient de précision de cette matrice. Cette matrice est une matrice 4x3, calculée pour toute la scène. Nous la calculons à partir de la donnée d'au moins six points non coplanaires, dont les coordonnées 2D du plan image et les coordonnées 3D de la scène sont connus. Nous estimons les coefficients de la matrice à l'aide de la méthode des moindres carrés. Le coefficient de précision de la matrice permet de quantifier la précision des coordonnées 2D, obtenues à l'aide des coordonnées 3D. Il est calculé à partir de la donnée de points supplémentaires, connus dans les deux systèmes de coordonnées.
- La matrice de calibration permettant le passage des coordonnées 2D, aux coordonnées 3D et inversement, ainsi que le coefficient de précision de cette matrice. Cette matrice n'est définie qu'à la condition que la zone donnée soit contenue dans un plan remarquable. Un plan remarquable est un plan dont on connaît au moins quatre points non alignés, avec leurs coordonnées 2D du plan image et leurs coordonnées 3D de la scène. Les différents niveaux du sol ou les murs, sont des exemples de plans remarquables. Cette matrice de calibration est définie pour tous les points d'un plan remarquable et permet pour tous ces points, de calculer leurs coordonnées 3D à partir de leurs coordonnées 2D. C'est une matrice inversible 3x3. Le coefficient de précision de la matrice permet de quantifier la précision du changement de coordonnées. Nous avons développé ces techniques de calibration en nous inspirant des travaux réalisés dans le cadre du projet européen Esprit HPCN PASSWORDS (Bogaert et al., 1996).

– Les points 2D délimitant le contour polygonal de la zone.

Le contenu de cet élément de contexte permet alors d'améliorer les performances de la tâche de raisonnement spatial.

2.3.3 Contexte dans la détection des régions mobiles

Utilité du contexte

La détection de mouvements et, de façon plus générale, le traitement d'images est une tâche complexe pour différentes raisons (se référer au chapitre 3). En traitement d'images, un simple algorithme peut rarement résoudre seul un problème donné. Plusieurs étapes sont souvent nécessaires pour calculer les résultats finaux, et chaque étape peut être réalisée de différentes manières. Les programmes de traitement d'images ont ainsi besoin de systèmes les pilotant (Shekhar et al., 1994), (Thonnat et al., 1994), (Strat, 1993). Ils ont besoin d'être sélectionnés, ordonnancés et reliés les uns aux autres. De plus, il est souvent nécessaire d'ajuster les paramètres des traitements d'images pendant leur exécution, afin d'obtenir des résultats satisfaisants en fin de traitement. Enfin, la plupart des systèmes d'interprétation doivent fonctionner en temps réel, rendant alors nécessaire la sélection de l'attention du système sur les régions d'intérêt des images (Buxton and Gong, 1995), (Howarth and Buxton, 1993). Pour piloter des traitements d'images, les systèmes utilisent traditionnellement les informations contextuelles provenant du domaine source IAI (Informations d'Acquisition d'Images). Ces informations permettent par exemple, de fixer, d'adapter les paramètres des algorithmes de traitement d'images et de faciliter la planification et le contrôle de ces traitements, afin de les utiliser dans les conditions pour lesquelles ils ont été conçus. Les informations contextuelles du domaine source de l'environnement de la scène sont également utilisées. Elles servent essentiellement à sélectionner l'attention du système sur les régions d'intérêt (*« Region Of Interest »*, en anglais). Toutes ces informations contextuelles permettent ainsi d'améliorer les performances de la tâche de détection de régions mobiles.

Représentation du contexte

En ce qui concerne le système que nous proposons, nous utilisons uniquement les informations contextuelles du domaine source IAI (Informations d'Acquisition d'Images). Effectivement, les informations contextuelles du domaine source IES (Informations sur l'Environnement de la Scène) sont principalement utilisées pour améliorer le temps de traitement du système

d'interprétation. Notre objectif étant de valider notre approche, et non de réaliser un système véritablement temps réel, nous laissons cette utilisation du contexte pour des travaux futurs. Nous avons ainsi défini un seul élément de contexte associé à la tâche de détection de régions mobiles, et relié à toutes les zones polygonales de la décomposition de l'espace. Cet élément de contexte contient des attributs globaux tels que :

- La taille de l'image (p. ex. 512x512), son type (p. ex. image couleur), son format (ppm).
- La cadence d'acquisition des images (p. ex. 4 images par seconde).
- Le taille minimale d'une région mobile (p. ex. 10 pixels).
- Les seuils indiquant, pour chaque couleur, à partir de quelle valeur la différence entre l'image courante et l'image de référence est significative.

Ces attributs sont décrits plus en détail dans la section 3.2.

2.3.4 Contexte dans le suivi de régions mobiles

Utilité du contexte

La tâche de suivi de régions mobiles est décrite dans le chapitre 4. Ses principaux problèmes proviennent d'erreurs dans la détection du mouvement, dues essentiellement à des irrégularités optiques, telles que des réflexions sur le sol, des ombres, des zones encombrées, des lumières clignotantes, des occultations, des zones faiblement contrastées (se référer à la section 4.2). Une solution pour résoudre ces problèmes consiste à utiliser les informations contextuelles du domaine source IES (Informations sur l'Environnement de la Scène). Premièrement, ces informations permettent d'indiquer à la tâche de suivi de régions mobiles l'existence et la localisation de ces irrégularités optiques. La tâche de suivi peut alors appliquer un traitement spécifique aux régions mobiles localisées dans les zones possédant un tel contexte. Par exemple, la tâche de suivi peut connaître à l'avance, la zone de réapparition d'une région mobile occultée par un obstacle. Deuxièmement, la tâche de suivi peut utiliser une incertitude *a priori*, définie comme une information contextuelle, qualifiant les résultats du suivi. Par exemple, une région mobile détectée dans une zone contenant de nombreuses irrégularités optiques a une incertitude *a priori* importante. Cette incertitude *a priori* peut être utilisée dans des réseaux probabilistes, tels que les réseaux bayesiens. L'utilisation

de contexte permet ainsi à la tâche de suivi, de calculer des pistes plus fiables représentant la trace des régions mobiles, problème essentiel de tout système d'interprétation.

Représentation du contexte

Nous avons défini un élément de contexte pour une zone polygonale donnée et associé à la tâche de suivi comprenant au moins quatre attributs :

- Les descriptions symboliques des irrégularités optiques dont l'influence s'exerce dans la zone donnée. Ces descriptions font référence à des descriptions plus complètes, directement utilisables par la tâche de suivi de régions mobiles et contenues dans une bibliothèque attachée au module de suivi. Cette bibliothèque a pour objectif d'être générique (indépendante de l'application) et donc d'être réutilisable pour différentes bases de contexte. Par exemple, la zone d'influence de l'irrégularité optique *« barrière »*, est la zone balayée par la barrière lorsqu'elle est soulevée.
- L'incertitude *a priori* quantifiant les résultats du suivi de régions mobiles.
- L'indication si la zone donnée est une zone potentielle d'apparition ou de disparition de régions mobiles, permettant d'anticiper l'initialisation et la terminaison des pistes des régions mobiles.
- L'ensemble des points 2D délimitant le contour polygonal de la zone donnée.

D'autres exemples d'attributs, plus complets, sont donnés dans la section 4.2.4. Le contenu de cet élément de contexte permet alors d'améliorer les performances de la tâche de suivi.

Exemple d'utilisation

La figure 2.2 montre un exemple d'utilisation du contexte pour la tâche de suivi de régions mobiles. Il s'agit de corriger une erreur de détection du mouvement, due à une occultation. Au centre de l'image A, on peut voir un individu évoluant derrière un banc. Sur l'image B, on a dessiné les régions mobiles détectées par le système d'interprétation. Ces régions constituent les données d'entrée de la tâche de suivi. On note alors que l'individu évoluant derrière le banc est mal détecté, le dossier du banc l'occultant partiellement. La région correspondant à la perception de l'individu est coupée en

A B

FIG. 2.2 – *L'image A est l'image courante et représente la donnée d'entrée du système. L'image B contient les régions mobiles détectées par le module de détection. Les deux régions du centre correspondent à la perception d'un individu occulté par le dossier d'un banc.*

deux. Pour corriger cette erreur, nous commençons par déterminer les zones, contenant les régions mobiles détectées. À ce niveau, on utilise comme informations contextuelles, le numéro des régions mobiles correspondant à leur indexation sur la décomposition de l'espace. On récupère alors l'élément de contexte de la tâche de suivi, lié à la zone contenant le dossier du banc. Cet élément de contexte contient la description de l'occultation. Le traitement associé à cette information consiste à réunir les deux régions mobiles détectées en une seule région. Ce traitement est spécifique à ce type d'occultation. La présence d'une occultation par exemple, par un dossier opaque nécessite la définition d'un traitement différent. La nouvelle région correspondant à la perception corrigée de l'individu, permet alors de ne pas perdre le suivi de l'individu, lorsqu'il passe derrière le banc. Ensuite, nous utilisons l'incertitude *a priori*, contenue dans le même élément de contexte, afin d'établir la confiance dans la tâche de suivi. Cette information indique la possibilité d'erreur dans le suivi de cette région mobile. Grâce au contexte, la tâche de suivi a ainsi pu améliorer ses résultats et les quantifier.

2.3.5 Contexte dans la reconnaissance de scénarios

Utilité du contexte

La tâche de reconnaissance de scénario est décrite dans les chapitres 5 et 6. Son objectif est de déterminer une description de scénario permettant d'expliquer les propriétés des objets mobiles. Ce qui rend cette tâche particulièrement délicate est le fossé séparant les descriptions de scénario et les propriétés

des objets mobiles. Ce fossé est dû à la nature des scénarios, qui sont des données abstraites dépendantes de l'application et dont la représentation est directement liée à leur expression en langage naturel. Un certain nombre d'informations, nécessaires à la description des scénarios, ne peuvent pas alors être calculées par des modules de vision. Ces informations souvent implicites pour les êtres humains, font partie du contexte de la scène. Un moyen d'améliorer la tâche de reconnaissance de scénarios est alors de fournir ces informations contextuelles, formant ainsi un pont entre les propriétés des objets mobiles et les descriptions de scénario. Le contexte de cette tâche peut ainsi être vu comme un ensemble de liens. Ces informations contextuelles appartiennent au domaine source IES (Informations sur l'Environnement de la Scène) puisque en général, leur signification dépend de la zone considérée. Par exemple comme le remarque (Howarth, 1995), la plupart des lieux sont nommés soit par leur localisation, comme *« une cuisine »*, ou par les actions s'y déroulant, telles que *« cuisiner »*. Ces informations fournissent par ce biais, à la tâche de reconnaissance de scénarios, de nombreux indices pour déterminer les comportements.

Représentation du contexte

Nous utilisons un élément de contexte pour une zone donnée et associé à la tâche de reconnaissance de scénarios, composé d'un ensemble de cinq attributs :

- Les seuils permettant de quantifier les propriétés des objets mobiles. Par exemple, ces seuils permettent de préciser, pour la zone donnée, la limite à partir de laquelle une vitesse est considérée comme étant faible ou élevée.

- Des indices permettant de sélectionner certaines descriptions de scénario. Par exemple, l'information contextuelle *« près d'un panneau »*, permet de renforcer la reconnaissance du scénario *« être stationnaire »*, scénario considéré comme normal dans une telle zone.

- Le coefficient d'intérêt de la zone indiquant, que les objets mobiles détectés dans la zone donnée sont particulièrement intéressants, et qu'il est nécessaire d'analyser leur comportement plus finement.

- Les listes des scénarios attendus, autorisés et tolérés dans la zone donnée. Par exemple, le scénario *« être stationnaire »* est interdit sur la chaussée. Ces scénarios sont des liens symboliques faisant référence

à une bibliothèque de descriptions de scénario, contenue dans le module de reconnaissance de scénarios. Ces informations symboliques indépendantes de l'application facilitent l'acquisition et la réutilisation du contexte. Ces listes permettent de renforcer la vraisemblance de la reconnaissance des scénarios et leur importance. Par exemple, si on reconnaît un scénario interdit, l'importance de ce scénario est augmentée.

– L'ensemble des points 2D du contour polygonal délimitant la zone.

Le contenu de cet élément de contexte permet alors d'améliorer les performances de la tâche de reconnaissance de scénarios.

Exemple d'utilisation

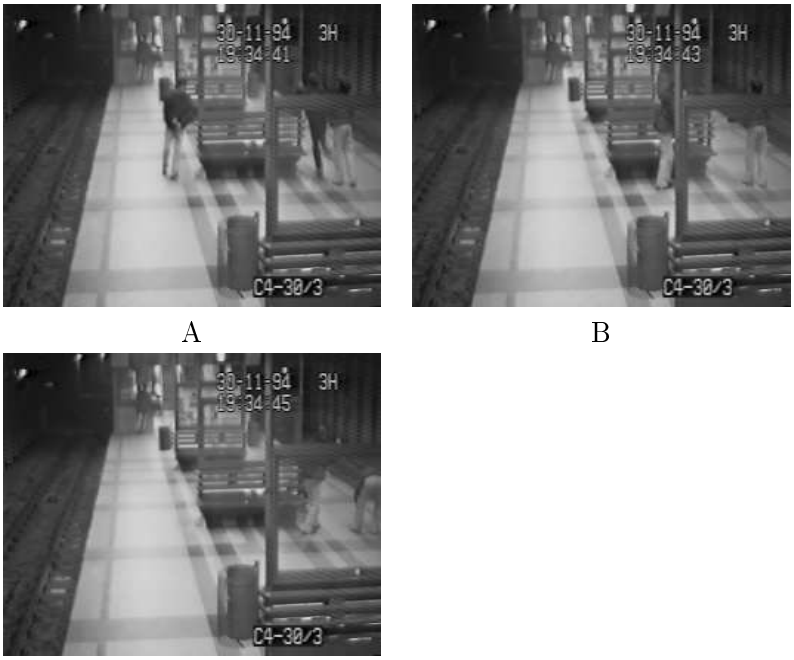

FIG. 2.3 – *Sur cette séquence d'images, trois individus tournent autour d'un banc. On essaie plus particulièrement d'analyser le comportement de l'individu au centre de l'image A.*

Sur la figure 2.3, la séquence d'images illustre l'utilisation des éléments de contexte associés à la tâche de reconnaissance de scénarios. L'objectif du

système est de comprendre que l'individu au centre de l'image A court autour du banc et de déclencher une alarme, ce scénario étant interdit.

- Sur l'image A, l'individu est détecté avec une vitesse élevée, évoluant sur le quai d'accès au métro. L'élément de contexte de cette zone indique qu'une telle vitesse correspond à l'action *« courir »*, et que ce scénario est toléré. Le module de reconnaissance de scénarios accroît alors le degré d'intérêt de l'individu.

- Sur l'image B, l'individu est détecté au milieu du banc avec une mauvaise qualité de détection. Le module de reconnaissance émet alors deux hypothèses concernant le comportement de l'individu : *« l'individu s'asseoit »* ou *« l'individu court »*. L'élément de contexte de la zone *« près du banc »*, indique que ce premier scénario est considéré comme normal, tandis que le deuxième est interdit.

- Sur l'image C, l'individu est détecté avec la même vitesse élevée, toujours dans la zone près du banc. Le scénario *« l'individu court »* est alors reconnu avec suffisamment de confiance. Ce scénario étant indiqué comme interdit par l'élément de contexte de la zone *« près du banc »*, le module de reconnaissance de scénarios déclenche alors une alarme.

Dans cet exemple, nous utilisons deux éléments de contexte associés à la tâche de reconnaissance de scénarios : ceux de la zone *« sur le quai »* et de la zone *« près du banc »*. Cette utilisation du contexte montre comment le module de reconnaissance focalise l'attention du système sur les objets mobiles d'intérêt et adapte ses actions en fonction de l'environnement de la scène. L'exemple décrit dans cette section serait particulièrement difficile à reconnaître sans l'utilisation du contexte.

2.4 MARES : un logiciel d'acquisition du contexte

La plupart des informations contextuelles décrites dans ce chapitre sont des informations *a priori*, acquises par un opérateur humain pendant l'installation du système d'interprétation. Cette tâche étant particulièrement astreignante et coûteuse en temps, nous avons développé une interface graphique permettant, d'une part à l'opérateur humain d'acquérir interactivement des informations contextuelles et, d'autre part, de construire la représentation du contexte de la scène. Nous avons appelé ce logiciel d'acquisition

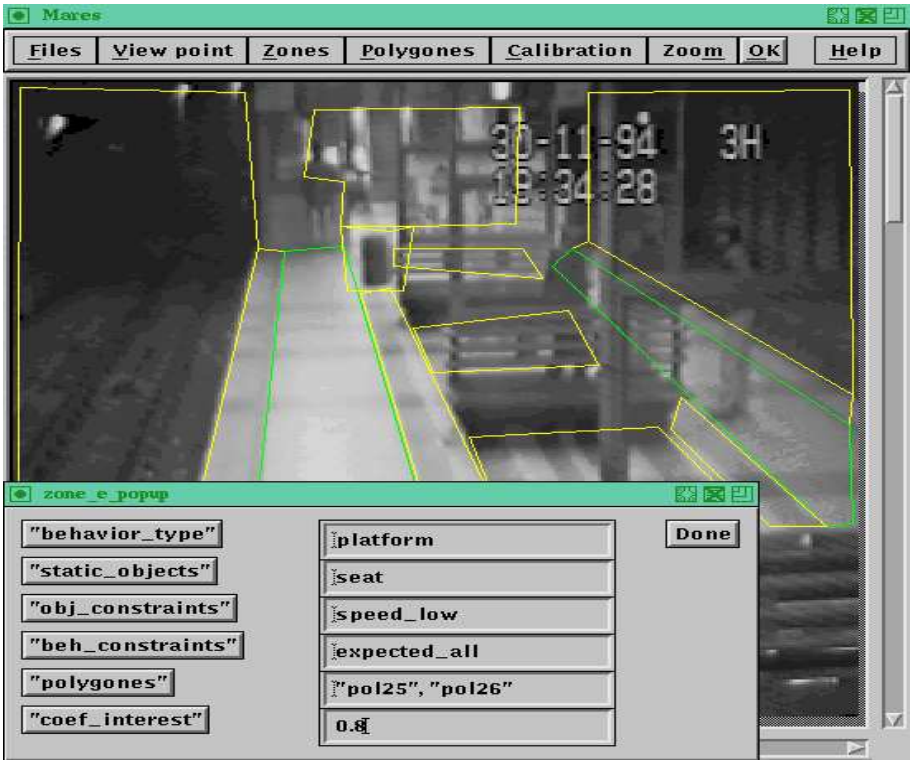

FIG. 2.4 – *L'opérateur humain utilise l'interface graphique, pour saisir les valeurs de l'élément de contexte, associé à la zone polygonale située sur le quai.*

« *MARES* », pour Module d'Acquisition et de Représentation d'Environnements Statiques. Il est implanté en langage C, utilise la boite à outils graphique MOTIF, et la boite à outils de dessin Knvas. Le manuel d'utilisation est en partie décrit dans (Nade, 1995). Pour générer la base de contexte, l'opérateur humain exécute une boucle de six opérations principales :

1) L'opérateur lance le logiciel. Une fenêtre s'affiche alors avec l'image du fond de la scène, ne contenant aucun objet mobile. Il choisit alors un fichier de descriptions, définissant la structure (c.-à-d. la classe) des éléments de contexte. Des fichiers définis par défaut permettent de lancer directement le logiciel.

2) L'opérateur sélectionne ensuite l'une des tâches du processus d'interprétation (p. ex. le suivi de régions mobiles).

3) Puis il dessine sur l'image de fond de scène un ensemble de polygones en cliquant à l'écran pour désigner les points des contours polygonaux. Pour signaler que cette opération est terminée, l'opérateur sauvegarde les polygones dessinés. Un tableau (une fenêtre avec des cases à remplir) s'affiche alors à l'écran, afin d'acquérir les informations contextuelles associées à la zone dessinée. Ce tableau correspond à l'élément de contexte de la tâche sélectionnée, associé aux zones polygonales dessinées.

4) L'opérateur remplit alors les cases du tableau à l'aide du clavier. Comme le montre la figure 2.4, sur la colonne de gauche du tableau sont indiqués les noms des champs, et sur celle de droite sont situés les emplacements des valeurs. Le fait qu'une case reste vide (et plus généralement que des zones ne soient pas associées à des éléments de contexte) n'est pas anormal. Dans ce cas, le système d'interprétation considère que l'information contextuelle n'est pas disponible, et continue son traitement sans tirer profit du contexte.

5) Lorsque l'opérateur remplit les cases correspondant à des matrices de calibration, une fenêtre supplémentaire s'affiche, afin de saisir les coordonnées 2D et 3D des points permettant le calcul de la matrice.

6) Une fois les cases remplies, l'opérateur sauvegarde la base de contexte pour clore la phase de saisie de l'élément de contexte. Il peut ensuite continuer à saisir des éléments de contexte du même type (retour à l'opération 3) ou de type différent (retour à l'opération 2).

Lorsque les éléments de contexte ont été saisis pour toutes les tâches du processus d'interprétation, le système construit alors la base de contexte. On obtient ainsi un ensemble d'éléments de contexte, visible sous différents points de vue et relié à quatre décompositions de l'espace en zones polygonales (une pour chaque tâche). Pour l'instant ces quatre décompositions ne sont pas reliées les unes aux autres. Cependant nous comptons étendre le système, afin de permettre la fusion des quatre décompositions en une seule en considérant l'intersection de tous les polygones. Pour cette raison dans la suite de ce mémoire, nous ne considérons la présence que d'une seule décomposition de l'espace, associée à tous les éléments de contexte. En fin de session, la base de contexte est sauvegardée sous la forme d'un fichier texte. L'opérateur peut recharger une base de contexte déjà saisie, la modifier et la compléter de différentes manières. Il peut grossir l'image de fond, afin d'affiner la saisie des polygones. Il peut déformer le contour polygonal d'une

zone déjà dessinée, et peut modifier le contenu d'un élément de contexte déjà saisi.

Ce logiciel est la solution que nous avons choisie afin de faciliter la tâche d'acquisition du contexte. Cette solution logicielle a été également choisie dans le cadre du projet européen Esprit VIEWS (Corrall, 1992), où l'utilisateur du système peut construire à l'aide d'un logiciel une carte (vue de dessus) de la scène, et peut également saisir les modèles *a priori* des objets mobiles de la scène.

2.5 Conclusion

Dans ce chapitre, nos contributions ont porté sur quatre points :

- Un **formalisme général**, permettant de définir et de délimiter les informations contextuelles. Ce formalisme permet une utilisation rationnelle et systématique du contexte.
- Une **représentation de la base de contexte**, possédant trois caractéristiques principales. Premièrement, elle est centralisée en un seul lieu, la décomposition de l'espace en zones polygonales. De plus, tous les éléments de contexte sont représentés sous un formalisme commun. Ces caractéristiques permettent aux tâches du processus d'interprétation de partager les mêmes informations contextuelles. Deuxièmement, la base de contexte est multi-points de vue, structurant ainsi les informations selon leur domaine d'utilisation. Troisièmement, les informations contextuelles sont représentées sous une forme symbolique, afin de permettre la réutilisation de bases de contexte déjà acquises.
- En ce qui concerne nos applications cibles, la réalisation de la base de contexte a nécessité la définition d'une **vingtaine d'attributs**, caractérisant ainsi les éléments de contexte des différentes tâches du processus d'interprétation.
- Un **logiciel graphique**, permettant la construction de la base de contexte.

Cependant l'acquisition du contexte reste fastidieuse, malgré l'aide apportée par le logiciel graphique. De plus, certaines applications nécessitent l'utilisation de capteurs modifiant le contexte pendant le déroulement du processus d'interprétation (p. ex. les caméras mobiles, les caméras pan-tilt-zoom). Par exemple, dans une application d'interprétation de séquences d'images prises à partir d'un avion, la caméra ne peut être que mobile.

Un axe de recherche consiste alors à étendre la base de contexte, afin d'acquérir et de mettre à jour automatiquement les informations contextuelles, pendant le déroulement du processus d'interprétation. Cette automatisation peut par exemple, être réalisée à l'aide de méthodes statistiques ou de méthodes d'apprentissage symbolique. Certains travaux (Johnson and Hogg, 1996), ont déjà abordé ce problème, en essayant d'apprendre automatiquement les chemins les plus communément employés par les objets mobiles.

Chapitre 3

Détection du mouvement

Ce chapitre a pour but de décrire le module choisi afin de détecter les régions mobiles. Notre objectif n'est pas de concevoir une nouvelle méthode de détection du mouvement, mais de choisir une méthode opérationnelle déjà existante, et de l'intégrer dans le système d'interprétation. D'après les spécifications décrites dans le chapitre 1, la méthode de détection du mouvement se doit d'être :

- **robuste** : pour faire face à des conditions réelles d'acquisition d'images. Cette méthode doit également être suffisamment générique pour traiter différents types de scène.
- **rapide** : pour être utilisée dans des applications traitant les séquences d'images à la volée (cadence de 4 à 5 images par seconde). La tâche de détection du mouvement est en général la tâche limitante du système d'interprétation et se doit d'être optimisée.
- **facile d'utilisation** : pour être rapidement intégrée au système d'interprétation et être facilement extensible si les besoins des applications cibles le nécessitent. En particulier, cette méthode doit posséder une phase d'initialisation réduite et doit pouvoir ajuster automatiquement ses paramètres.

Par contre, nous ne cherchons pas à déterminer les paramètres du mouvement des objets mobiles, tels que le terme de rotation. Les applications cibles ayant en général de mauvaises conditions d'acquisition des séquences d'images, telles qu'une faible sensibilité de la caméra, les méthodes classiques d'estimation des paramètres du mouvement ne peuvent pas s'appliquer. Par exemple, la méthode de détermination du mouvement 3D à partir du champ

de mouvement apparent ne s'applique qu'à condition que le mouvement apparent soit suffisamment précis. De plus, étant donnée l'échelle de temps considérée (de l'ordre de la dizaine de minutes), ces méthodes ne déterminent que le mouvement instantané des objets mobiles. Nous nous intéressons plutôt à la détermination du mouvement moyen, tel que *« l'individu avance droit devant lui »*, prenant en compte les déplacements des objets sur des intervalles de temps suffisamment grands. La détermination du mouvement moyen nécessite alors une analyse complète de la séquence d'images, et est réalisée si besoin, au niveau plus abstrait de la reconnaissance des scénarios. La méthode recherchée n'a ainsi pour objectif que la localisation des régions mobiles correspondant à la perception du mouvement instantané des objets.

Afin de déterminer des éléments permettant de choisir une méthode remplissant les spécifications précédentes, nous commençons dans la section 3.1 par décrire la problématique du processus de détection du mouvement. Dans la section 3.2 nous décrivons le module de détection des régions mobiles que nous avons à notre disposition, en soulignant ses points forts et ses points faibles. Enfin, dans la section 3.3 nous proposons plusieurs méthodes afin d'améliorer les performances de ce module.

3.1 Analyse du problème

La détermination du mouvement instantané se décompose en général en trois étapes. La première étape consiste à détecter les éléments de l'image en mouvement. La deuxième étape regroupe ses éléments, afin d'obtenir une segmentation de l'image en régions mobiles. La troisième étape a pour objectif d'estimer les paramètres du mouvement réel des objets mobiles, en particulier l'équation de leur mouvement 3D. Comme le montre (François, 1991), il existe quatre approches mettant en œuvre le processus de détermination du mouvement :

- **Approche par mise en correspondance** : cette approche consiste à extraire préalablement des primitives 2D (contours, coins, motifs correspondants à la texture d'une zone, régions de l'image), puis à mettre en correspondance ces primitives, avec celles extraites aux instants précédents. Cette mise en correspondance est réalisée à l'aide d'hypothèses *a priori* sur la structure des surfaces observées (p. ex. surfaces supposées planaires) et sur la nature des mouvements des objets mobiles (p. ex. mouvement supposé rigide). Dans un second temps, ces hypothèses *a priori* permettent également de reconstruire la structure 3D des objets, ainsi que leur mouvement 3D. En particulier, cette approche

est utilisée en vision stéréoscopique, où la structure de l'environnement est plus facilement déterminable, en raison de la connaissance de la profondeur de la scène. L'inconvénient de cette approche est l'utilisation de connaissances *a priori*. Ces connaissances sont spécifiques aux applications et sont nécessaires pour construire les régions mobiles à partir du regroupement des primitives 2D. De plus, le nombre de primitives 2D facilement reconnaissables et caractérisant un objet mobile est souvent faible. Cette approche permet d'obtenir un champ de mouvement épars, puisqu'il est calculé à partir d'un petit nombre de primitives 2D. L'obtention de ce champ est alors relativement peu coûteuse en temps de calcul, mais souffre de phénomènes d'instabilité. Cependant, cette approche permet de résoudre efficacement certaines applications. Par exemple dans (Motamed and Vannoorenberghe, 1997), les auteurs calculent les flux de mouvements d'une foule à partir de la détermination et le suivi de motifs texturaux.

– **Approche différentielle** : cette approche repose sur l'hypothèse de base de l'invariance de la luminance d'un point, lors de son déplacement dans l'espace échantillonné par la séquence d'images (Bouthémy, 1988). Cette hypothèse conduit à une équation différentielle reliant le gradient spatio-temporel de l'intensité lumineuse au vecteur vitesse. Cette équation est connue sous le nom d'Équation de Contrainte du Mouvement Apparent (ECMA) :

$$\overrightarrow{\nabla I}.\overrightarrow{v} + \frac{\partial I}{\partial t} = 0 \tag{3.1}$$

avec $\overrightarrow{\nabla I} = (\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y})$ le gradient spatial de l'intensité lumineuse, et $\frac{\partial I}{\partial t}$ le gradient temporel. Elle permet alors de mesurer la composante du vecteur vitesse parallèle au gradient d'intensité en chaque point de l'image, constituant ainsi le flot optique de la séquence d'images. Dans (François, 1991), l'auteur énumère différents problèmes posés par l'ECMA :

  – La non prise en compte des phénomènes d'ombrage et de la variation de la luminosité.
  – La supposition d'hypothèses fortes sur la fonction d'intensité. Par exemple, cette équation ne s'applique pas aux zones à forte texture, car ces zones ne respectent pas l'hypothèse de linéarité locale de l'intensité. De même, l'équation ne permet pas de mesurer la vitesse des points des zones uniformes, car le gradient spatial de l'intensité est nul en ces points.

- La faiblesse du modèle aux frontières. L'ECMA ne peut pas s'appliquer au niveau des contours, car l'hypothèse de la linéarité de la fonction d'intensité n'est pas vérifiée.
- Les restrictions sur le type de mouvement des objets mobiles. Par exemple, les grands déplacements et les problèmes d'occultation ne sont pas pris en compte par le modèle.
- L'assimilation abusive entre le champ des vitesses apparentes et le mouvement projeté. Le mouvement projeté est la réelle projection sur le plan image 2D du mouvement 3D, tandis que le mouvement apparent (seul mesurable) est le mouvement 2D perçu dans l'image, au travers des variations temporelles de l'intensité lumineuse. En général, ces deux notions étant différentes, l'ECMA ne permet alors de mesurer qu'une approximation du mouvement projeté. Si les conditions d'acquisition de la séquence d'images sont mauvaises, telles qu'une résolution grossière de la caméra, un niveau de bruit élevé ou un rapport *« taille de l'objet / profondeur relative »* faible, cette approximation est imprécise.
- Un coût de traitement important. Le gradient spatio-temporel de la fonction d'intensité doit être calculé en chaque point de l'image.

Cependant, il existe de nombreuses méthodes permettant en partie de résoudre ces problèmes. En particulier, l'emploi de méthodes multi-résolutions permet de calculer les vitesses apparentes pour de grandes amplitudes de déplacement. Elles permettent également de réduire le temps de traitement, en calculant le gradient de la fonction d'intensité uniquement sur les zones à forte granularité, et de préciser ce calcul si besoin, en prenant une granularité plus fine. De même, les méthodes markoviennes permettent d'améliorer le calcul des vitesses apparentes. En particulier, elles permettent d'intégrer plusieurs sources de mesure du mouvement et de spécifier des interactions spatiales locales complexes. Par exemple, ces méthodes peuvent conduire à minimiser la somme de deux termes d'énergie. Le premier terme est alors relatif à l'adéquation du champ aux variations spatio-temporelles de l'intensité, et le second terme favorise la continuité du champ de vecteur (Ricquebourg, 1993). Une étude comparant les performances de différentes méthodes de flots optiques peut être trouvée dans (Barron et al., 1994). En théorie, toutes ces méthodes permettent de calculer des champs de mouvement apparent denses, cependant comme le montre cette étude, ces résultats ne concordent pas toujours avec la réalité.

- **Approche par comparaison à une image de référence** : pour employer cette approche il est nécessaire de posséder une image de référence, correspondant à l'image du fond de la scène sans objet mobile. Cette méthode consiste à faire la différence entre l'image courante et l'image de référence, afin d'obtenir les points en mouvement. Un problème dur de cette approche est la mise à jour de l'image de référence (Karmann and Brandt, 1989). Cette mise à jour doit compenser le changement de l'illumination de la scène, et doit intégrer à l'image de référence les objets mobiles devenu statiques. Cette approche a comme première qualité sa robustesse. Elle permet par exemple d'éliminer une partie des phénomènes d'ombrage. Elle est de plus facile à mettre en œuvre, ne nécessitant pas en général de nombreux réglages de paramètres. Le principal inconvénient de cette méthode est la nécessité d'utiliser une caméra fixe. Cette approche est néanmoins largement utilisée, en particulier à cause de sa robustesse. Par exemple dans (Sato et al., 1993), (Motamed and Vannoorenberghe, 1997), (Azarbayejani et al., 1996), les auteurs utilisent cette approche, qu'ils combinent à des techniques sophistiquées pour mettre à jour l'image de référence. Cette mise à jour est réalisée en se basant sur l'historique des changements de luminosité. Cette approche par comparaison possède différentes variantes. Par exemple dans (Dawson-Howe, 1996), l'auteur utilise une première caméra à grand angle de vue pour mettre à jour l'image de référence, et une seconde caméra, mobile et à plus forte résolution, pour obtenir l'image courante. Cette méthode permet, en partie, de contourner l'inconvénient de devoir utiliser une caméra fixe.

- **Autres approches** : diverses approches sont également utilisées afin de déterminer le mouvement dans une scène. Par exemple, l'approche par transformées, comme la transformée de Fourier, consiste à utiliser les propriétés de l'évolution spatio-temporelle du signal dans le domaine fréquentiel.

Cette section n'a pas pour objectif d'étudier ou d'énumérer exhaustivement toutes les méthodes de détermination du mouvement. En tant qu'utilisateur, notre but est plutôt de déterminer des éléments pour choisir un module de détection des régions mobiles. Il s'avère que les méthodes les plus fréquemment utilisées proviennent des approches différentielles et de comparaison à une image de référence. Les méthodes relatives à l'approche différentielle peuvent être robustes, mais imposent une mise en œuvre souvent complexe. Les méthodes de l'approche par comparaison à une image

de référence sont plus faciles à élaborer, mais ne peuvent s'appliquer qu'aux applications utilisant une caméra fixe. Certaines méthodes, comme celle de (Ricquebourg, 1993), combinent ces deux approches.

3.2 Module choisi de détection des régions mobiles

3.2.1 Description du module

Nous utilisons un module qui a été développé au cours du projet européen Esprit HPCN PASSWORDS (Bogaert et al., 1996). Utilisant une caméra fixe monoculaire couleur, ce module traite des séquences d'images et détecte les régions mobiles. Il se singularise par la mise à jour de l'image de référence en compensant le changement de l'illumination, et par l'élimination d'une partie des phénomènes d'ombrage dans les régions mobiles détectées. La compensation du changement d'illumination permet en particulier de traiter des scènes extérieures. Les facteurs de changement de l'illumination sont calculés à partir de la différence logarithmique entre l'image courante et l'image de référence. Ces facteurs sont lissés au cours du temps afin d'éviter de grandes variations discontinues. Pour déterminer les phénomènes d'ombrage, chaque région mobile est segmentée en régions de couleur. Après une phase de classification, les régions de couleur correspondant à de l'ombre sont éliminées. Les régions mobiles ainsi détectées sont caractérisées par leur boite englobante (hauteur, largeur et position) et par leurs régions de couleur.

3.2.2 Résultats

Ce module a été testé sur de nombreuses séquences d'images, prises dans différentes conditions d'acquisition. La qualité des résultats est directement dépendante de ces conditions. Si elles sont mauvaises (p. ex. pluie, présence d'ombres, de reflets) de nombreuses régions correspondant à du bruit sont détectées comme des régions mobiles. En particulier, dans certaines séquences d'images de métro, de grandes zones de reflets sur le sol sont considérées comme des régions mobiles. Un second inconvénient provient de la nécessité d'utiliser une caméra fixe. Cette restriction ne convient pas à bon nombre d'applications.

Le principal avantage de cette méthode est sa facilité de mise en œuvre. Elle ne demande que très peu de réglages concernant les paramètres du traitement. Par exemple, la taille minimale d'une région mobile est un des paramètres à régler. Ce paramètre est directement dépendant de la distance entre la caméra et la scène, et demande à être ajusté à chaque changement

de scènes. Bien que ne permettant pas encore un traitement temps réel, ce module permet de traiter une image (512x512) couleur, en près de deux secondes sur un Sun Sparc 10. Grâce à ce module, nous avons pu détecter avec suffisamment de précision les régions mobiles correspondant à différentes séquences d'images.

3.3 Améliorations

3.3.1 Améliorations des méthodes déjà existantes

Notre objectif n'est pas de développer de nouvelles méthodes de détection du mouvement, mais d'étudier différentes approches permettant d'utiliser au mieux les méthodes déjà existantes. Dans ce but, nous prévoyons à court terme, de mettre en œuvre trois techniques :

- l'ajustement automatique des paramètres de la méthode de détection du mouvement. De manière plus générale, nous comptons dans un second temps utiliser les techniques de pilotage de programmes de traitement d'images, afin d'employer la méthode de détection dans les conditions pour lesquelles elle a été conçue. Ces techniques facilitent le traitement de nouvelles scènes par le système d'interprétation, en permettant d'adapter la méthode de détection au contexte de la nouvelle scène.

- le contrôle de la mise à jour de l'image de référence par le système d'interprétation. Cette technique permet au système d'interprétation de signaler au module de détection de régions mobiles la présence de deux situations à prendre en compte. La première situation est rencontrée lorsqu'un objet précédemment mobile devient statique. Il s'agit alors pour le module de détection d'intégrer cet objet dans l'image de référence. Inversement, la deuxième situation se produit lorsqu'un objet faisant précédemment partie du fond de la scène devient mobile. Dans ce cas, le module doit ôter l'objet de l'image de référence.

- la sélection de l'attention du module de détection de régions mobiles. Dans cette technique, il s'agit pour le système d'interprétation de signaler au module de détection, les régions d'intérêts (« *Regions Of Interest* », R.O.I. en anglais) et les régions bruitées. Les R.O.I. sont des portions de l'image contenant effectivement des objets mobiles. Le module de détection peut alors appliquer des traitements sophistiqués sur ces régions afin de mieux délimiter les régions mobiles. Par région

bruitée, nous entendons une portion d'image contenant du bruit, tel qu'un reflet, une ombre, et ne pouvant pas contenir d'objet mobile. Connaissant la présence de ces régions, le module de détection peut éviter de les traiter. Le module n'a plus alors qu'à appliquer un traitement peu coûteux sur les régions restantes (c.-à-d. ni d'intérêts, ni bruitées), afin de détecter toute nouvelle apparition de mouvement. Cette technique est particulièrement importante, car le temps de traitement du système global d'interprétation est avant tout limité par le module de détection.

Ces techniques traduisent l'intérêt de faire communiquer entre eux les modules du système d'interprétation. Le module de détection bénéficie ainsi des résultats obtenus par les autres modules, ces résultats étant contenus dans la base de contexte (se référer à la section 2.3). Certaines de ces techniques ont déjà commencé à être implantées dans le système d'interprétation.

3.3.2 Évolution des modules de détection

Notre objectif est également de pouvoir changer facilement de module de détection ou de pouvoir en combiner plusieurs. Nous espérons ainsi améliorer les performances de ce module et de pouvoir prendre en compte toute évolution. Effectivement, le module de détection peut évoluer du fait de l'utilisation de nouveaux capteurs (p. ex. caméras infrarouges), de nouveaux algorithmes de détection (p. ex. flot optique) ou de l'utilisation de plusieurs capteurs (p. ex. avoir plusieurs angles de vue d'une même scène). Dans cette situation, ce module serait en mesure de fournir des propriétés supplémentaires sur les données d'entrée des autres modules du système d'interprétation.

Un premier axe de recherche est alors de garantir une détection minimale, quelque soit le type d'application, moyennant un changement éventuel de module de détection. Dans ce but, nous avons déjà spécifié un corpus faisant interface (c.-à-d. gérant les communications) entre des modules de traitement d'images et d'interprétation, afin de reconnaître des objets statiques complexes (Ossola, 1996). Il s'agit alors d'étendre ces travaux à l'interprétation de séquences d'images. Un second axe de recherche consiste à anticiper l'évolution du module de détection, en répercutant cette évolution sur les autres modules du système d'interprétation (p. ex. fusion des propriétés relatives à un même objet mobile).

3.4 Conclusion

Ce chapitre ne contient pas à proprement parler de contribution. Son objectif est de fournir des éléments permettant de choisir une méthode de détection de régions mobiles. Les approches les plus prometteuses sont l'approche différentielle et l'approche par comparaison à une image de référence. Nous avons ainsi testé différents algorithmes et choisi finalement d'intégrer au système d'interprétation le module développé dans le cadre du projet européen Esprit HPCN PASSWORDS. Les raisons de ce choix sont, d'une part la facilité de sa mise en œuvre et d'autre part sa robustesse. En effet, à l'aide de ce module nous avons pu traiter une grande diversité de séquences d'images.

Néanmoins, le problème de la détermination du mouvement reste un problème difficile. Sur de longues séquences d'images, un faible taux d'erreur de détection équivaut globalement à un très grand nombre d'erreurs. Les performances de ce module en qualité des résultats obtenus et en temps de traitement limitent les performances de tout le système d'interprétation. Il est alors nécessaire d'améliorer le module de détection de régions mobiles. Nous avons ainsi prévu deux types d'amélioration. À court terme, nous comptons appliquer des techniques de pilotage de programmes de traitement d'images. À plus long terme, nous envisageons de concevoir une interface entre le module de détection et le reste du système d'interprétation, afin de faciliter l'évolution de ce module. Ces deux types d'amélioration ne sont pas antagonistes et peuvent être réalisés concomitamment.

Chapitre 4

Suivi de régions mobiles

Le but de ce chapitre est de proposer une nouvelle méthode de suivi de régions mobiles représentant des objets en mouvement et de décrire le module de suivi implantant cette méthode. La méthode de suivi recherchée doit obtenir avec des exigences de traitement temps réel, un suivi robuste et simultané de plusieurs objets mobiles, évoluant avec des mouvements indépendants dans un environnement encombré et complexe, sur une longue séquence temporelle. Par exemple, un de nos objectifs est de suivre plusieurs individus s'entrecroisant dans une station de métro, sur une séquence d'images dépassant le quart d'heure. Les problèmes majeurs d'une méthode de suivi sont d'une part, l'estimation du mouvement de l'objet suivi (appelé cible) permettant de prédire sa nouvelle position et d'autre part, le traitement des associations ambiguës (appelé en anglais, *« the data association problem »*) entre les cibles déjà existantes et les régions mobiles détectées dans la nouvelle image. La valeur d'une méthode de suivi, sa robustesse, réside dans l'efficacité du traitement des associations ambiguës.

Le problème de suivi des régions mobiles est un problème clé du processus global d'interprétation de séquences d'images, dans la mesure où la perte du suivi d'un objet mobile bloque la chaîne de traitements, mettant en échec tout le système d'interprétation. Dans ce chapitre, la section 4.1 présente un état de l'art des méthodes de suivi d'objets mobiles rigides et non rigides. La section 4.2 décrit ensuite les principes de la méthode de suivi proposée et définit précisément les modèles utilisés relatifs aux objets mobiles et à leur mouvement. Dans la section 4.3, on calcule les correspondances entre cibles et régions mobiles nouvellement détectées à l'aide d'une matrice d'ambiguïté. On traite également le cas de la mise à jour des cibles non ambiguës. Dans la section 4.4, on essaie de résoudre les situations ambiguës. Pour cela, on

introduit la notion de cible composée, permettant de geler temporairement le suivi des cibles ambiguës. Enfin, la section 4.6 décrit un exemple de suivi de deux individus dans une scène de parking.

4.1 État de l'art

Dans la littérature, il existe trois grandes familles de méthodes de suivi selon le type d'objet suivi. Cette section présente les principes de ces méthodes.

4.1.1 Suivi d'objets rigides

De nombreux travaux abordant le suivi d'objets rigides (p. ex. véhicules, robots) ont conduit à des méthodes performantes et robustes. Il existe principalement trois méthodes :

- Une première méthode consiste à détecter des primitives particulières sur les objets mobiles (p. ex. arêtes, coins) et à suivre ces primitives d'images en images (Wang and Brady, 1995). Cette méthode ne peut s'appliquer que sur des objets mobiles très particuliers, possédant des primitives nombreuses, faciles à détecter et pouvant s'organiser sous la forme de modèle d'objet.

- La seconde méthode consiste à mettre en correspondance la région mobile détectée 2D avec un modèle géométrique 3D de l'objet mobile. Pour cela, on estime le centre et la direction du mouvement de la région mobile, puis on met par exemple, en correspondance les segments de droite de la région mobile avec ceux du modèle. Cette mise en correspondance est souvent délicate, due au bruit des régions détectées. Dans (Du et al., 1993), les auteurs calculent un degré de cohérence de la mise en correspondance, évaluant les distances, les différences d'orientation et de longueur des segments appariés. Dans (Koller et al., 1993), les auteurs utilisent un modèle géométrique paramétré permettant une adaptation du modèle. Ils utilisent également un modèle de l'illumination pour calculer l'ombre des objets mobiles portée au sol.

- La troisième méthode consiste à suivre les contours des régions mobiles à l'aide d'un modèle déformable des contours. Par exemple, dans (Meyer and Bouthemy, 1992), on approxime la région mobile par un polygone convexe et on suit les points du polygone. Pour cela, on met globalement en correspondance le polygone prédit avec le polygone

nouvellement détecté en calculant la distance entre les deux polygones. L'intérêt de cette méthode est la prise en compte des occultations partielles des régions mobiles. La perte d'une partie du polygone est compensée par la présence de l'autre partie. Dans (Bascle et al., 1994), les auteurs procèdent de même, mais utilisent des courbes cubiques B-splines à la place des polygones. Dans (Koller et al., 1994), les auteurs proposent un traitement spécifique des situations d'occultation partielle. Cette méthode de suivi des contours possède principalement deux inconvénients : l'initialisation du premier contour et la gestion des occultations totales. Cependant, cette méthode peut être également utilisée dans le cas d'objet non rigide.

La suivi d'un objet mobile rigide se caractérise également par l'utilisation d'un filtre de Kalman pour estimer la nouvelle position de l'objet mobile (Hutber, 1995), (Du et al., 1993). Le filtre de Kalman consiste à estimer l'état d'un ensemble de mesures en tenant compte du bruit (Bar-Shalom and Fortmann, 1988), (Welch and Bishop, 1995). La méthode de filtrage utilise un modèle du calcul de l'état et des modèles du bruit relatif à ce calcul et aux mesures. Elle est constituée d'un cycle de deux opérations : (1) mise à jour du filtre (estimation de l'état à l'instant suivant et prédiction de l'erreur sur l'estimation réalisée) et (2) correction du filtre en tenant compte des nouvelles mesures dont on pondère (c.-à-d. filtre) l'influence selon les modèles du bruit. L'utilisation du filtre de Kalman nécessite alors la définition du modèle de mouvement. Dans (Koller et al., 1993), les auteurs utilisent un mouvement à trois degrés de liberté du centre de gravité de l'objet mobile, couplé avec un mouvement de rotation angulaire et de glissement. L'efficacité du filtre dépend essentiellement de l'adéquation entre les modèles utilisés et la réalité. Ce problème de modélisation est délicat, car les mouvements des objets mobiles réels suivent rarement des lois facilement modélisables.

4.1.2 Suivi d'objets non rigides

Les objets mobiles non rigides (p. ex. êtres humains) ne possèdent pas de modèle géométrique précis de leur forme. Donc contrairement au suivi d'objets rigides, on ne peut pas employer des méthodes de suivi utilisant de tels modèles. À la place, ces méthodes utilisent des modèles dynamiques ou temporaires des objets mobiles. Dans la littérature, trois méthodes de suivi d'objets mobiles non rigides sont généralement proposées :

– La première méthode utilise des modèles déformables de contours, comme dans le cas du suivi d'objets mobiles rigides. La différence entre

les deux types de méthode réside dans l'utilisation de modèles de déformation préférentielle. Par exemple dans (Baumberg and Hogg, 1995), on utilise des courbes cubiques B-splines pour représenter les modèles de contour. Selon le modèle de contour, on utilise un ensemble de vecteurs de déformation permettant de déformer, dans une certaine direction, la portion de la courbe comprise entre deux points de contrôle. De cette façon, on peut suivre un homme en train de marcher et reconnaître son mode de déplacement (p. ex. la marche) et la direction de son mouvement. L'intérêt de cette méthode réside dans l'apprentissage des modèles de contour et des déformations associées. Ses points faibles sont la dépendance des modèles par rapport à l'angle de prise de vue (p. ex. on doit définir un modèle pour un homme marchant de face et un autre pour un homme marchant de biais) et la non prise en compte des situations d'occultation.

– La seconde méthode utilise un modèle temporaire des régions mobiles représentant l'objet mobile sur l'image (Woodfill and Zabih, 1991). Ce modèle temporaire est défini par l'ensemble des pixels (c.-à-d. intensité des points de l'image) appartenant à la région mobile. À chaque nouvelle image, on compare ce modèle temporaire à l'intensité des pixels des régions mobiles nouvellement détectées. Puis, une fois que la correspondance est établie avec une région mobile, on met à jour le modèle temporaire en tenant compte des pixels de cette région mobile. Il existe différentes variantes de cette méthode. Par exemple dans (Choi et al., 1997), les auteurs utilisent en plus, un histogramme de couleur pour traiter les problèmes d'occultation dynamique entre des objets mobiles de couleurs différentes. Cette méthode est utilisée en particulier pour suivre des joueurs de football. Dans (Intille and Bobick, 1995), les auteurs incluent dans le modèle temporaire des objets mobiles le proche environnement des régions mobiles pour ne pas risquer de perdre une partie de l'objet à suivre. Ils utilisent également une modélisation du fond de la scène leur permettant ainsi d'extraire les objets statiques de l'environnement pouvant se mélanger aux objets mobiles. De cette façon, ils obtiennent un suivi plus fiable. Dans (Pentland, 1995), (Azarbayejani et al., 1996), les auteurs attachent à chaque point du modèle temporaire, la distribution de l'intensité de couleur (dans l'espace de couleur YUV) et sa distribution spatiale dans le plan image selon les coordonnées (x,y). Ils modélisent de même, chaque point du fond de la scène. Ils utilisent ces distributions pour prédire les évolutions des modèles (en particulier, le mouvement des objets mobiles) et pour com-

penser les changements d'illumination et la présence d'ombres. En modélisant les différentes régions de couleur d'un individu, ils arrivent à suivre précisément ses mouvements (p. ex. le mouvement de ses mains).

– La troisième méthode est proposée par l'équipe de P. Huttenlocker (Huttenlocker and Rucklidge, 1992). Elle utilise également un modèle temporaire des régions mobiles représentant l'objet mobile, mais ce modèle conserve les arêtes des régions mobiles plutôt que l'intensité de chaque pixel. De plus, cette méthode essaie d'abord d'ajuster le modèle temporaire de l'objet mobile à la région mobile nouvellement détectée, puis calcule la distance entre les deux ensembles d'arêtes en autorisant une mise en correspondance partielle des deux ensembles. Le pourcentage autorisé de mise en correspondance est un des paramètres de la méthode. Cette tolérance permet de traiter les cas d'occultation partielle des objets mobiles. Cette méthode se caractérise également par l'utilisation d'un modèle original du mouvement des objets mobiles. Pour cela, on calcule les classes de transformations (principalement des translations pour des raisons de temps de calcul) qui permettent de minimiser la distance entre le modèle temporaire de l'objet mobile et les régions de la nouvelle image. On détermine ainsi la région mobile correspondant à l'objet mobile et le mouvement de l'objet mobile en récupérant la classe de transformations. Cette méthode est utilisable quelque soit l'ampleur du déplacement de l'objet mobile. La seule restriction porte sur le changement du modèle temporaire de l'objet mobile, qui se doit d'être faible.

La plupart de ces méthodes se caractérisent par l'utilisation d'un filtre de Kalman comme modèle du mouvement, pour prédire la nouvelle position des objets mobiles. Ces méthodes se caractérisent également par les fortes conditions restrictives qu'elles imposent, pour permettre un suivi efficace (p. ex. caméra en face de l'objet mobile en train de se déplacer, utilisation de plusieurs caméras, objet mobile se détachant bien du fond de la scène). Ces conditions restrictives s'expliquent par la difficulté du problème du suivi d'objets mobiles et par la spécificité des applications envisagées. Seule la méthode proposée par P. Huttenlocker intègre dans ses principes, un traitement des occultations partielles (le cas des occultations dynamiques peut, sous certaines restrictions, être également traité). Néanmoins, on peut reprocher à cette méthode de modèle temporaire à base d'arêtes, qu'elle n'utilise pas d'historique du modèle de mouvement. Elle autorise effectivement n'importe quel déplacement de l'objet mobile suivi, pourvu que son modèle temporaire

de forme change peu. Cette contrainte lui empêche de traiter efficacement les problèmes d'occultation totale. Cette méthode nécessite également une bonne résolution des images.

4.1.3 Suivi d'objets sans modèle

Les méthodes de suivi d'objets sans modèle consiste à suivre des objets mobiles dont on ne connaît que les coordonnées de la position. Les précédentes méthodes de suivi utilisent un modèle des objets à suivre afin d'éviter des associations ambiguës. Les méthodes de suivi ne possédant pas de modèle d'objet se caractérisent pas le traitement sophistiqué des associations ambiguës. Elles n'utilisent pas de propriétés liées au traitement des images et au suivi de primitives (p. ex. coins, arêtes). Ces méthodes sont particulièrement utilisées dans l'imagerie radar. Elles peuvent aussi s'utiliser pour suivre des objets possédant un modèle dans le cas d'une mauvaise détection des objets mobiles (cas en général des applications de vidéo-surveillance). Ces méthodes de suivi s'appliquent également lorsqu'on utilise plusieurs capteurs et que certains de ces capteurs ne peuvent détecter que la position des objets mobiles. Dans la littérature, on propose principalement quatre méthodes de suivi d'objets sans modèle :

- La méthode du filtre d'association des données à probabilité jointe, appelée *"Joint Probabilistic Data Association Filter"* (JPDAF) en anglais (Bar-Shalom and Fortmann, 1988), traite efficacement le problème d'association des données à l'aide de filtres de Kalman. Lorsqu'une cible correspond à plusieurs points nouvellement mesurés, le filtre associé combine l'influence de l'ensemble des points correspondants, à l'aide de probabilités conditionnelles. Ce filtre cumule ainsi les informations relatives à plusieurs points. Dans ce cas, on utilise un filtre de même nature que celui associé à une cible ayant une correspondance avec un seul point nouvellement mesuré. Ce problème de cohérence est en général reproché à cette méthode (Zhang, 1993).

- La méthode des hypothèses multiples de suivi, *"Multiple Hypothesis Tracking"* (MHT) en anglais, traite également les cas d'association ambiguë. Cette méthode conserve toutes les informations relatives aux associations entre les cibles et les points nouvellement mesurés, puis attend de recevoir de nouvelles informations pour éliminer les associations incohérentes. Ces associations sont définies comme des hypothèses et correspondent à des filtres de Kalman distincts. Cette méthode maintient ainsi en parallèle, plusieurs mondes concurrents (ensembles d'hy-

pothèses incompatibles). Pour éviter une explosion combinatoire, cette méthode d'une part ne conserve que les hypothèses d'association les plus cohérentes, et d'autre part, fixe un seuil du passé (généralement 2 ou 3 images précédentes) des informations utilisées pour le calcul des associations. On reproche en général à cette méthode la complexité de sa mise en œuvre. Cependant, dans (Cox and Hingorani, 1996), on propose une implantation efficace des MHT.

- La méthode de stratégie de recherche par faisceaux, appelée *"beam search strategy"* en anglais (Zhang, 1993), duplique physiquement toutes les cibles (et les filtres de Kalman associés) correspondant à des associations ambiguës. Elle continue ensuite le suivi de ces cibles, jusqu'à ce que leur piste devienne incohérente ou se termine normalement. Pour établir la cohérence d'une piste, cette méthode utilise principalement l'estimation de l'erreur associée au filtre et la durée d'existence de la cible. Cette méthode examine ainsi plusieurs hypothèses sans maintenir de liens de concurrence entre elles. Ceci permet d'avoir une implantation simple de la méthode, mais peut poser des problèmes de cohérence. Par exemple, lorsqu'on souhaite étudier le comportement d'une cible, il est nécessaire de pouvoir différencier une cible issue d'une duplication, d'une cible réellement suivie.

- La méthode du filtre de Kalman distribué (Rao et al., 1993), appelée *"Decentralized Kalman Filter"* (DKF) en anglais, permet de combiner plusieurs filtres de Kalman pour augmenter la robustesse du suivi. Chaque filtre est associé de façon indépendante à un capteur et à un module de suivi, et réalise ses propres prédictions quant aux nouvelles positions des objets mobiles. Ensuite, une seconde étape combine ces prédictions pour obtenir un suivi global des objets mobiles. Les ambiguïtés sont ainsi résolues d'elles mêmes, par l'utilisation de différentes sources d'informations. L'intérêt de cette méthode est sa robustesse. Par exemple, une implantation efficace de cette méthode a été proposée dans le cadre du projet européen Esprit SKIDS. Le système avait pour but de suivre plusieurs individus s'entrecroisant avec des robots dans une pièce. Quatre caméras étaient disposées aux quatre angles de la pièce et des barrières optiques étaient placées aux endroits sensibles, en particulier près des zones de sortie. L'ensemble de ce système a permis de suivre efficacement des individus sur de longues séquences d'images. De même, dans (Hutber, 1995), l'auteur propose une méthode multi-capteurs comparable, développée dans la cadre du projet

Eureka Prometheus, pour l'équipement d'un véhicule routier. Cette méthode diffère de la précédente principalement dans le type de filtre de Kalman utilisé. Cette méthode utilise une modélisation précise du mouvement des véhicules, car ce mouvement est plus facilement modélisable que celui des individus. Ce modèle étant non linéaire, la méthode de suivi utilise des filtres de Kalman étendus (plus sophistiqués et pouvant effectuer un filtrage non forcément linéaire).

Ces méthodes de suivi d'objets sans modèle sont très génériques, car elles requièrent peu de contraintes d'utilisation. Elles peuvent ainsi être utilisées afin de suivre des objets mobiles non rigides. Cependant, comme elles ne possèdent pas d'informations sur les objets à suivre, elles peuvent mener à de nombreuses situations ambiguës et à une perte de cibles, lorsque ces situations deviennent trop difficiles à gérer.

4.2 Méthode proposée de suivi

Cette section a pour but de déterminer le type de méthode de suivi à utiliser. Plus précisément, il s'agit de déterminer si on doit utiliser une méthode générique (sans modèle) ou si on peut utiliser une méthode de suivi plus précise. Pour cela, on énumère d'abord les problèmes que la méthode de suivi doit gérer et les conditions d'utilisation où l'on souhaite utiliser cette méthode. À la suite de ces deux études, on justifie alors le choix du type de méthode et le modèle d'objet mobile retenu. Puis, on présente les deux sources d'informations permettant d'améliorer la méthode générale de suivi. En fin de section, on détermine un modèle de mouvement des objets mobiles, permettant de prédire la nouvelle position d'une cible.

4.2.1 Problèmes rencontrés

Un algorithme de suivi de régions mobiles doit prendre en compte deux types de problème : les problèmes de détection des régions mobiles et ceux plus classiques de leur suivi.

Les environnements des scènes traitées (p. ex. scènes intérieures et extérieures, scènes encombrées) posent de multiples **problèmes de détection** des régions mobiles tels que : des reflets, des ombres portées au sol, des manques de contraste de l'intensité lumineuse (p. ex. un fond possédant une couleur absorbante), des objets statiques de l'environnement en léger mouvement (p. ex. un arbre, une barrière, un escalier mécanique), des changements de luminosité (p. ex. une lumière que l'on allume, un feu clignotant),

des changements des conditions météorologiques (p. ex. un nuage, la pluie), la mise à jour de l'image de référence (p. ex. un objet mobile s'intégrant dans le fond de la scène)... Si ces problèmes de détection des régions mobiles ne sont pas complètement gérés par le module de traitement d'images, alors ils doivent être pris en compte au niveau du module de suivi.

Ensuite, l'algorithme de suivi de régions mobiles doit également régler les **problèmes inhérents à tout algorithme de suivi** :

- Apparition d'une cible non détectée auparavant et initialisation d'une piste.
- Disparition d'une cible suivie dans les images précédentes et terminaison d'une piste.
- Absence temporaire - partielle ou totale - d'une cible, due aux problèmes de détection des régions mobiles et suspension d'une piste.
- Occultation - partielle ou totale - d'une cible, due à la présence d'un objet statique de l'environnement (occultation dite statique) ou due à la présence d'un autre objet mobile (occultation dite dynamique).
- mélange du suivi de plusieurs cibles progressant les unes à côté des autres.

Comme conséquence de ces problèmes, une région mobile peut ainsi correspondre à un bruit dû à une mauvaise détection (p. ex. un artefact), un objet mobile (p. ex. un individu), une partie d'un objet mobile (p. ex. le bras d'un individu) et un groupe d'objets mobiles (p. ex. une foule).

4.2.2 Conditions d'utilisation

Dans les applications que l'on considère, une partie des objets mobiles traités (p. ex. êtres humains), ne possède pas certaines caractéristiques permettant d'utiliser des méthodes précises de suivi :

- Ces objets n'ont pas de primitives (p. ex. points caractéristiques, coins, arêtes) que l'on puisse suivre tout au long du traitement. Pour suivre ces objets, on ne peut alors pas se servir de telles primitives.
- Ces objets sont non rigides. On ne possède donc pas de modèle géométrique précis de leur forme et par conséquent, on ne peut pas utiliser de façon générale, une méthode de suivi utilisant un modèle *a priori* des objets mobiles.

– Ils peuvent avoir un mouvement irrégulier, changeant totalement d'une image à l'autre. Cette caractéristique est particulièrement vraie dans les applications d'analyse de sports collectifs (p. ex. matchs de football). Dans la méthode de suivi, on ne peut donc pas utiliser un modèle fiable du mouvement des objets.

Pour ces raisons, on ne peut pas utiliser les méthodes développées pour le suivi d'objets rigides.

De plus, on souhaite que la méthode de suivi soit générale - utilisable quelque soit l'angle de vue de la caméra et quelques soient les conditions d'acquisition des images (p. ex. en particulier, on souhaite pouvoir suivre un individu, même s'il est occulté sur une partie de la séquence d'images). On ne peut donc pas utiliser une méthode basant le suivi sur des modèles déformables de contour, comme celles proposées par (Meyer and Bouthemy, 1992), (Baumberg and Hogg, 1995). En effet, ces méthodes suivent globalement la forme des objets mobiles et permettent de compenser l'absence temporaire d'une partie de la forme de l'objet mobile par la présence de l'autre partie. Cependant pour des images bruitées, le suivi de la forme de l'objet mobile, même global, peut être profondément perturbé par l'absence répétée d'une partie de sa forme (p. ex. cas d'une occultation prolongée).

On souhaite également que la méthode de suivi puisse traiter des images de mauvaise qualité (p. ex. faible résolution), utilisant une seule caméra et qu'elle soit peu coûteuse en temps, pour pouvoir suivre en temps réel plusieurs objets mobiles à la fois. Pour ces raisons, les méthodes utilisant des modèles dynamiques comme celle proposée dans (Huttenlocker and Rucklidge, 1992), ne conviennent pas. De même, les méthodes utilisant plusieurs capteurs, comme celle proposée dans (Rao et al., 1993), ne conviennent pas également.

4.2.3 Présentation générale

Étant donné ces considérations, on a pour objectif de développer une méthode de suivi qui soit générique, rapide et robuste. Plus précisément :

– Générique : la méthode doit permettre de suivre n'importe quel objet mobile et en particulier les objets non rigides sans modèle *a priori*.

– Rapide : le suivi doit être temps réel et ainsi pouvoir traiter le flot d'images à la fréquence de leur arrivée (c.-à-d. de 1 à 25 images par seconde selon l'application).

- Robuste : elle doit pouvoir utiliser des images de mauvaise qualité (p. ex. faible résolution). Elle doit également gérer les problèmes classiques de suivi (p. ex. occultations statiques et dynamiques) et permettre de suivre les objets mobiles sur de longues séquences.

Développer une méthode générique de suivi est un objectif ambitieux, mais cette étape est nécessaire pour traiter certaines classes d'applications. En effet ces conditions peuvent paraître trop contraignantes, néanmoins elles correspondent aux conditions réelles d'utilisation de nombreuses applications. Une étude de marché a été réalisée dans ce sens au cours du projet européen Esprit HPCN PASSWORDS.

Pour atteindre cet objectif, on se propose alors de développer une méthode de suivi traitant de façon complète les situations ambiguës (comme le font les méthodes de suivi d'objets sans modèle) et n'utilisant comme seul modèle temporaire des objets mobiles, que la position et la taille des régions mobiles associées (comme le ferait une méthode rudimentaire de suivi d'objets non rigides). Dans ce cadre, on appelle **cible** toute région mobile suivie. Suivre une cible consiste à maintenir sa **piste** tout au long de la séquence d'images. L'utilisation de ce modèle temporaire simple des objets mobiles, permet à la fois d'avoir une précision suffisante du suivi et de pouvoir s'adapter à n'importe quelle condition d'utilisation. Puis dans un second temps, si les conditions d'utilisation le permettent, on peut alors affiner ce modèle temporaire, par exemple en tenant compte des couleurs des régions mobiles. L'utilisation de ce modèle temporaire permet ainsi de suivre l'**apparence** des objets mobiles (c.-à-d. les régions mobiles perçues dans la séquence d'images) plutôt que d'essayer de suivre les objets mobiles dans leur intégralité, avec la possibilité de se tromper. L'apparence d'un objet mobile est définie comme étant toute région mobile ou tout ensemble de régions mobiles proches les unes des autres. En cas de proximité de plusieurs régions mobiles, on suit alors une même région mobile de deux manières différentes :

- On suit individuellement la région mobile qui peut alors correspondre à l'apparence d'un bruit, d'un seul objet mobile ou d'un groupe d'objets mobiles ayant fusionnés.

- On suit globalement la région mobile qui est alors perçue comme appartenant à un groupe de régions mobiles. La région mobile peut correspondre à l'apparence d'une partie d'un objet mobile ou à l'apparence d'un objet mobile appartenant à un groupe se dissociant.

Cette définition de l'apparence, permet à une **cible** de pouvoir correspondre au suivi d'un bruit, d'un objet mobile, d'une partie d'un objet mobile

ou d'un groupe d'objets mobiles.

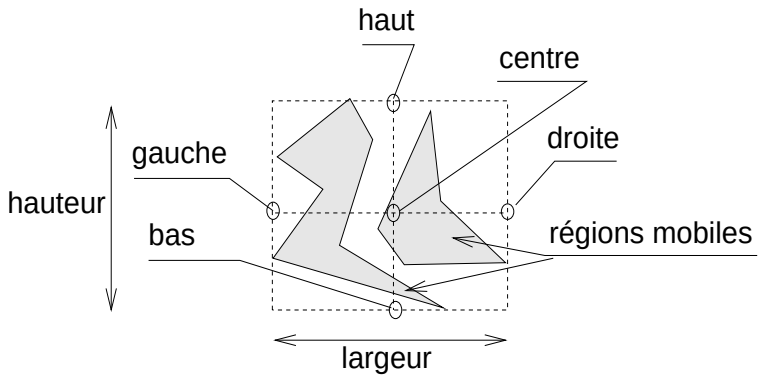

FIG. 4.1 – *Les cinq points génériques d'une région mobile : les milieux des quatre côtés de la boite englobante et le centre.*

On caractérise l'apparence d'un objet mobile par la donnée de ses dimensions (c.-à-d. hauteur et largeur) et de cinq points génériques décrits sur la figure 4.1 : le centre, le point du haut, du bas, de gauche et de droite de la boite englobante. Ces points sont dits génériques car ils sont définissables pour toute apparence d'objet mobile. Selon les résultats de la méthode de suivi, on peut également définir des points génériques supplémentaires. L'intérêt de cette méthode est de séparer le domaine du suivi, du domaine de l'interprétation qui est dépendant de l'application par nature. En effet, l'apparence des objets mobiles est une notion simple mais générique, permettant de ne conserver que ce qui est utile au suivi et d'éliminer ce qui est sujet à interprétation donc à erreur. On obtient ainsi un module de suivi plus générique, repoussant la prise des décisions liées aux modèles des objets mobiles au niveau du module de reconnaissance des scénarios. On ne prend de décision que lorsqu'on a suffisamment d'informations fiables sur l'objet. Ce gèle des décisions permet de ne pas avoir à gérer la consistance du suivi et donc de ne pas revenir en arrière sur les décisions passées. Cette nouvelle méthode se situe ainsi à la frontière des méthodes de suivi d'objets non rigides et des méthodes de suivi d'objets sans modèle.

Comme indiqué sur la figure 4.2, le suivi des cibles est réalisé au cours d'une boucle classique, constituée de trois étapes : prédiction, calcul des correspondances, mise à jour.

1) Au temps t, on prédit la nouvelle position des cibles existantes (c.-à-d. suivies

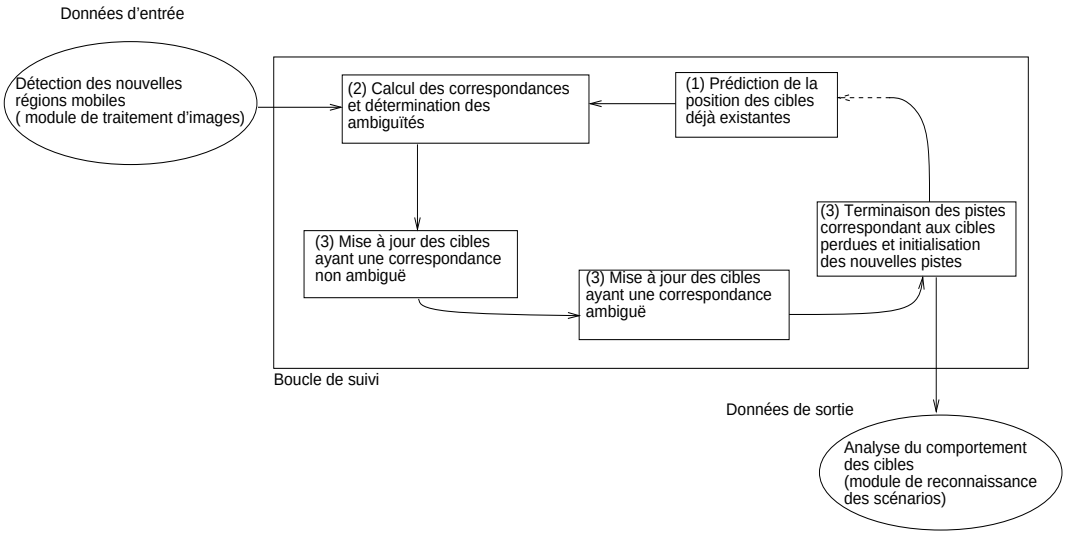

FIG. 4.2 – *Boucle de traitement de la méthode de suivi proposée.*

jusqu'au temps *t-1*).

2) On calcule les correspondances entre ces prédictions et les régions mobiles nouvellement détectées au temps t.

3) On met à jour les cibles ayant une nouvelle région mobile qui leur correspond.

Lorsqu'on ne connaît pas exactement la correspondance d'une cible déjà existante, on qualifie cette situation comme étant ambiguë. Cette boucle de traitement définit les étapes principales de la méthode de suivi.

4.2.4 Améliorations proposées

On qualifie cette méthode de suivi comme étant générique, car elle n'impose aucune condition particulière d'utilisation. Cependant une fois le cas générique traité, on peut envisager dans un second temps d'améliorer certaines étapes de la méthode de suivi à l'aide d'informations supplémentaires.

Pour cela, on utilise principalement deux sources d'informations :

- les informations contextuelles de la scène contenant en particulier les informations *a priori* de la scène,

- les résultats du module de reconnaissance de scénarios.

Premièrement, pour pallier en partie les problèmes de suivi, on utilise un ensemble d'informations sur la scène où se déroule le suivi des régions mobiles, appelé **contexte** (se référer au chapitre 2). Ces informations sont attachées à des zones prédéfinies de la scène, délimitant ainsi la portée d'utilisation des informations. Il existe de nombreuses façons d'utiliser ces informations :

- **zones de terminaison de pistes** : on les utilise pour déterminer quand on peut clore une piste et éliminer la cible associée (p. ex. *"zone de sortie"*).

- **zones d'initialisation de pistes** : on les utilise pour déterminer quand on peut débuter une piste et construire la cible associée (p. ex. *"zone d'entrée"*).

- **zones de changement de cibles** : elles permettent d'anticiper le changement de la cible suivie. Par exemple, si on détecte un individu disparaissant dans une rangée de voitures, on peut s'attendre à ce qu'il ressorte au volant d'une voiture et que sa forme ait beaucoup changée. On peut alors faire le lien entre la cible représentant l'individu et la cible représentant la voiture.

- **zones d'occultation statique** : elles préviennent des occultations avec des objets statiques de l'environnement. Elles indiquent le type de l'occultation et les méthodes pour la réparer. Par exemple, lorsqu'un individu est occulté partiellement par un banc et que le bas de l'individu est séparé de son haut, on utilise ces informations pour reformer l'individu en une seule région mobile. De même, lorsqu'un individu est occulté complètement par un pilier, on utilise le contexte pour prévoir la zone de réapparition de l'individu.

- **zones d'encombrement** : elles permettent d'anticiper le mélange de plusieurs pistes.

- **zones à trajectoires prédéfinies** : elles renseignent sur la forme attendue de la trajectoire des objets mobiles et sur leur vitesse. Ces informations permettent de corriger la prédiction des positions d'un objet mobile. Par exemple, si on sait qu'un individu est dans une zone à grande vitesse, on renforcera l'utilisation de sa vitesse pour calculer sa prochaine position. De même, si les pistes de deux individus se mélangent dans une zone de croisement (respectivement de couloir), on

s'attendra qu'ils se séparent en ayant croisé (respectivement en ayant gardé parallèle) leur trajectoire.

- **zones de bruit liées et se mélangeant aux objets mobiles suivis** (p. ex. ombre, reflet sur le sol) : elles permettent de corriger la forme globale de la cible. Par exemple, dans une zone d'ombre, elles permettent de prévoir que la taille des objets mobiles risque de s'allonger dans une certaine direction.

- **zones d'artefact non liées à des objets mobiles suivis** (p. ex. lumière clignotante, escalier mécanique, barrière, panneau publicitaire ou arbre en léger mouvement) : elles permettent de considérer certaines cibles comme correspondant totalement à du bruit et de les éliminer.

- **zones de bruit liées mais ne se mélangeant pas aux objets mobiles suivis** (p. ex. reflet à distance) : elles permettent également de considérer certaines cibles comme correspondant totalement à du bruit et de les éliminer.

Deuxièmement, pour améliorer la méthode de suivi, on utilise les résultats du module de reconnaissance de scénarios (se référer au chapitre 6). Ces résultats portent essentiellement sur quatre domaines :

- Ces résultats aident à la **résolution des correspondances ambiguës**. Lorsqu'une cible correspond à plusieurs régions mobiles nouvellement détectées, on établit les associations obtenant ainsi plusieurs cibles temporaires. On analyse ensuite le comportement des objets mobiles associés à ces cibles temporaires et on ne conserve que la cible temporaire reliée au comportement le plus vraisemblable. Cette méthode a pour effet de renforcer le suivi des cibles qui correspondent à un objet mobile ayant un comportement intéressant. L'utilisation du module de reconnaissance de scénarios permet ainsi de compenser la simplicité du modèle des objets mobiles.

- Ils permettent de **filtrer certaines valeurs aberrantes** utilisées par la méthode de suivi. En effet, cette méthode base son raisonnement sur le modèle des objets mobiles (c.-à-d. taille et position) et sur leur trajectoire, qui peuvent avoir temporairement des valeurs aberrantes. Le module de reconnaissance de scénarios effectuant un diagnostic sur ces valeurs permet d'éviter ces valeurs aberrantes et de maintenir des valeurs moyennes cohérentes.

- Ils indiquent si une cible peut **correspondre à du bruit** et être alors éliminée. Pour cela le module de reconnaissance de scénarios examine si les cibles ont un comportement correspondant à un bruit. Par exemple, ce comportement peut être régulier pour un bruit relatif à un feu clignotant, ou anarchique pour un reflet au sol.

- Dans certains cas, ils permettent d'indiquer le **type de l'objet mobile** associé à la cible. Cette information permet d'adapter la méthode de suivi en fonction du type de la cible. Par exemple, dans une application analysant les comportements d'individus et de véhicules, on peut utiliser par défaut la méthode de suivi générique sur tout objet mobile. Puis, si un objet est identifié comme étant un véhicule, on peut utiliser une méthode de suivi utilisant un modèle *a priori* de véhicule. Cette idée est utilisée dans (Prokopowicz et al., 1994) dont l'application cible a pour but de munir un robot d'un module de suivi pouvant réaliser trois tâches : repérage, suivi et poursuite d'objets mobiles. Les auteurs proposent d'utiliser différents types de méthode (p. ex. suivi de contour, utilisation de modèles temporaires) en fonction de la tâche à effectuer, du type d'objet mobile et du contexte de la scène.

Ces informations permettent également d'établir un coefficient de confiance dans le suivi effectué. Par exemple dans une zone d'encombrement, on s'attend à ce que le suivi soit peu fiable. Toutes ces informations permettent dans plusieurs cas de corriger des erreurs de détection ou de suivi et de compenser avantageusement la simplicité du modèle d'objet mobile utilisé.

4.2.5 Mouvement d'une cible

Cette section a pour but de définir le modèle de mouvement utilisé pour suivre les objets mobiles. Ce modèle repose sur la définition de la trajectoire d'une cible. Cette section commence par définir la représentation de la trajectoire utilisée, puis à justifier le choix du modèle de mouvement. À partir de ces considérations, on explique comment calculer la vitesse d'une cible. Enfin, on décrit les moyens utilisés pour déterminer la position d'une cible en corrigeant les erreurs potentielles de détection et de suivi.

La trajectoire d'une cible

On représente la trajectoire d'une cible par une approximation polygonale des positions de la cible. Elle est définie comme étant une succession de segments de droite dont les extrémités sont les positions de la cible, marquant

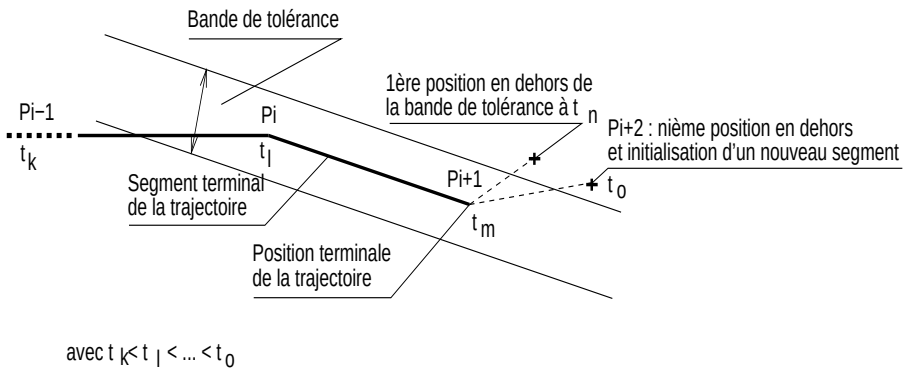

FIG. 4.3 – *Utilisation d'une bande de tolérance autour du segment terminal de la trajectoire d'une cible.*

un net changement de direction du mouvement. Un segment est construit à travers une phase d'initialisation et à travers une phase de prolongement. La phase d'initialisation consiste à rechercher deux positions de la cible suffisamment distantes l'une de l'autre. Une fois initialisé, on prolonge le segment avec la nouvelle position de la cible, si cette dernière se situe dans le prolongement du segment. La nouvelle position devient alors une des extrémités du segment. Si la nouvelle position n'est pas dans le prolongement du segment, on clôt ce segment et on en initialise un nouveau. Le dernier segment de la trajectoire est appelé segment terminal (ou segment courant). Pour chaque segment, on conserve les dates de création de chacune de ses extrémités, ce qui permet d'avoir une approximation de la vitesse moyenne de déplacement de la cible le long du segment. Dans le but d'obtenir des segments significatifs (suffisamment longs en distance et en durée), on utilise une bande de tolérance entourant le segment. Dans cette bande de tolérance le prolongement est considéré comme valide, exception faite des cas où la cible reste sur place ou retourne en arrière. De même, comme l'indique la figure 4.3, on accepte qu'un certain nombre de positions soient en dehors de la bande de tolérance avant de clore la phase de prolongement du segment. On ne prend alors pas en compte certaines positions qui peuvent correspondre à des positions aberrantes. On ne clôt un segment que lorsque le changement de direction est suffisamment prononcé.

La trajectoire d'une cible constitue ainsi l'historique de son suivi et rassemble une approximation de l'ensemble de ses positions passées.

Mouvement et vitesse de la cible

Le choix d'un modèle de mouvement de la cible est indépendant des autres choix réalisés pour la méthode de suivi. Tout au long de leur suivi, les cibles ne possèdent pas un mouvement appartenant à un modèle constant et précis. Leur direction et leur vitesse peuvent changer radicalement entre deux instants successifs. Pour cette raison, on suppose que le mouvement de la cible est *a priori* linéaire. Même si cette hypothèse se confirme et que le mouvement de la cible est régulier, on n'essaie pas d'affiner cette approximation avec des termes mesurant la rotation de son mouvement ou son accélération, parce que la précision de la détection des régions mobiles ne le permet pas. Si l'hypothèse de mouvement linéaire se révèle erronée (cas d'un changement de direction), on considère que le mouvement change de forme linéaire. On considère ainsi que le mouvement de la cible est linéaire par morceaux. L'historique de ce mouvement est alors complètement conservé à travers la représentation de la trajectoire.

Pour calculer la vitesse de la cible, on se base sur ce modèle du mouvement (c.-à-d. représenté par la trajectoire), en considérant que la vitesse est constante le long des segments et qu'elle change au niveau des extrémités. Le vecteur vitesse d'une cible est obtenu à partir du segment terminal de la trajectoire de la piste (à partir des dates de création de ses extrémités et de la longueur du segment). On appelle cette vitesse, *vitesse courante*. Comme les segments sont en général de faible longueur, cette vitesse est calculée sur un faible intervalle de temps et correspond ainsi à une vitesse instantanée.

Position d'une cible

On appelle position d'une cible le barycentre de ses cinq points génériques (se référer à la figure 4.1). On calcule la position de ces points génériques, au cours de la phase de suivi de la cible et plus précisément, au cours de l'association de la cible avec la région mobile nouvellement détectée lui correspondant. Lors de cette opération, on effectue séparément la mise en correspondance de chacun des points génériques avec les points correspondants de la région mobile nouvellement détectée : on prédit la nouvelle position des points génériques (la méthode de prédiction est décrite ci-après, dans la section 4.3.1), puis on compare cette position prédite avec la position du point correspondant de la région mobile nouvellement détectée. Ensuite, on sélectionne le point générique de la cible ayant la meilleure correspondance et on en déduit sa nouvelle position. Si les autres points génériques ont également une bonne correspondance, on utilise cette correspondance pour calculer leur

position. Sinon, si un autre point générique a une mauvaise correspondance, on estime sa position à partir du point sélectionné (c.-à-d. ayant la meilleure correspondance) et des dimensions de la cible. L'estimation des dimensions de la cible (c.-à-d. sa hauteur et sa largeur) est un problème délicat, lié au calcul des points génériques. Pour cette estimation, on utilise deux techniques : on calcule la moyenne des valeurs afin d'éviter de trop grandes irrégularités et on ne tient pas compte des valeurs aberrantes. Pour déterminer qu'une valeur est aberrante, on utilise les informations contextuelles de la scène (p. ex. indication de la présence d'une occultation) et une phase de diagnostic permettant de prédire l'évolution des dimensions de la cible.

Chaque point générique de la cible peut ainsi être calculé indépendamment des autres points. Si un point générique est mal détecté, sa position est estimée à partir des autres points. Cette méthode est ainsi peu sensible aux occultations partielles ou au changement de forme des cibles de courte durée. C'est un des moyens pour traiter les cas d'occultation partielle. Elle renseigne également sur la nature d'une occultation partielle d'une cible. Les points génériques de la cible ayant une mauvaise mise en correspondance de leur position donnent un indice sur la présence et la localisation d'une éventuelle occultation.

4.3 Les trois étapes

Cette section a pour but de présenter les trois étapes du suivi d'une cible : (1) une étape de prédiction de la nouvelle position des cibles déjà existantes, (2) une étape de calcul des correspondances entre les cibles déjà existantes et les régions mobiles nouvellement détectées, puis (3) une étape de mise à jour des cibles non ambiguës à l'aide de ces correspondances. Dans la première partie de cette section, on commence par expliquer comment prédire la nouvelle position d'une cible. Ensuite dans la seconde partie, on définit une distance entre cibles et régions mobiles, servant au calcul des correspondances, et permettant de construire une matrice d'ambiguïté, utilisée afin de comparer globalement toutes les correspondances et de déterminer les cibles ambiguës. Enfin, dans la troisième partie de cette section, on décrit les différents cas de mise à jour d'une cible non ambiguë : maintien, initialisation et terminaison d'une piste.

4.3.1 Prédiction d'une nouvelle position

L'étape de prédiction sert principalement à estimer la position d'une cible à l'instant suivant pour la comparer avec la nouvelle position mesurée. Le

calcul de la prédiction est basé sur la vitesse courante de la cible. Si la trajectoire est déterminée avec un degré de vraisemblance (appelé aussi degré de confiance) suffisant, on utilise cette vitesse. C'est le cas, dès que le mouvement de la cible est suffisamment régulier (4 à 5 positions approximativement dans la même direction dans le cas d'un individu en train de marcher, représentant une durée d'une seconde). Dans ce cas, cette méthode donne de bons résultats. Dans le cas contraire, quand le degré de vraisemblance est faible, on prend comme position prédite du point de la cible son ancienne position. Le degré de vraisemblance de la trajectoire est calculé à l'aide de la théorie sur les ensembles flous. Il est établi grâce à une étape de diagnostic, réalisée par le module de reconnaissance des scénarios (cf. chapitre 6).

Généralement, les modules de suivi d'objets rigides utilisent le filtre de Kalman pour estimer la position d'une cible. Cette utilisation dans le cas de suivi d'êtres humains soulève plusieurs objections :

- La principale objection est due à la nature du déplacement d'un être humain. Ce déplacement se caractérise par des phases de mouvements réguliers, entrecoupées de mouvements erratiques correspondant à des phases de changements. Pendant les phases de mouvements réguliers, la valeur du bruit relatif au calcul du modèle de l'état du filtre de Kalman doit être très faible afin de renforcer l'influence de la prédiction de la nouvelle position. L'état du filtre est alors principalement obtenu à partir de cette prédiction. Pendant les phases de mouvements erratiques, la valeur du bruit relatif au calcul du modèle de l'état du filtre de Kalman doit être très importante afin de donner plus de poids aux nouvelles mesures. L'état du filtre est alors principalement obtenu à partir de ces nouvelles mesures. Ces deux valeurs du bruit relatif au calcul du modèle de l'état sont antagonistes. Or, ce changement de phases est aléatoire et est difficilement modélisable à l'aide d'un filtre de Kalman (Rao et al., 1993). L'utilisation d'un filtre de Kalman peut alors conduire à des erreurs de prédiction pendant ces phases. Pour résoudre ce problème, on peut par exemple (Hutber, 1995), décider de relier toute cible perdue avec une cible réapparaissant à proximité. Une autre méthode consiste à maintenir en parallèle plusieurs filtres de Kalman, et à utiliser celui dont l'incertitude est la plus faible. Par exemple, A. Pentland utilise 15 filtres modélisant tous les mouvements possibles d'un conducteur à son volant (Pentland, 1995). Cette méthode ne peut s'appliquer que dans des environnements très contraints, tels que l'habitacle d'un véhicule, où le nombre de mouvements différents est faible.

Par contre, l'utilisation de la trajectoire permet d'adapter le calcul de la vitesse (donc de la prédiction), en fonction de ces phases de changements.

- Une seconde objection est due à l'imprécision de la détection des régions mobiles. Cette imprécision est telle que si on incorpore une donnée aberrante, le calcul du mouvement peut être modifié de façon significative, perturbant le calcul des positions prédites. Si le changement de mouvements est très important, il est alors nécessaire de vérifier que ce changement ne correspond pas à une erreur de détection. Donc dans ce cas, l'utilisation d'un filtre de Kalman nécessite l'utilisation d'un contrôle préalable des données fournies au filtre.

- Une troisième objection vient de la prédiction des positions des cibles sur des grands intervalles de temps. Lorsqu'une cible est perdue (ou occultée) sur plusieurs images, on souhaite pouvoir prédire sa nouvelle position à partir de la date de sa disparition et en tenant compte du temps de suspension de son suivi. Dans le cas d'un filtre de Kalman, il n'existe qu'une seule façon de tenir compte de ce temps de suspension : prendre comme mesures les positions prédites. Dans le cas de l'utilisation de la trajectoire, on peut adapter la prédiction en utilisant différentes vitesses selon l'importance du temps de suspension : vitesse instantanée (calculée à partir du dernier segment de la trajectoire) pour un temps court, vitesse moyenne (calculée à partir d'un plus grand nombre de segments) pour un temps plus long. Ces calculs sont expliqués par la suite.

- Enfin les modèles que l'on possède sur les cibles sont très génériques et très pauvres. Une cible doit pouvoir correspondre aussi bien au suivi d'une région mobile qu'au suivi d'un ensemble de régions mobiles. L'utilisation d'un filtre de Kalman peut alors conduire à des incohérences de modèle. Par exemple, quand deux cibles fusionnent et donnent naissance à une nouvelle cible unique, il est alors nécessaire de combiner les filtres des cibles se mélangeant pour obtenir un filtre relatif à la nouvelle cible. La construction d'un filtre combinant les mesures de plusieurs cibles est également utilisée par la méthode du JPDAF (Bar-Shalom and Fortmann, 1988). Ce problème de cohérence est en général reproché à cette méthode (Zhang, 1993).

Malgré ces objections, le filtre de Kalman est utilisé dans de nombreux systèmes pour estimer la nouvelle position d'êtres humains en mouvement et

peut donner de bons résultats (Rao et al., 1993), (Choi et al., 1997), (Intille and Bobick, 1995). Cependant, l'utilisation du filtre de Kalman dans ces systèmes se justifie par des conditions d'utilisations plus souples (p. ex. systèmes multi-capteurs, bonne résolution des images). Par ailleurs, ces systèmes combinent en général plusieurs méthodes pour pallier les erreurs de prédiction.

La méthode de prédiction des positions qui est proposée dans ce chapitre, n'utilise pas de filtre de Kalman. À la place, on calcule la vitesse courante d'une cible à partir de sa trajectoire et cette méthode donne des résultats satisfaisants. Le principal avantage de la méthode de prédiction utilisant la trajectoire est la possibilité d'adapter la prédiction selon le contexte, le type de cible. Par exemple, lorsque une cible est occultée sur un grand laps de temps, on calcule sa vitesse moyenne comme indiqué ci-après sur la figure 4.6, en prenant en compte plusieurs segments du passé lointain de la trajectoire. Ce point est particulièrement important lorsqu'on souhaite traiter différents types de scène avec la même méthode de suivi.

Néanmoins, la méthode de prédiction des positions est indépendante du reste du module de suivi des régions mobiles. Il est donc envisageable, dans des travaux futurs, de combiner différentes méthodes de prédiction (filtre de Kalman et trajectoire).

4.3.2 Calcul des correspondances

Distance entre une cible et une région mobile

Pour calculer une correspondance entre une cible existante à l'instant *t-1* et une région mobile nouvellement détectée à l'instant t, on regarde deux critères :

- **Distance spatiale** : on détermine d'abord si la position prédite de la cible est proche de la position de la région mobile nouvellement détectée. Plus exactement, on calcule la position prédite des points génériques de la cible et on détermine dans quelle mesure, la région définie par ces points génériques se superpose à la région mobile. S'il n'y a pas superposition entre les deux régions, on calcule la distance qui les sépare, que l'on pondère par leur taille. Ce calcul de distance entre la cible projetée à l'instant t et la région mobile nouvellement détectée est traduit par la formule suivante :

si intersection(cible projetée, région mobile) $\neq \emptyset$
alors distance = 1 - $\frac{intersection}{union}$(cible projetée, région mobile)
sinon distance = 1 + $\frac{distance}{taillemaximale}$(cible projetée, région mobile)

La distance entre les régions est calculée comme étant la distance minimale entre deux bords de chacune des boites englobant les régions. Cette distance n'est pas au sens strict une distance car elle ne vérifie pas l'inégalité triangulaire. Cette propriété n'est pas vérifiée pour permettre de tenir compte de la proportion de recouvrement des régions. Cette définition de la distance permet de relativiser le calcul de la distance en fonction des surfaces des régions. L'utilisation de ce critère permet de définir une zone (un disque) de probabilité de présence autour de la cible projetée à l'instant t. Une amélioration de cette méthode consiste à utiliser d'autres formes de zones de probabilité de présence. Par exemple, dans le système Soccer (Schirra and Stopp, 1993), on propose d'utiliser une zone délimitée par une ellipse centrée sur la position prédite et dont le grand axe est dans le sens du mouvement.

– **Similarité des caractéristiques** : On calcule également la distance entre les caractéristiques de la cible et de la région mobile nouvellement détectée. En ce qui concerne l'implantation actuelle du module de suivi développé, ces caractéristiques sont principalement la taille des régions (hauteur et largeur) et secondairement leurs couleurs. Le choix de ces caractéristiques dépend du module de traitement d'images (car il les calcule) et l'importance de ces caractéristiques, les unes par rapport aux autres, dépend de l'application (car elle établit leur pouvoir discriminant). Par exemple, dans nos applications cibles (métros, parkings), les couleurs des régions mobiles sont en générales peu discriminantes car les individus suivis sont souvent de mêmes couleurs sombres. On peut envisager de calculer d'autres caractéristiques, afin d'obtenir un critère de similarité efficace. Dans ce but, le module de suivi présenté dans ce rapport permet l'ajout de caractéristiques supplémentaires. En ce qui concerne les modules de suivi développés par la communauté scientifique, d'autres caractéristiques sont utilisées. Pour le suivi d'objets rigides, en particulier pour les véhicules, ces caractéristiques correspondent aux propriétés des modèles des objets rigides (Koller et al., 1993), (Du et al., 1993). Pour le suivi d'objets non rigides, ces caractéristiques sont le contour des régions mobiles (Baumberg and Hogg, 1995), la disposition des couleurs à l'intérieur des régions mobiles (Azarbayejani et al., 1996) ou la disposition des arêtes contenues dans les régions mobiles (Huttenlocker and Rucklidge, 1992). Ces caractéristiques supplémentaires sur les objets non rigides nécessitent une bonne qualité des séquences d'images et ne sont donc pas calculables dans les

applications cibles choisies.

Une fois ces deux critères calculés, il s'agit de déterminer l'importance d'un critère par rapport à l'autre. Comme l'objectif est de développer un module de suivi générique, le critère de similarité des caractéristiques n'est pas suffisamment fiable et peut conduire à des erreurs de mise en correspondance. Par conséquent, dans l'implantation actuelle du module de suivi développé, on a donné plus d'importance au critère de distance spatiale. Cet ajustement des poids des critères est un des paramètres du module de suivi, dépendant du type d'application.

La matrice d'ambiguïté

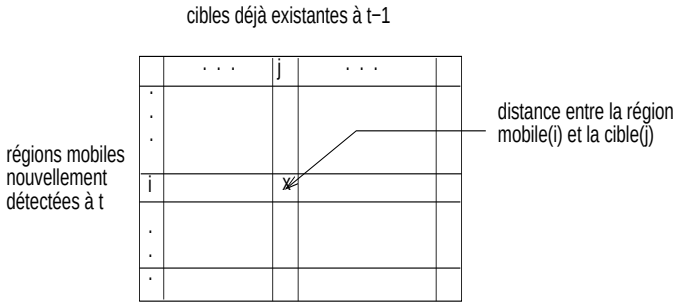

FIG. 4.4 – *La matrice d'ambiguïté.*

Pour comparer l'ensemble des distances calculées entre les cibles déjà existantes et les régions mobiles nouvellement détectées, on utilise une matrice, appelée généralement matrice d'ambiguïté ou matrice des associations (Cox and Hingorani, 1996). Comme l'illustre la figure 4.4, les colonnes de la matrice représentent les cibles déjà existantes à l'instant *t-1* et les lignes représentent les régions mobiles nouvellement détectées à l'instant *t*. Les éléments de la matrice mesurent la distance entre une cible et une région mobile nouvellement détectée.

La fonction principale de la matrice est de déterminer le nombre de régions mobiles nouvellement détectées qui sont proches d'une cible (c.-à-d. distance suffisamment faible). Selon ce nombre, une cible peut être dans quatre états différents :

– État **visible** : si une seule région mobile nouvellement détectée lui correspond et si cette région mobile ne correspond pas à d'autres cibles.

- État **perdu** : si aucune région mobile nouvellement détectée lui correspond.

- État **occulté** : si aucune région mobile nouvellement détectée lui correspond et si la cible se situe dans une zone possédant une information contextuelle indiquant qu'il y a une possibilité d'occultation statique. Ici le terme d'occultation pour une cible ne concerne que les occultations statiques et totales. Le cas des occultations partielles est traité par l'étape de mise en correspondance, qui admet les correspondances partielles. Le cas des occultations dynamiques est traité à travers les cibles ambiguës (voir ci-après, section 4.4).

- État **ambigu** : s'il existe plusieurs régions mobiles nouvellement détectées qui lui correspondent ou s'il existe une seule région mobile nouvellement détectée qui lui correspond et que cette région mobile correspond également à une autre cible.

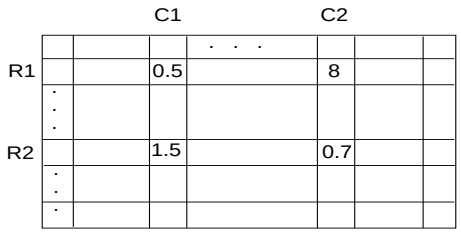

FIG. 4.5 – *La cible C1 est ambiguë car elle est suffisamment proche des régions mobiles nouvellement détectées R1 et R2.*

L'utilisation de la matrice d'ambiguïté permet de raisonner globalement sur les correspondances entre cibles et régions mobiles nouvellement détectées. Plus précisément, elle permet d'ajuster ces correspondances de deux manières différentes :

- Elle permet, par exemple, de réduire le nombre de cibles ambiguës. Sur la figure 4.5, la cible C1 est très proche de la région mobile R1, moyennement proche de R2 et éloignée des autres régions. La cible C1 devrait donc être considérée comme une cible ambiguë. Cependant, la cible C2 est très proche de R2 et beaucoup plus loin des autres régions. Dans cette situation, on considère que C1 est uniquement associée à R1 et que C2 est uniquement associée à R2. Pour cela, on augmente la distance entre C1 et R2. La cible C1 n'est plus alors considérée comme

une cible ambiguë. Une fois que toutes les correspondances sont calculées, on effectue ainsi une phase d'ajustement pour toutes les situations permettant de réduire le nombre d'ambiguïtés. Ces ajustements ne sont malheureusement pas toujours valides et doivent être réglés en fonction de la qualité d'acquisition des séquences d'images : plus les images sont de bonne qualité et plus on peut se permettre d'ajuster des situations.

– L'utilisation de la matrice d'ambiguïté permet également d'éviter de perdre certaines cibles déjà existantes. Un impératif pour le module de suivi est de ne pas perdre les cibles ayant un important degré de vraisemblance. Ces cibles sont considérées par le système d'interprétation comme correspondant à des objets mobiles intéressants. C'est le cas des cibles ayant une durée de vie importante (c.-à-d. elles sont suivies depuis suffisamment longtemps). Pour ne pas perdre ces cibles, le module de suivi favorise leurs correspondances avec des régions mobiles nouvellement détectées qui ne leur sont pas trop éloignées (c.-à-d. on diminue leur distance dans la matrice d'ambiguïté). Ces ajustements peuvent être également dangereux et sont paramétrés dans le module de suivi, en fonction de la qualité d'acquisition des séquences d'images.

Dans le calcul des éléments de la matrice d'ambiguïté, on ne tient compte que des cibles visibles ou occultées. Ceci permet de donner une priorité à ces cibles effectivement suivies. Le calcul des correspondances entre les cibles perdues ou ambiguës et les régions mobiles nouvellement détectées est effectué ultérieurement dans l'algorithme de suivi, après la mise à jour des cibles effectivement suivies.

4.3.3 Mise à jour des cibles non ambiguës

Cette étape de mise à jour des cibles non ambiguës consiste de façon générale à gérer les pistes associées aux cibles : maintenir, initialiser ou clore les pistes selon l'état des cibles.

Maintien d'une piste

On maintient une piste si la cible associée est visible, (totalement) occultée ou si elle est perdue depuis peu. Dans le cas d'une cible visible, le maintien de la piste consiste à prolonger la trajectoire de la cible, avec la région mobile nouvellement détectée lui correspondant. Dans le cas d'une cible occultée ou perdue depuis peu, on suspend sa trajectoire et on attend les

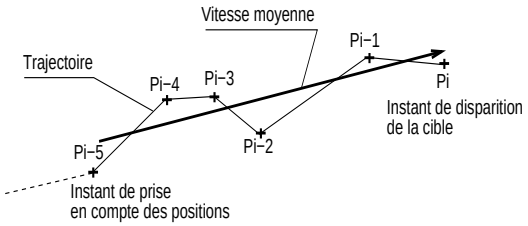

FIG. 4.6 – *La vitesse moyenne est calculée à partir de la composante principale du nuage de points, constitué des extrémités des segments de l'approximation polygonale de la trajectoire. Le nombre de points pris en compte, dépend du temps de suspension de la cible. Plus ce temps est important, plus on tient compte du passé.*

images suivantes pour voir si la cible réapparaît. En attendant la réapparition de la cible, on comptabilise le temps de suspension de la trajectoire. Au cours des images suivantes, lorsqu'on calcule les correspondances entre cette cible (occultée ou disparue depuis peu) et les régions mobiles nouvellement détectées, on prédit la position de réapparition de la cible soit à l'endroit de sa disparition, soit à l'endroit où elle aurait dû être, si elle avait maintenu une vitesse constante pendant son temps de suspension. Si ce temps de suspension est court, on utilise la vitesse courante calculée à partir du segment terminal de la trajectoire de la cible. Sinon, la vitesse utilisée est une vitesse moyenne, calculée à partir des derniers segments de la trajectoire comme le montre la figure 4.6. Dans le cas d'une occultation statique, le contexte à un rôle important pour déterminer l'endroit de réapparition de la cible (Brémond and Thonnat, 1996a). Il indique si la cible est plutôt attendue à l'endroit de sa disparition, ou s'il faut tenir compte du déplacement de la cible. Il permet également d'indiquer si ce déplacement est uniforme ou s'il faut s'attendre à un changement de direction de la cible (se référer au chapitre 2).

Initialisation d'une nouvelle piste

On initialise de nouvelles pistes à partir des régions mobiles nouvellement détectées, qui n'ont pas encore été associées avec des cibles déjà existantes. D'abord, on vérifie que cette région mobile non associée ne correspond pas à une cible occultée ou perdue. Si c'est le cas, on établit les correspondances et on met à jour la trajectoire de la cible, ainsi redevenue visible. Si ce n'est pas le cas, avant d'initialiser une nouvelle piste, on calcule trois types de critère

sur le contexte de la zone où est apparue la région mobile non associée et sur la forme de la région mobile :

- La zone peut contenir du bruit (p. ex. présence de reflets). Si c'est le cas, on élimine la région mobile non associée.
- La zone est considérée comme une zone d'apparition d'objets mobiles. Par exemple, les zones du bord de l'image où près d'une entrée peuvent être des zones d'apparition d'objets mobiles. Si c'est le cas, on initialise la piste.
- La forme de la région mobile non associée peut correspondre à du bruit. Par exemple, lorsque la taille de la région mobile est suffisamment petite, elle peut correspondre à du bruit et dans ce cas, on élimine la région mobile non associée.

Ces critères ont une valeur entière, quantifiant l'influence du critère calculé pour un cas donné, et sont pondérés par un coefficient, quantifiant l'influence du type de critère. Ces coefficients sont prédéfinis et dépendants de l'application. Par exemple, le type de critère relatif aux zones d'apparition d'objets mobiles est particulièrement important et possède un coefficient élévé. Selon la combinaison de ces critères, on élimine la région mobile non associée ou on crée une nouvelle cible et on initialise sa piste associée.

Terminaison d'une piste

Pour que la mise à jour des pistes soit complète, il reste à traiter les cas de terminaison d'une piste. Cette terminaison de pistes concerne (1) les cibles occultées ou perdues depuis un certain laps de temps. Pour déterminer l'opportunité de l'élimination d'une cible, on calcule principalement cinq types de critère :

- La nature de la cible. On élimine plus facilement une cible perdue qu'une cible occultée, car cette dernière est plus susceptible de réapparaître.
- Le temps de suspension du suivi. Plus ce temps est important, plus on a tendance à éliminer la cible.
- Si la zone où a disparu la cible est une zone de terminaison. Cette information est indiquée par le contexte de la zone. Si c'est le cas, on élimine rapidement la cible perdue.

- Si la zone où a disparu la cible est une zone d'occultation. Cette information est également indiquée par le contexte de la zone. Si c'est le cas, on conserve plus longtemps la cible en question.

- Le degré d'importance que représente la cible pour le système d'interprétation. Ce degré est établi par le module de reconnaissance de scénarios. Si la cible correspond à un objet mobile impliqué dans un scénario jugé intéressant par ce module, alors on conserve la cible (se référer au chapitre 6). Ce critère est un retour d'informations du module de reconnaissance de scénarios vers le module de suivi.

En combinant l'ensemble de ces critères, on décide si on élimine la cible et termine la piste associée ou si on continue de maintenir son suivi.

La terminaison de pistes concerne également dans certaines situations (2) les cibles visibles. Cette élimination est due à la nature des correspondances entre ces cibles et la scène. Effectivement, une cible peut correspondre à du bruit (p. ex. un reflet), sans que le module de suivi puisse le déterminer. Cependant, le module de reconnaissance des scénarios possède la connaissance pour déterminer si une cible correspond à du bruit. Pour cela, il analyse le comportement de l'objet mobile associé à la cible. Si ce comportement correspond au comportement d'un bruit, la cible est éliminée même si elle est visible. Cette méthode est un deuxième effet retour du module de reconnaissance de scénarios vers le module de suivi.

Caractéristique générale de la mise à jour

L'étape de mise à jour des cibles est une étape classique dans tout module de suivi. Ce qui fait l'originalité de ce module de suivi est l'adaptation de cette étape à plusieurs caractéristiques du système :

- le contexte de l'application,

- les propriétés de la cible (p. ex. son état, sa taille, sa durée d'existence),

- l'historique du suivi de la cible contenu dans sa trajectoire,

- les résultats de l'interprétation du comportement de l'objet mobile associé à la cible.

Cette adaptation s'effectue à travers la combinaison de différents critères. L'utilisation de critères similaires par une méthode de suivi n'est pas une idée

nouvelle dans la communauté scientifique. Par exemple, dans (Hutber, 1995), on définit également des zones de terminaison de pistes. Dans la méthode de recherche par faisceaux (Zhang, 1993), on calcule le support d'existence d'une cible pour déterminer si la cible est cohérente et si on peut l'éliminer. Cependant, ce support d'existence est uniquement calculé à partir d'informations provenant des propriétés du suivi de la cible. Ce qui caractérise la méthode de suivi présentée ici, est la généralisation de ce principe à d'autres sources d'informations (p. ex. contexte, analyse des comportements).

4.4 Résolution des situations ambiguës

Cette section expose les différentes techniques de résolution des ambiguïtés rencontrées au cours du processus du suivi. L'efficacité obtenue pour résoudre ces ambiguïtés établit la robustesse de la méthode de suivi. La tâche centrale de cette résolution consiste à récupérer les cibles ambiguës, car elles contiennent l'historique du suivi des objets mobiles. Ces cibles sont donc les cibles les plus pertinentes pour le système d'interprétation de séquences d'images. Cette section commence par définir les cibles composées qu'on utilise comme support de résolution des situations d'ambiguïté. On énumère ensuite tous les cas où l'on essaie de résoudre ces situations d'ambiguïté, ainsi que les techniques employées. Le nombre de cas à traiter est important, car il permet d'adapter finement la méthode générale de suivi et d'obtenir alors de bons résultats. Puis on propose différentes améliorations à expérimenter, permettant de généraliser ce traitement par cas. Enfin, on compare ces techniques de résolution aux autres méthodes de suivi.

4.4.1 Cibles composées

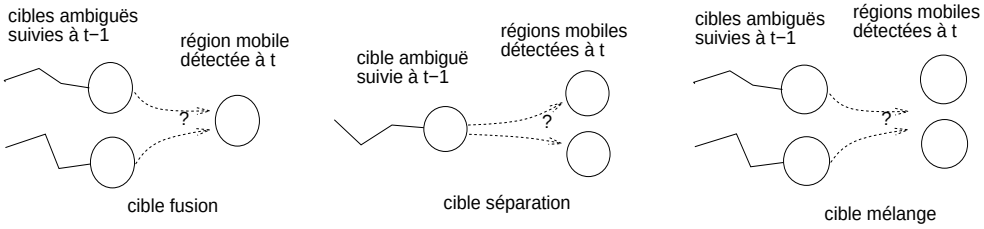

FIG. 4.7 – *Les trois types de cible composée*

Une cible ambiguë est une cible déjà existante qui correspond à plusieurs régions mobiles nouvellement détectées ou qui correspond à une région mobile

nouvellement détectée ayant d'autres correspondances. Pour résoudre une ambiguïté, on définit une cible composée qui rassemble toutes les entités intervenant dans la situation d'ambiguïté. Comme le montre la figure 4.7, il existe trois types d'ambiguïté définissant trois types de cible composée :

- Type **séparation** : on définit une cible composée de type séparation dans la situation où, une seule cible suivie à *t-1* correspond à plusieurs régions mobiles nouvellement détectées. La cible séparation est alors composée de la cible ambiguë s'étant divisée et d'un ensemble de cibles temporaires correspondant aux régions mobiles nouvellement détectées.

- Type **fusion** : on définit une cible composée de type fusion dans la situation où, plusieurs cibles suivies à *t-1* correspondent à la même région mobile nouvellement détectée. La cible fusion est alors composée de l'ensemble des cibles ambiguës ayant fusionnées et d'une cible temporaire correspondant à la région mobile nouvellement détectée.

- Type **mélange** : on définit une cible composée de type mélange dans la situation où, plusieurs cibles suivies à *t-1* correspondent à plusieurs régions mobiles nouvellement détectées. La cible mélange est alors composée de l'ensemble des cibles ambiguës s'étant mélangées et d'un ensemble de cibles temporaires correspondant aux régions mobiles nouvellement détectées.

Une cible composée rassemble ainsi un ensemble de cibles ambiguës et un ensemble de régions mobiles nouvellement détectées. Dans une situation d'ambiguïté, chaque région mobile nouvellement détectée donne naissance à une cible que l'on appelle **cible temporaire**. Le principe d'une cible composée consiste à suspendre le suivi des cibles ambiguës et à suivre les cibles temporaires, jusqu'à ce qu'on ait suffisamment d'informations pour résoudre l'ambiguïté. Lorsque ces informations sont disponibles et qu'on est en mesure d'établir les correspondances avec les cibles ambiguës, on met à jour ces cibles ambiguës et on reprend leur suivi. À ce moment là, les cibles temporaires deviennent inutiles et sont éliminées. Dans une cible composée, les cibles ambiguës contiennent l'historique du suivi des régions mobiles jusqu'à l'instant où l'ambiguïté a été rencontrée, tandis que les cibles temporaires représentent les traces courantes des régions mobiles suivies. On utilise les cibles composées pour geler la prise de décision quant à l'association des cibles ambiguës aux cibles temporaires.

Cette décision d'association est prise ultérieurement, soit lorsqu'on est en mesure de résoudre l'ambiguïté ou soit lorsqu'on est obligé de le faire. La

première situation est rencontrée quand on est en possession d'informations supplémentaires permettant de résoudre l'ambiguïté. La seconde situation est rencontrée par exemple, quand une seconde ambiguïté survient obligeant de résoudre la première. Les deux paragraphes suivants décrivent ces situations.

4.4.2 Première situation de résolution d'une ambiguïté

Pour déterminer si la première situation est rencontrée, on calcule à chaque traitement d'une nouvelle image, les correspondances entre les cibles ambiguës et les cibles temporaires d'une même cible composée. Pour chaque cible composée, on calcule ainsi une matrice d'ambiguïté dont les colonnes représentent les cibles ambiguës et les lignes représentent les cibles temporaires. Puis on procède comme pour le calcul des correspondances entre cibles et régions mobiles nouvellement détectées, en utilisant deux critères : distance spatiale et similarité des caractéristiques. Lorsqu'une correspondance entre une cible ambiguë et une cible temporaire n'est plus ambiguë, on associe ces deux cibles et on met à jour la cible ambiguë. Cette mise à jour consiste principalement à prolonger la trajectoire de la cible ambiguë à partir de la trajectoire de la cible temporaire. On récupère ainsi la portion de trajectoire de la cible ambiguë, correspondant à la suspension du suivi de la cible pendant la situation d'ambiguïté. Une fois cette mise à jour réalisée, on élimine la cible temporaire. Si toutes les cibles ambiguës ont pu être associées, on élimine également la cible composée. De cette manière, l'ambiguïté représentée par une cible composée se résout sans approximation, car la trajectoire de la cible ambiguë a été complètement retrouvée.

4.4.3 Deuxième situation de résolution d'une ambiguïté

On force la résolution d'une situation d'ambiguïté et de sa cible composée associée, lorsque :

- le temps de suspension des cibles ambiguës est trop important,
- les cibles temporaires de la cible composée s'éparpillent,
- la cible composée participe à une seconde ambiguïté.

Dans ces cas là, on s'oblige à résoudre l'ambiguïté (ou la première ambiguïté selon le cas), pour forcer l'établissement des correspondances suffisamment tôt, évitant ainsi de raisonner sur des informations trop anciennes. Le but de cette opération est de récupérer les cibles ambiguës (représentant

l'historique du suivi) et de leur trouver une correspondance avec une cible temporaire (représentant l'état actuel du suivi).

Le temps de suspension est trop important

Lorsque le temps de suspension des cibles ambiguës est trop important, il devient alors difficile de récupérer les cibles ambiguës, car les informations les concernant deviennent trop vieilles pour être utilisées dans le calcul des correspondances. Le point dur de ce cas de figure est de déterminer la durée limite du temps de suspension. Cette durée dépend en particulier de :

- la qualité des images : plus les images sont bruitées, plus il est important de résoudre les ambiguïtés au plus tôt.
- l'uniformité du mouvement : plus le mouvement est uniforme, plus on peut se permettre d'attendre pour résoudre les ambiguïtés.
- la vitesse des objets mobiles : plus leur vitesse est grande, plus il est important de résoudre les ambiguïtés au plus tôt.

La durée limite du temps de suspension est alors calculée en combinant l'ensemble de ces critères. Lorsque qu'une cible composée atteint cette durée limite, on force la résolution de l'ambiguïté associée. On calcule alors une matrice d'ambiguïté dont les colonnes représentent les cibles ambiguës et dont les lignes représentent les cibles temporaires. S'il y a effectivement correspondance, on associe alors les cibles ambiguës aux cibles temporaires. Sinon, on détache les cibles temporaires qui perdent ainsi toute possibilité de récupérer leur historique. On élimine ensuite la cible composée. Cette technique pouvant conduire à des pertes d'une partie des pistes est peu utilisée (c.-à-d. on autorise généralement une durée limite importante du temps de suspension).

Les cibles temporaires s'éparpillent

Quand une cible temporaire s'écarte de sa cible composée, on essaie d'abord de trouver une cible ambiguë qui appartient à la même cible composée et qui lui correspond. Si cette recherche est un succès, on établit la correspondance. Si c'est un échec, cas de la figure 4.8, la cible temporaire devient une cible visible ayant perdu son historique et on la détache de la cible composée.

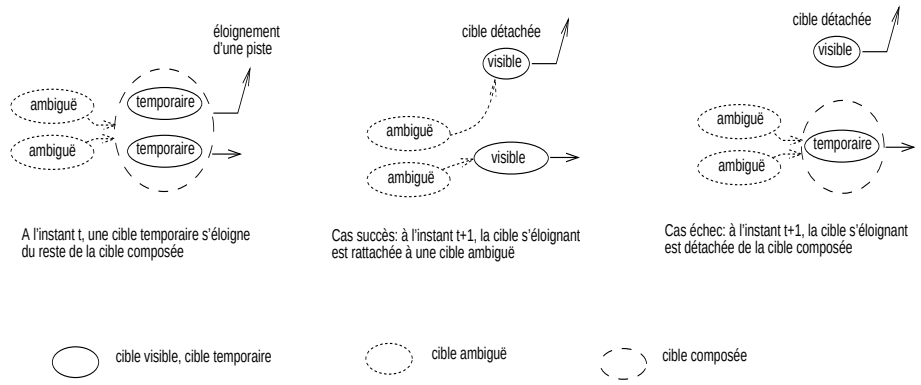

FIG. 4.8 – *Une cible temporaire s'écarte de sa cible composée sans qu'on puisse lui faire correspondre une cible ambiguë.*

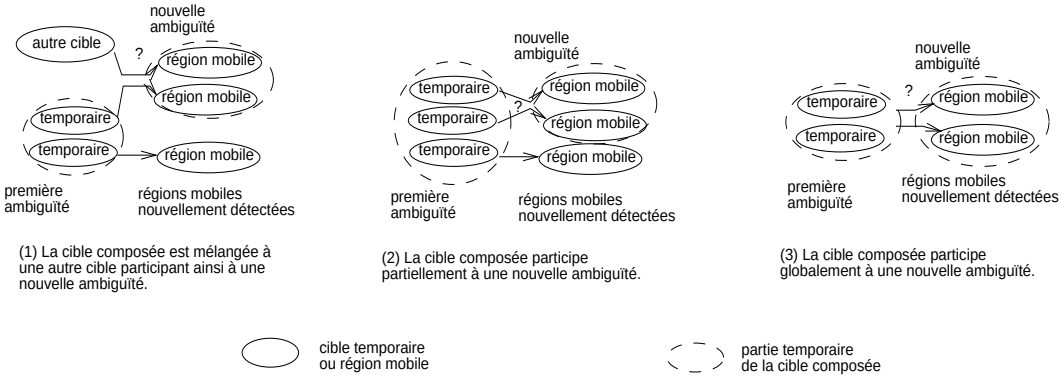

FIG. 4.9 – *Les trois cas de confrontation à une seconde ambiguïté*

La cible composée se mélange à une seconde ambiguïté

Une cible composée peut participer à une seconde ambiguïté de trois manières différentes. Comme le montre la figure 4.9, la cible composée peut (1) soit être mélangée à d'autres cibles, (2) soit être mélangée seule et partiellement ou (3) soit être mélangée seule et globalement, selon qu'une partie ou que la totalité des cibles temporaires sont mêlées à la deuxième ambiguïté. On résout les cas de mélange à d'autres cibles et du mélange partiel, en considérant que les cibles temporaires, mêlées à la deuxième ambiguïté se détachent de la première cible composée. On est alors ramené au cas précédent, où les cibles temporaires s'écartent de leur cible composée. Ces cibles temporaires sont alors considérées comme de simples cibles visibles n'ayant

plus de connexions avec la première ambiguïté.

Cas du mélange global d'une cible composée dans une seconde ambiguïté

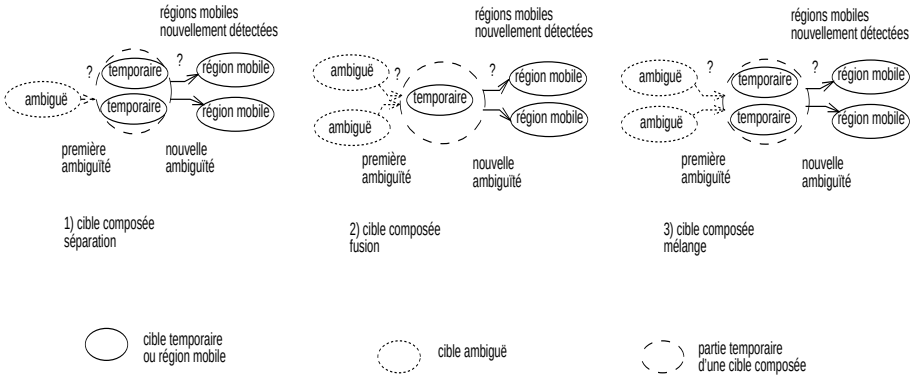

FIG. 4.10 – *Les trois sous-cas de mélange global à une seconde ambiguïté*

Le cas du mélange global d'une cible composée est le plus intéressant pour le processus de suivi, car cette configuration est la plus courante et permet de mieux récupérer les cibles ambiguës de la cible composée. On résout cette configuration en construisant une seconde cible composée dont la partie temporaire est formée des régions mobiles nouvellement détectées. Le problème réside dans la détermination des cibles ambiguës de cette seconde cible composée. Pour cela, on transforme les cibles ambiguës de la première ambiguïté en des cibles ambiguës de la seconde ambiguïté. Ceci revient à prolonger la trajectoire de ces cibles ambiguës de la première ambiguïté, pendant l'intervalle de temps compris entre l'apparition des deux ambiguïtés. On prolonge les trajectoires des cibles ambiguës en utilisant les trajectoires des cibles temporaires de la première ambiguïté. Ce prolongement est plus ou moins précis selon le type de la première cible composée. La figure 4.10 illustre les trois sous-cas possibles :

- Dans le cas d'une cible composée séparation, on duplique la cible ambiguë autant de fois qu'il y a de cibles temporaires. Puis, on prolonge chaque copie de la cible ambiguë avec les trajectoires des cibles temporaires. Ces copies prolongées de la cible ambiguë deviennent les cibles ambiguës de la seconde cible composée.

– Dans le cas d'une cible composée fusion, on prolonge toutes les cibles ambiguës avec l'unique cible temporaire. Ces cibles prolongées deviennent les cibles ambiguës de la seconde cible composée.

– Dans le cas d'une cible composée mélange, on calcule la trajectoire moyenne des cibles temporaires. Cette moyenne est une approximation de la trajectoire des cibles ambiguës entre les deux instants d'apparition des ambiguïtés. Cette approximation est d'autant plus précise que les cibles temporaires sont proches les unes des autres. Puis on prolonge toutes les cibles ambiguës avec cette approximation de la trajectoire. Ces cibles prolongées deviennent alors les cibles ambiguës de la seconde cible composée.

Dans ces trois cas, les cibles ambiguës de la première cible composée sont récupérées en tant que cibles ambiguës de la seconde cible composée. Une fois que la cible composée associée à la seconde ambiguïté est ainsi construite, on élimine la première cible composée.

Cas particulier de mélange global d'une cible composée dans une seconde ambiguïté

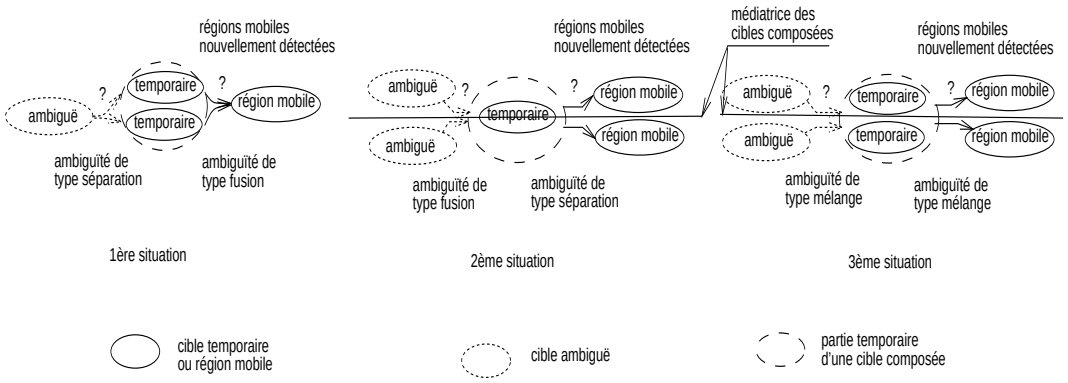

FIG. 4.11 – *Trois situations particulières où les cibles ambiguës conservent leur disposition au sein de la cible composée.*

Dans chaque cas précédent de mélange global conduisant à une seconde ambiguïté, il existe des situations particulières où la deuxième ambiguïté permet de résoudre la première ambiguïté. Ces situations sont importantes, car elles permettent de résoudre de nombreuses occultations dynamiques. Elle surviennent lorsqu'il y a conservation de la disposition des cibles ambiguës

au sein de la cible composée. Comme le montre la figure 4.11, ces situations peuvent se produire lorsque :

- (1) une ambiguïté de type séparation se remélange en une ambiguïté de type fusion,
- (2) une ambiguïté de type fusion se remélange en une ambiguïté de type séparation,
- (3) une ambiguïté de type mélange se remélange en une seconde ambiguïté de type mélange.

Pour qu'il y ait conservation de la disposition des cibles ambiguës quand les situations 2 et 3 surviennent, il faut également que l'une des deux conditions suivantes soit satisfaite :

- si les cibles ambiguës ont des **trajectoires parallèles** de même sens, on considère que leur ordre au sein des cibles composées fusion ou mélange se conserve. Cette condition est à rapprocher du principe *« de rigidité »* des objets mobiles (Zhang, 1993). On ordonne alors les cibles ambiguës le long de la médiatrice de la cible composée, en projetant leur barycentre sur cette droite. Cette médiatrice est définie par le barycentre des cibles temporaires et par la direction moyenne des cibles ambiguës. Dans le cas où des cibles ambiguës auraient le même ordre le long de cette médiatrice, on ordonne ces cibles ambiguës le long de la normale à cette médiatrice. Puis on ordonne de même les régions mobiles nouvellement détectées. On vérifie alors que les cibles ambiguës correspondent bien aux régions mobiles nouvellement détectées ayant le même ordre, en utilisant le critère de similarité des caractéristiques. Enfin si le critère est vérifié, on associe les cibles aux régions mobiles.
- si les cibles ambiguës ont des **trajectoires toutes sécantes**, on considère que ces trajectoires seront maintenues entre les instants d'apparition des deux ambiguïtés. On calcule alors une matrice d'ambiguïté dont les colonnes représentent les cibles ambiguës et dont les lignes représentent les régions mobiles nouvellement détectées. Si il y a effectivement correspondance, on associe alors les cibles ambiguës aux régions mobiles leur correspondant.

De façon générale, pour qu'il y ait conservation de la disposition des cibles ambiguës, il est également nécessaire que le laps de temps, pendant lequel la disposition de ces cibles a été suspendue (c.-à-d. laps de temps

séparant les deux ambiguïtés), soit suffisamment court. La durée maximale de ce laps de temps est un des paramètres du système qui appartient au contexte de l'application, car il dépend du type de la scène. Par exemple, on acceptera que les cibles représentant un groupe d'individus conservent leur ordre pendant un temps plus long dans un couloir de métro (lieu où les trajectoires sont assez parallèles), que pour une scène se déroulant sur un quai de métro (lieu où les trajectoires sont peu parallèles).

Cette section décrit les différentes manières de résoudre les ambiguïtés survenant pendant l'étape d'association des cibles déjà existantes avec les régions mobiles nouvellement détectées. Ce qui caractérise ainsi la méthode de suivi proposée dans ce rapport, c'est le traitement de tous ces cas d'ambiguïté dans un unique cadre, celui des cibles composées.

4.5 Comparaison à d'autres méthodes de suivi

On ne compare la méthode proposée des cibles composées uniquement qu'aux méthodes de suivi d'objets mobiles sans modèle, car seules ces méthodes proposent des techniques de résolution d'ambiguïté. Les autres méthodes de suivi essayent plutôt d'éviter l'apparition de ces situations d'ambiguïté. On compare ainsi la méthode des cibles composées aux trois méthodes suivantes : JPDAF, MHT et la recherche par faisceaux.

Lorsque qu'une cible correspond à plusieurs régions mobiles nouvellement détectées, **la méthode du JPDAF** (Bar-Shalom and Fortmann, 1988), calcule un filtre combinant l'influence de toutes les régions mobiles correspondant à la cible. Ce filtre permet à l'étape suivante de prédire la nouvelle position de la cible en tenant compte de l'ambiguïté passée. Comme dans la méthode des cibles composées, la JPDAF utilise un filtre qui combine les informations de plusieurs régions mobiles. Cependant, ce filtre ne permet pas de temporiser la décision d'association et les correspondances sont établies de suite, même si ces associations restent ambiguës. De plus, le filtre utilisé pour combiner plusieurs mesures est de même nature que les filtres associés aux cibles, correspondant à une seule région mobile. Ce point pose un problème de cohérence au niveau de la notion de filtre associé à une cible. Cette méthode n'est donc pas adaptée à ce suivi de régions mobiles, ne permettant pas de suspendre la prise de décision d'association.

La **méthode des MHT** (Cox and Hingorani, 1996), comme la méthode proposée des cibles composées, maintient l'ambiguïté pour la résoudre ul-

térieurement. Cependant la méthode des cibles composées se singularise de cette méthode sur la manière de résoudre l'ambiguïté. La méthode des MHT conserve toutes les informations (c.-à-d. les filtres) relatives aux correspondances ambiguës, et attend de recevoir des informations fiables pour prendre une décision. Pour éviter l'explosion combinatoire cette méthode, d'une part ne conserve que les filtres les plus cohérents, et d'autre part fixe un seuil du passé des filtres conservés pour le calcul des correspondances. Par contre, la méthode des cibles composées conserve les informations de façon linéaire en fonction du nombre d'ambiguïté, car les informations sur le passé sont condensées à chaque étape sous la forme de trajectoires et de cibles composées. Elle permet ainsi de pouvoir résoudre une ambiguïté ayant eu lieu dans le lointain passé.

La **méthode de recherche par faisceaux** (Zhang, 1993), conserve également les informations sur les ambiguïtés pour les résoudre ultérieurement. Elle duplique physiquement toutes les cibles ambiguës et concurrentes, puis poursuit le suivi de ces cibles jusqu'à ce que leur piste devienne incohérente (et sont alors éliminées) ou se terminent normalement. La méthode des cibles composées procède de manière inverse. Premièrement, elle ne duplique des informations ambiguës que si ces informations appartiennent à des cibles distinctes et réellement suivies. Elle globalise et conserve uniquement les informations du passé nécessaires à la résolution future de l'ambiguïté. Deuxièmement, elle gèle la prise de décision évitant de suivre des cibles qui se révéleront incohérentes par la suite. Par contre, la méthode de recherche par faisceaux suit des cibles concurrentes, sans différencier les cibles issues d'une duplication des cibles effectivement suivies. Ce problème de cohérence impose au module de reconnaissance des scénarios de maintenir à jour les états de plusieurs mondes concurrents.

Néanmoins, les méthodes des MHT et de recherche par faisceaux permettent de geler plusieurs ambiguïtés successives. Quand une seconde ambiguïté est rencontrée, la méthode des cibles composées ne résout pas toujours la première ambiguïté avec succès. De plus, comme elle approxime les informations du passé, elle peut effectuer une résolution imprécise conduisant plus tard à des erreurs de suivi. Dans la section 4.7 suivante, on propose des améliorations à la méthode des cibles composées, prenant en compte ces problèmes.

4.6 Résultats

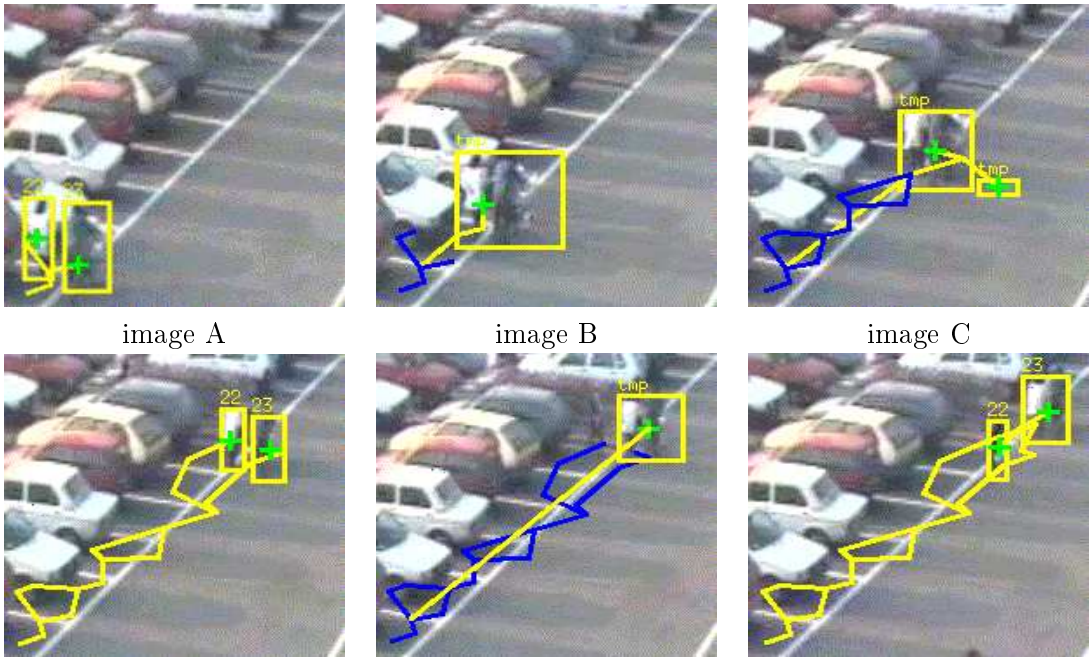

FIG. 4.12 – *Cette séquence d'images montre deux individus marchant côte à côte. Selon leur disposition, ils sont suivis individuellement ou en tant que groupe à l'aide de cibles composées. Sur la figure, les trajectoires des cibles visibles et des cibles temporaires sont dessinées en gris clair tandis que les trajectoires suspendues des cibles ambiguës sont dessinées en noir.*

Dans le cadre de cette thèse, on a développé un module de suivi écrit en C++, implantant l'algorithme de suivi décrit dans ce chapitre. Dans l'état actuel du système, le module de suivi utilise des informations contextuelles, mais il n'est pas encore directement connecté avec la base de contexte construite par le logiciel d'acquisition du contexte (voir chapitre 2). Ces connexions prévues à courts termes, permettraient par exemple à l'aide d'une calibration de la scène, d'utiliser les propriétés 3D des objets mobiles au lieu de leurs propriétés 2D. Le module de suivi a été testé sur des séquences d'images de métro et de parking, prises dans le cadre du projet européen Esprit HPCN PASSWORDS (Bogaert et al., 1996). Sur la figure 4.12, on peut

voir deux individus marchant côte à côte, le long d'une rangée de voitures :

- Sur l'image A, le groupe d'individus se sépare. Le processus de suivi utilise alors une cible composée séparation et suit individuellement les deux individus, dont les cibles correspondantes sont numérotées 22 et 23.

- Sur l'image B, les deux individus se regroupent. Le processus de suivi utilise alors une cible composée fusion et suit globalement le groupe d'individus. La trajectoire du groupe est en gris clair car elle correspond à une cible temporaire, tandis que les trajectoires des individus sont en noir, car elles correspondent à des cibles ambiguës. Effectivement, sur l'image B, le processus de suivi a suspendu le suivi des individus pour suivre temporairement le groupe.

- Sur l'image C, après une succession de fusion et de séparation, les deux individus sont en groupe. Ce groupe se mélange avec du bruit (une ombre portée au sol, non éliminée par le module de traitement d'images). Le processus de suivi suit temporairement les deux pistes mais n'établit pas les correspondances, car le critère de similarité des caractéristiques n'est pas vérifié. Il utilise pour cela une cible composée mélange.

- Sur l'image D, les deux individus se sont de nouveaux séparés. Le processus de suivi est effectivement capable d'établir les correspondances avec les cibles initiales (numérotées 22 et 23), car les critères de similarité des caractéristiques et de distance spatiale (entre les positions prédites des anciennes cibles et les régions mobiles nouvellement détectées) sont vérifiés. Pour établir ces correspondances, on tire également profit de la situation particulière de conservation de la disposition des cibles ambiguës. Sur cette séquence, les directions des individus sont parallèles et permettent de supposer que l'ordre des individus est conservé au sein de la cible composée (c.-à-d. l'individu numéroté 22 reste à gauche par rapport à la direction moyenne du mouvement et l'individu numéroté 23 reste à droite). Ceci permet de fournir un indice supplémentaire pour établir les correspondances.

- Sur l'image E, les deux individus sont de nouveaux mélangés dans un groupe. Le processus de suivi utilise alors une cible composée fusion. On peut remarquer la présence, à gauche du groupe, d'un troisième individu sortant de la rangée de voitures. Cet individu n'est pas détecté, en raison de son manque de contraste avec l'image du fond de la scène.

– Sur l'image F, les deux individus sont toujours en groupe. Mais le troisième individu passant juste à côté du groupe vient d'être détecté. Le critère de similarité des caractéristiques est moyennement vérifié pour la cible numéroté 23 (elle représente le groupe d'individus). Le critère de distance spatiale est également moyennement vérifié pour la cible numéroté 22 (elle représente le troisième individu), et ce changement de direction est autorisé (c.-à-d. retour en arrière). Par la combinaison fortuite de ces raisons, le processus de suivi établit les mauvaises correspondances et mélange le troisième individu au groupe initial d'individus.

Cette séquence d'images montre l'utilisation des cibles composées pour résoudre différentes ambiguïtés de suivi. Elles permettent, dans une majorité de cas, de récupérer les cibles ambiguës. Les cibles composées peuvent encore être mieux exploitées en suivant les objets mobiles en même temps, en tant qu'individu et en tant que groupe. Par exemple, sur cette séquence d'images, ceci permettrait de suivre le groupe d'individus comme une cible à part entière et non pas comme une cible temporaire.

De plus, sur de longues séquences d'images, il est nécessaire d'être plus robuste. Pour cela, on peut utiliser des informations supplémentaires fournies par la base de contexte et par le module de reconnaissance de scénarios. Par exemple, pour l'image F, la base de contexte peut indiquer la possibilité d'apparition d'un troisième individu, dans les zones bordant la rangée de voitures. Le module de reconnaissance de scénarios peut également indiquer la non conformité d'une trajectoire réalisée par un individu rebroussant chemin. Ces méthodes permettraient ainsi de récupérer certaines erreurs de détection des régions mobiles.

4.7 Améliorations de l'étape de résolution des ambiguïtés

Le problème d'association des cibles existantes avec les régions mobiles nouvellement détectées reste un problème difficile, dû essentiellement aux erreurs de détection, obligeant souvent à prendre des décisions sans avoir toutes les informations nécessaires. Prises trop tôt, ces décisions peuvent conduire à des erreurs de suivi, bloquant ainsi tout le système d'interprétation. Pour améliorer la résolution du problème d'association des cibles, deux axes de recherches peuvent être considérés : (1) conserver plus d'informations sur le suivi des cibles et (2) retarder la prise de décision.

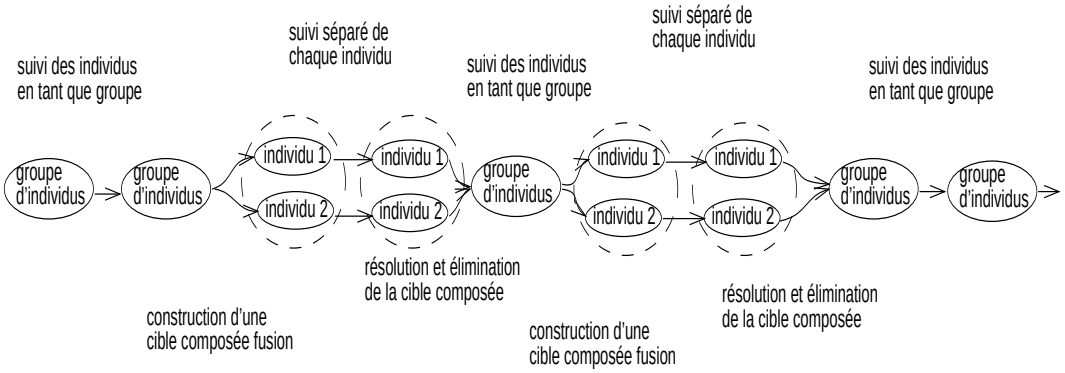

FIG. 4.13 – *Deux individus marchant côte à côte, donnent naissance à une succession de fusion et de séparation de leur piste.*

Pour conserver plus d'informations sur le suivi des cibles, on propose de garder les cibles composées ayant déjà permis de résoudre une ambiguïté. Cette idée repose sur la constatation que dans de nombreux cas, la cible composée correspond à un objet mobile. Ces cas conduisent à une succession de motifs constitués (1) de la construction d'une cible composée puis (2) de la résolution et de l'élimination de la cible composée. Par exemple, comme le montre la figure 4.13, deux individus marchant côte à côte ont tendance à se mélanger, puis à se séparer, puis à se mélanger de nouveau... Conserver la cible composée permet alors d'effectuer parallèlement le suivi de deux manières différentes, selon que les cibles soient mélangées (on utilise alors la cible composée) ou séparées (on utilise alors les cibles temporaires). Cette méthode de double suivi des cibles permet d'éliminer certains cas d'ambiguïté et d'avoir un suivi plus précis. Il permet également au module de reconnaissance des scénarios de maintenir en même temps, l'analyse du comportement des objets mobiles pouvant correspondre soit à la cible composée, soit aux cibles temporaires.

Pour retarder la prise de décision, on propose de définir des cibles composées formées à partir d'autres cibles composées. La rencontre d'une seconde ambiguïté n'impose plus alors la résolution de la première ambiguïté. À la place, comme le montre la figure 4.14, on construit une cible composée dont l'une (ou plusieurs) de ses cibles ambiguës est la cible composée associée à la première ambiguïté. Cette méthode permet de généraliser le gel des résolutions d'ambiguïté et peut permettre d'éviter certaines approximations du suivi. Ces deux améliorations n'ont pas encore été mises en œuvre.

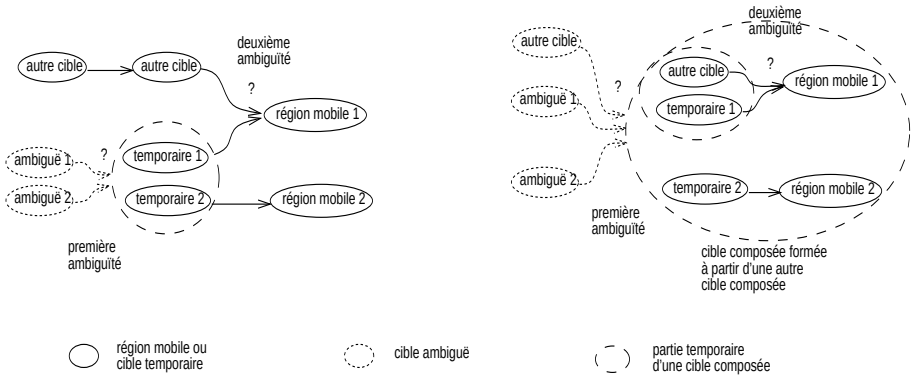

FIG. 4.14 – *On construit une cible composée formée à partir d'une précédente cible composée, pour éviter de résoudre de suite la première ambiguïté.*

4.8 Conclusion

Ces travaux ont permis de proposer une **méthode de suivi** utilisant les régions détectées par le module de détection présenté au chapitre 3. Cette méthode possède quatre caractéristiques principales :

- Elle utilise des **modèles dynamiques** et rudimentaires sur les objets mobiles, plutôt que des connaissances *a priori*. Ces modèles dynamiques sont d'une part un modèle minimal de la perception de l'objet mobile comprenant la position et les dimensions des régions mobiles associées et d'autre part, un modèle du mouvement de l'objet mobile (supposé linéaire par morceaux) constitué de sa trajectoire et permettant de calculer sa direction et sa vitesse. Cette absence de connaissances *a priori* permet de suivre aussi bien les objets rigides (p. ex. les véhicules) que des objets non rigides (p. ex. les êtres humains). Pour cette raison nous considérons que cette méthode de suivi est générique.

- Elle prend en compte les **erreurs de détection**, telles que des reflets et des ombres. Pour cela, la méthode proposée suit la perception du mouvement (c.-à-d. les régions mobiles) plutôt que les objets réels de la scène. L'identification des régions suivies en tant que bruit ou en tant qu'objets réels est confiée au module de reconnaissance de scénarios.

- Elle gère les **problèmes spécifiques au suivi**, tels que les occultations statiques et dynamiques, conduisant à des ambiguïtés dans le processus de suivi. Pour résoudre ces ambiguïtés, nous avons développé

la méthode des *cibles composées*. Une cible composée modélise une ambiguïté en permettant de suivre globalement tout un ensemble de cibles ambiguës. Elle permet alors de geler le suivi de ces cibles et d'attendre des informations supplémentaires pour réaliser ultérieurement les correspondances.

La méthode de suivi proposée adapte ainsi la résolution des ambiguïtés aux différentes situations rencontrées. Bien que nous ayons dû traiter de nombreuses situations, beaucoup d'entre elles correspondent à des situations rares et pathologiques du processus de suivi. Les situations les plus utiles sont les situations particulières de conservation de la disposition des cibles, où la rencontre d'une seconde ambiguïté permet de résoudre une première ambiguïté. Ces situations sont les plus courantes, correspondant généralement à des situations d'occultation dynamique. Dans ces situations là, la méthode de suivi proposée est capable de récupérer la correspondance des cibles ambiguës dans de bonnes conditions. Ces situations d'occultation dynamique sont un problème délicat, que tout processus de suivi doit aborder. Par exemple, dans (Toal and Buxton, 1992), les auteurs proposent d'utiliser une relation spéciale pour gérer les problèmes d'occultation. Cette relation permet de conserver l'objet mobile occulté, d'attendre son émergence, puis s'il y a lieu de restaurer son historique (c.-à-d. sa trajectoire). Ce qui caractérise la méthode proposée dans ce mémoire, c'est le traitement de tous les cas d'ambiguïté dans un unique cadre, celui des cibles composées.

- Elle utilise des **informations complémentaires** afin d'améliorer les performances du suivi et de tirer profit des conditions particulières d'une application. Ces informations comprennent :

  - Information contextuelles de la scène : Ces informations telles que les zones de terminaison de pistes, permettent d'améliorer la robustesse du suivi.
  - Résultats de l'analyse des comportements des objets mobiles : Ces informations permettent de déterminer si les régions suivies associées aux objets mobiles correspondent à du bruit ou à des objets réels de la scène. Elles permettent également de renforcer le traitement des associations ambiguës.
  - Propriétés plus précises sur les objets mobiles. Elles comprennent en particulier les propriétés 3D des régions mobiles et un modèle

dynamique précis des objets mobiles (contenant par exemple leur contour, leur histogramme de couleur et les arêtes détectées dans les régions mobiles associées). Bien que ces propriétés ne soient pas actuellement prises en compte par la méthode de suivi proposée, nous avons montré tout au long de ce chapitre comment ces propriétés pouvaient être utilisées.

Ces travaux ont conduit également à la réalisation et à l'intégration au système d'interprétation d'un **module de suivi** de régions mobiles (écrit en langage C++). Plusieurs tests de validation ont permis de déterminer les limites de la méthode proposée. Ces limites dépendent en particulier de la qualité de l'acquisition des images et au type de scènes traitées.

Parmi les **travaux prospectifs**, nous envisageons trois axes de recherche. Le premier axe consiste à mettre en œuvre les améliorations proposées dans ce chapitre à la section 4.7, afin d'obtenir un module de suivi robuste.

Un second axe consiste à prendre en compte les évolutions possibles du module de détection pouvant être relié à plusieurs capteurs, tels que différents angles de prise de vue de la scène. Il s'agit alors de pouvoir utiliser et fusionner ces différentes sources d'informations. Le cadre générique de la méthode de suivi devrait faciliter l'utilisation de ces diverses informations.

Enfin, un troisième axe de recherche est d'adapter la méthode de suivi au type de l'objet suivi. Dans certaines applications, de nombreux objets mobiles sont rigides et possèdent un modèle *a priori* de leur forme et de leur mouvement améliorant les performances de la méthode de suivi. Nous comptons alors combiner une méthode de suivi n'utilisant pas de modèles *a priori* afin de suivre la majorité des objets mobiles (cette méthode se doit d'être peu coûteuse en temps et de pouvoir suivre tout type d'objet), avec une méthode de suivi tirant profit du modèle *a priori* pour suivre certains objets mobiles identifiés comme rigides et importants (cette méthode se doit d'être robuste et précise).

Chapitre 5

Passage du numérique au symbolique

Ce chapitre traite le problème de la transformation des propriétés numériques, calculées à partir de traitements d'images, en propriétés symboliques, utiles à la reconnaissance d'actions, première étape de la reconnaissance de scénarios. Ces propriétés symboliques correspondent soit à la perception d'actions, telles que *« s'asseoir »*, soit à la perception d'absence d'actions, telle que *« attendre »*. Les modules de détection du mouvement et de suivi calculent des données numériques sur les régions mobiles, appelées mesures. Ces mesures renseignent sur la perception des actions réalisées par les objets mobiles associés à ces régions mobiles. Les actions n'étant pas instantanées, il n'est possible au début d'une action que d'émettre des hypothèses sur sa reconnaissance à l'aide de diagnostics, puis d'attendre des informations supplémentaires pour confirmer les hypothèses. De plus, les mesures sur les régions mobiles sont souvent imprécises et incomplètes, rendant nécessaire la prise en compte de l'incertitude de ces hypothèses. Pour ces raisons, la reconnaissance d'actions repose sur des raisonnements de diagnostic abductif et de gestion de l'incertitude.

La reconnaissance des actions représentées par des propriétés symboliques et perçues à partir de mesures, est un problème essentiel du processus global d'interprétation de séquences d'images. Le module de reconnaissance de scénarios a besoin de propriétés symboliques suffisamment sûres et ayant été valides pendant un intervalle de temps suffisamment long pour mener à bien son traitement. Sans ce passage du numérique au symbolique, ce module est dans la plupart des cas, incapable de donner une interprétation de la scène. Dans ce chapitre, la section 5.1 présente un état de l'art sur la perception

d'actions à partir de traitements d'images et de raisonnements abductifs. Dans la section 5.2, on définit et justifie un modèle de raisonnement à base de diagnostics abductifs. On décrit ensuite une implantation de ce modèle dans la section 5.3, et un exemple d'utilisation de cette étape de diagnostic dans la section 5.4.

5.1 État de l'art

Cette section présente deux grandes familles de méthodes dédiées à la perception d'actions. En ce qui concerne la première famille, le mécanisme de base consiste à reconnaître une action en calculant directement, à l'aide de traitements numériques, des mesures précises et détaillées sur le mouvement des objets mobiles et en les comparant à des modèles *a priori* d'action. Ces méthodes ne peuvent s'appliquer que sur des cas particuliers d'actions, dépendant par exemple de l'angle de vue de la caméra. On présente ensuite des méthodes permettant de calculer l'imprécision des mesures pour améliorer le processus global de reconnaissance d'actions. La deuxième famille de méthodes utilise le raisonnement abductif. On présente alors différents formalismes servant de cadre à l'abduction et à sa mise à jour. On finit cette section en décrivant plusieurs systèmes ayant mis en œuvre ces formalismes.

5.1.1 Reconnaissance d'actions à partir de traitements d'images

De nombreux travaux ont été réalisés dans la communauté vision par ordinateur afin de reconnaître les mouvements du corps humain. Ces travaux se caractérisent par l'utilisation de traitements sophistiqués permettant de calculer les propriétés du mouvement et de les comparer à un modèle. Il existe cinq types de méthode :

- Un premier type de méthode consiste à rechercher les étapes caractéristiques du mouvement. Par exemple dans (Kuniyoshi and Inoue, 1993), les auteurs cherchent à reconnaître le mouvement d'une main saisissant un cube et le déplaçant. Pour cela, ils utilisent des traitements d'images spécialisés dans la détection de propriétés relatives à des étapes caractéristiques du mouvement à reconnaître. Par exemple, pour déterminer si la main a saisi le cube, ils calculent la distance et la possibilité de contact entre la main et le cube. L'exécution d'un traitement spécifique, nécessite au préalable de déterminer dans quelle situation se trouve la main, puis de planifier la reconnaissance de la tâche supposée se produire.

- Un autre type de méthode consiste à calculer des propriétés physiques, telles que la périodicité d'un phénomène, relatives à l'évolution dans le temps de paramètres de mouvements cycliques d'un individu. Ces méthodes utilisent des modèles *a priori* de l'évolution des paramètres du mouvement, qui ne sont calculables que si le mouvement est cyclique. De plus, elles requièrent souvent que le mouvement soit plan et que la caméra soit en face de ce plan du mouvement, pour faciliter la modélisation du mouvement. Un des problèmes consiste à repérer le début du mouvement par rapport au cycle. Les paramètres du mouvement sont en général obtenus à partir de l'étude des caractéristiques géométriques de l'individu et des différentes parties le constituant (p. ex. l'orientation d'une jambe par rapport au tronc). Par exemple dans (Rohr, 1994), l'auteur utilise un modèle 3D d'être humain composé de cylindres, pour calculer les paramètres du mouvement d'un piéton, projeté dans le plan du mouvement. Ces paramètres sont rassemblés dans un seul vecteur dont l'évolution dans le temps permet de repérer par rapport au modèle d'évolution du mouvement, l'étape courante du piéton. Dans (Campbell and Bobick, 1995), les auteurs utilisent la méthode de l'affichage de points lumineux en mouvement (*« Moving Light Displays »* (MLD) en anglais), permettant de suivre des marqueurs 3D lumineux sur l'individu examiné. Ils calculent ainsi plusieurs paramètres du mouvement, et arrivent à reconnaître neuf types d'enchaînement de pas de danse en ballet classique, comme *« plié »* ou *« relevé »*. Cette méthode étant très précise, elle permet de discriminer des mouvements similaires. Dans (Polana and Nelson, 1994), en s'inspirant de techniques d'analyse de texture, les auteurs placent une grille constituée de cellules sur l'individu en mouvement et étudient l'évolution dans le temps de l'intensité lumineuse de chaque cellule. Ils arrivent ainsi à discriminer plusieurs types de mouvement : marcher, courir, nager, skier, osciller, sauter. L'avantage de cette dernière méthode est qu'elle peut être utilisée avec une mauvaise qualité des images, comme une faible résolution.

- Un autre type de méthode consiste à reconnaître un mouvement à l'aide d'un modèle *a priori* de contour déformable. Dans (Baumberg and Hogg, 1995), les auteurs utilisent un modèle correspondant à l'allure d'un individu en train de marcher, pour chaque angle de prise de vue, pour chaque type de démarche. Un des intérêts de cette méthode, est la possibilité d'apprentissage des modèles *a priori* de contour déformable.

- Un autre type de méthode, proposé par (Bobick and Davis, 1996), uti-

lise des modèles *a priori* d'action. Les auteurs accumulent pendant le temps de durée d'une action, les points de l'individu en mouvement projetés sur le plan image, qu'ils comparent ensuite aux modèles d'action. Ces modèles dépendent de l'angle de prise de vue et ne permettent pas de discriminer des actions ayant les mêmes projections, comme *« s'asseoir »* et *« s'accroupir »*. Un autre inconvénient de cette méthode est la nécessité de déterminer le début de l'action, pour commencer à accumuler les points de mouvement.

– Enfin, un dernier type de méthode à la mode depuis peu, consiste à utiliser les Modèles de Markov Cachés (HMM *« Hidden Markov Models »*, en anglais). Ces modèles permettent de reconnaître l'enchaînement d'actions dans le temps, en tenant compte du bruit du vecteur de mesures servant à la reconnaissance. Ces mesures caractérisent les actions à reconnaître et sont par exemple, les coordonnées angulaires du bras d'un individu. Les HMM sont des automates d'états finis, dont les états représentent les actions. À chaque état est associée une distribution de probabilités, permettant de décider si le vecteur de mesures courant est caractéristique de l'état en question. De même, à chaque transition est associée une distribution de probabilités, indiquant quand l'automate doit changer d'état courant. Par ces moyens, il est possible de reconnaître des actions, même si le vecteur de mesures est bruité. Par exemple dans (Starner and Pentland, 1995), les auteurs utilisent les HMM afin de reconnaître près de 40 mots du langage des signes à partir d'une caméra. De même dans (Brand et al., 1997), les auteurs reconnaissent d'autres actions complexes, comme celles d'un ouvrier réparant un appareil électronique. Cependant, un inconvénient majeur des HMM réside dans la phase de développement des automates. Effectivement, les distributions de probabilités sont spécifiques à une action et nécessite une période d'apprentissage. De plus les HMM permettent difficilement de prendre en compte l'amplitude de mouvements, tels qu'un geste plus ou moins grand. Pour résoudre ce problème, les auteurs dans (Wilson and Bobick, 1997) utilisent des HMM paramétrés et ainsi peuvent reconnaître et quantifier le geste d'un individu, désignant un objet avec son doigt.

Ces méthodes permettent de reconnaître des mouvements et des actions, mais uniquement dans certaines conditions. Ces mouvements ou actions sont cycliques, souvent caricaturaux, et dépendent de l'angle de prise de vue de la caméra. De plus, leur reconnaissance nécessite généralement une bonne

qualité des images et les séquences d'images correspondantes sont alors prises en laboratoire.

5.1.2 Imprécision

Pour améliorer la reconnaissance des actions accomplies par les objets mobiles, un principe consiste à prendre en compte l'imprécision des mesures utilisées pour cette reconnaissance. Plusieurs types de méthodes ont ainsi été développés :

- Un premier type de méthodes utilise le contexte de la scène (se référer au chapitre 2), comprenant en particulier les informations sur les capteurs et le système d'acquisition des images. Par exemple dans (Miura and Shirai, 1993), les auteurs calculent l'imprécision d'une donnée selon sa profondeur dans la scène. Plus une donnée concerne un objet mobile se situant au fond de la scène, plus cette donnée est considérée comme imprécise.

- Un autre type de méthodes utilise des informations relatives à la reconnaissance des objets mobiles. Ce type de méthodes est principalement utilisé dans le cas d'objets rigides, et consiste à quantifier la correspondance entre le modèle *a priori* d'objet mobile et les mesures caractérisant l'objet mobile courant.

- Un autre type de méthodes utilise les informations relatives au suivi des objets mobiles et au calcul des paramètres de leur mouvement. Une méthode classique consiste à estimer l'imprécision des mesures réalisées, à partir de la qualité de la mise en correspondance entre la prédiction des positions des objets mobiles et de leurs positions réelles mesurées. Une autre méthode, proposée dans (Kollnig et al., 1994), consiste à calculer plusieurs paramètres du mouvement des objets mobiles et à classer les valeurs de ces paramètres à l'aide de la théorie sur les ensembles flous. Par exemple, ces auteurs qualifient une vitesse comme étant nulle, petite, normale, rapide ou très rapide.

L'imprécision des mesures calculées sur les régions mobiles permet de pondérer leur influence sur la reconnaissance d'actions.

5.1.3 Formalismes servant de cadre à l'abduction

Les méthodes de reconnaissance d'actions décrites précédemment supposent de manière *a priori*, des conditions particulières d'utilisation, per-

mettant de guider la reconnaissance sans avoir à gérer son incertitude. Les méthodes suivantes ne présupposent pas de conditions particulières, mais utilisent des hypothèses à l'aide de raisonnement abductif et mettent à jour la validité de ces hypothèses. Cette section présente ainsi différents formalismes servant de cadre à l'abduction et à sa mise à jour.

Abduction

L'abduction est généralement définie comme étant un raisonnement de l'effet vers les causes (Dubois and Prade, 1992b) :

$$\frac{\begin{array}{l}\text{si } h \text{ alors } e \\ e \text{ est observé}\end{array}}{h \text{ peut être vraie}}$$

Avec h, une hypothèse sur l'existence d'une cause et e, un effet observé ou mesuré. Il existe principalement trois approches permettant un raisonnement abductif :

- Premièrement, dans l'**approche relationnelle**, on suppose connues les relations entre toutes les hypothèses et tous les effets possibles. Pour un ensemble d'effets observés $\mathcal{E}_{\mathcal{OBS}} = \{ e_j \}$, il s'agit alors de déterminer l'ensemble des hypothèses $\mathcal{H}_{\mathcal{SOL}} = \{ h_i \}$, en relation avec ces effets. Dans le cadre de la théorie sur les ensembles flous, ce problème revient à calculer la fonction d'appartenance de h_i à $\mathcal{H}_{\mathcal{SOL}}$. Par exemple, la formule suivante permet de calculer cette fonction d'appartenance (Dubois and Prade, 1992a) :

$$\mu_{\mathcal{H}_{\mathcal{SOL}}}(h_i) = max_j[\, min(\mu_{\mathcal{R}}(h_i, e_j),\, \mu_{\mathcal{E}_{\mathcal{OBS}}}(e_j))\,]$$

  Avec $\mu_{\mathcal{R}}(h_i, e_j)$, une quantification de la relation floue *« si l'hypothèse h_i est vraie, alors l'effet e_j peut être observé »*, et $\mu_{\mathcal{E}_{\mathcal{OBS}}}(e_j)$, le degré d'appartenance de e_j à $\mathcal{E}_{\mathcal{OBS}}$.

- Deuxièmement, dans l'**approche logique**, une explication abductive d'une formule **e**, conjonction d'effets, dans le cadre d'une théorie $\mathcal{T}$, consiste à déterminer une conjonction d'hypothèses **h** qui soit valide et qui soit telle que :

$$\mathcal{T}, h \models e$$

Ce modèle de raisonnement est théorique et nécessite d'être étendu pour aborder des problèmes du monde réel. Dans (Piechowiak et al., 1994), les auteurs ont étendu ce modèle en ajoutant à chaque formule, une variable représentant l'intervalle de temps pendant lequel la formule est valide. Le raisonnement abductif prend ainsi en compte des contraintes temporelles. Dans le cadre de la logique possibiliste, l'adaptation au monde réel revient à calculer la nécessité que **h** soit une explication de **e**. Par exemple, cette nécessité peut se calculer à travers la formule suivante (Dubois and Prade, 1992b) :

$$max_j[\, min(\mathcal{N}(h_i \rightarrow e_j), \mathcal{N}(e_j))\,] \tag{5.1}$$

Avec $\mathcal{N}(h_i \rightarrow e_j)$ la nécessité que l'hypothèse h_i soit la cause de l'effet e_j, et $\mathcal{N}(e_j)$ la nécessité que e_j soit observé. Comme le font remarquer les auteurs proposant cette formule, l'approche relationnelle peut être considérée comme équivalente à l'approche logique au point de vue applicatif (c.-à-d. une même méthode de calcul des hypothèses solutions). L'approche logique s'adapte bien à la modélisation de problèmes traitant du monde réel. Par exemple, dans (Cayrac et al., 1995), les auteurs diagnostiquent les causes possibles de pannes dans un système constitué de plusieurs composants électroniques. À l'aide de la logique possibiliste (formule 5.1) et d'un modèle représentant le réseau des composants, ils génèrent, en fonction d'un symptôme observé, une hypothèse de panne relative à l'état des composants.

– Troisièmement, dans l'**approche conditionnelle**, on calcule la probabilité que l'hypothèse **h** soit vraie, sachant que l'effet **e** est observé. Cette probabilité conditionnelle se calcule à l'aide de la formule de Bayes :

$$\begin{aligned} P(h \mid e) &= \frac{P(e \mid h).P(h)}{P(e)} \\ &= \frac{1}{1 + \frac{P(e|\neg h)}{P(e|h)}.\frac{P(\neg h)}{P(h)}} \end{aligned}$$

En logique possibiliste, on se ramène à ce calcul en considérant les mesures possibilistes conditionnelles (Dubois and Prade, 1992b) :

$$\Pi(h \mid e) = min(1, \frac{\Pi(e \mid h)}{\Pi(e \mid \neg h)}.\frac{\Pi(h)}{\Pi(\neg h)})$$

Le formalisme le plus communément utilisé pour mettre en œuvre ces approches est le cadre probabiliste, car il offre un cadre strict et rigoureux qui a été éprouvé depuis longtemps. Ce formalisme est décrit en détail dans le livre de (Pearl, 1988). Cependant, dans ce rapport de thèse ces trois approches ont plutôt été abordées dans le cadre de la théorie des ensembles flous et de la logique possibiliste, car ce cadre offre de plus grandes possibilités d'expression de problèmes du monde réel. Néanmoins, d'autres formalismes peuvent également être utilisés pour mener à bien des raisonnements abductifs. Par exemple, les fonctions de croyance de Dempster-Shafer, les probabilités qualitatives et les probabilités symboliques (Pacholczyk and Pacholczyk, 1996) offrent un cadre intermédiaire entre numérique et symbolique. D'autres formalismes, comme les plausibilités (Friedman and Halpern, 1996), proposent une théorie permettant de rassembler les probabilités et la logique possibiliste. Cependant, bien que différents dans leurs propriétés secondaires, tous ces formalismes ne s'opposent pas (Friedman and Halpern, 1995), (Dubois and Prade, 1993) et sont théoriquement équivalents. Ainsi l'utilisation d'un formalisme plutôt qu'un autre ne s'impose pas pour mettre en œuvre des raisonnements abductifs dans les applications d'interprétation de séquences d'images.

Gestion de l'incertitude

Les approches présentées précédemment décrivent des façons de sélectionner des hypothèses permettant d'expliquer un ensemble d'effets. Il reste à présenter les méthodes permettant de gérer ces hypothèses, lorsque les effets observés changent au cours du temps. Une étude complète des raisonnements gérant l'incertitude se trouve dans (Pearl, 1988). Cette section présente les deux principales approches.

- La première approche consiste à utiliser les **réseaux bayesiens** (Pearl, 1988), (Charniak, 1991). Ces réseaux sont des graphes directs acycliques, dont les nœuds sont des variables aléatoires représentant des propositions et les arcs sont des matrices de probabilités conditionnelles représentant les dépendances entre les variables. Ces réseaux sont principalement utilisés pour mettre à jour l'incertitude des propositions. La mise à jour consiste à constater que le monde a changé et à

modifier le degré d'incertitude des propositions précédemment calculé pour tenir compte de ces changements. La mise à jour du réseau est bi-directionnelle. La formule des probabilités conditionnelles permet de déduire la probabilité des effets à partir de celles des causes (raisonnement déductif), tandis que la formule de Bayes permet de déduire la probabilité des causes à partir de celles des effets (raisonnement abductif). Pour utiliser un réseau bayesien, il est nécessaire de fournir initialement les probabilités *a priori*, affectées aux racines du graphe, ainsi que les matrices de probabilités conditionnelles. L'initialisation du réseau est d'autant plus dure qu'il est souvent difficile de donner un sens aux probabilités *a priori* et aux matrices de probabilités conditionnelles, par rapport au monde réel. Une alternative consiste à utiliser la théorie de Dempster-Shafer (Pearl, 1988), qui modélise l'ignorance et n'oblige pas à fournir les probabilités de toutes les propositions.

– La seconde approche consiste à utiliser les **réseaux d'hypothèses** mis sous forme de systèmes de maintien de la vérité TMS (*« Truth Maintenance Systems »* en anglais) et en particulier des ATMS (*« Assumption-based Truth Maintenance Systems »* en anglais) (Pearl, 1988), (Charpillet et al., 1992). Ces réseaux sont des graphes dont les nœuds représentent des propositions ou des hypothèses et les arcs des implications logiques de dépendance, du type :

SI propositions (prémisses) *ET* hypothèses *ALORS* propositions (conclusions)

Les propositions peuvent être vraies ou fausses, tandis que les hypothèses sont supposées *a priori* comme étant justes. À chaque nœud, on attache un ensemble d'environnements (un environnement correspondant à une combinaison d'hypothèses), tels que si l'un d'entre eux est vérifié alors la proposition associée au nœud est également vérifiée. Ces réseaux sont utilisés afin de maintenir la cohérence des hypothèses. Par exemple, lorsqu'une nouvelle hypothèse est ajoutée, on calcule les nouveaux environnements cohérents. Sachant qu'une proposition-conclusion est vraie, on déduit alors (abduction) les environnements valides. Les ATMS permettent donc de réaliser des raisonnements abductifs, mais le cadre logique modélise mal la complexité du monde réel. Pour cela, diverses extensions de la logique classique sont proposées. Par exemple, dans (Dubois et al., 1990), les auteurs présentent un ATMS possibiliste.

5.1.4 Réseaux probabilistes et analyse de séquences d'images

Dans (Ghallab et al., 1992), (Grandjean, 1991), les auteurs utilisent un réseau probabiliste pour reconnaître des objets réels dans une scène en faisant correspondre les primitives de traitements d'images détectées dans la scène (p. ex. des segments et faces 3D) aux modèles des objets réels. Ils commencent par générer une hypothèse *a priori* de l'objet à reconnaître. Dans un second temps, ils calculent de nouvelles primitives ou propriétés sur les objets à reconnaître afin d'infirmer ou de confirmer l'hypothèse émise. Ils estiment ensuite les similitudes et ambiguïtés dans l'appariement entre les primitives nouvellement mesurées et les primitives attendues correspondant à l'hypothèse émise sur l'objet. En combinant ces similitudes et ambiguïtés sur l'objet à l'aide de la formule de Bayes, ils calculent la validité de l'hypothèse. Ils construisent ainsi de façon incrémentale, une hypothèse sur identification des objets de la scène.

Dans (Djian et al., 1996), les auteurs pilotent un robot mobile qui reconnaît de façon incrémentale les objets statiques contenus dans une pièce, à l'aide d'un réseau bayesien. Ils possèdent un ensemble d'hypothèses *a priori* correspondant aux différentes identifications possibles de ces objets. À chaque image, le système calcule des primitives de traitements d'images, comme des segments de droites, relatives aux objets à reconnaître. Puis ils déterminent par une phase de propagation des probabilités si ces mesures nouvellement calculées renforcent ou diminuent la probabilité de validité des hypothèses. Le système ainsi réalisé permet de reconnaître les portes et fenêtres de la pièce.

Dans (Gong and Buxton, 1993), (Buxton and Gong, 1995), les auteurs utilisent des réseaux bayesiens dynamiques pour reconnaître un scénario de trafic routier. Ils ont construit un système comprenant deux composants. Le premier composant rassemble les opérateurs de base, que sont par exemple la proximité, la vitesse, le changement d'orientation sur les objets mobiles suivis. À partir des valeurs calculées par ces opérateurs de base, le deuxième composant génère dynamiquement un réseau bayesien correspondant aux comportements probables des objets mobiles suivis, comme *« doubler »*, *« suivre »*, ou *« comportement inconnu »*. Ces réseaux sont initialisés par les valeurs des opérateurs de base et à chaque étape, les auteurs calculent la probabilité que les hypothèses de comportement soient valides connaissant les valeurs nouvellement calculées des opérateurs de base. Si une hypothèse de comportement est jugée inintéressante, le réseau bayesien associé est alors éliminé. Le système obtenu permet alors de reconnaître en temps réel des scénarios tels que deux véhicules se doublant autour d'un rond-point.

5.1.5 Réseaux d'hypothèses et traitement du signal

Dans (Charpillet et al., 1992), les auteurs utilisent un ATMS pour détecter les défauts de pièces mécaniques dans les centrales nucléaires. Ils analysent les signaux électriques parcourant ces pièces mécaniques, puis déterminent un ensemble de défauts possibles en se basant sur le contexte et l'historique de la pièce mécanique.

Dans (Dekneuvel et al., 1992), (Ghallab et al., 1992), les auteurs utilisent un RMS à base de logique possibiliste (*« Reason Maintenance System »*, en anglais), extension des TMS, pour maintenir la cohérence dans un réseau d'hypothèses. À partir de la reconnaissance d'objets réels (p. ex. *« l'objet est un meuble »*) et d'événements (p. ex. *« l'objet est en mouvement »*), ils génèrent des propositions hypothétiques, appelées croyances, telles que p. ex. *« l'objet est poussé »*, qu'ils essaient par la suite d'infirmer ou de confirmer. Pour cela, ils utilisent des informations contextuelles ou des informations supplémentaires provenant d'un module de reconnaissance des objets et des événements. Les croyances permettent de guider l'interprétation de la scène. Dès qu'une croyance est jugée incohérente, l'interprétation associée est abandonnée et le système essaie de générer une nouvelle croyance. Le robot mobile piloté par ce système arrive ainsi à reconnaître le laboratoire dans lequel il évolue.

5.2 Modèle proposé

Cette section explique le modèle du processus proposé pour réaliser le passage des données numériques calculées sur les régions mobiles, aux propriétés symboliques caractérisant les objets mobiles associés aux régions.

5.2.1 Objets mobiles

Régions et objets mobiles

La première étape consiste à déterminer les régions mobiles ou les groupes de régions mobiles correspondant à des objets mobiles. Étant donné le peu de connaissances *a priori* sur les objets mobiles, nous supposons que toute région et que tout groupe de régions mobiles proches les unes des autres, peuvent constituer des objets mobiles. Dans un second temps, nous calculons les propriétés relatives à ces objets. Si ces propriétés ont un degré de vraisemblance élevé (défini ci-après), nous considérons alors qu'elles sont effectivement caractéristiques d'objets mobiles et nous poursuivons normalement l'analyse du comportement de ces objets. Si ce n'est pas le cas, ces

propriétés ont un faible degré de vraisemblance et nous ne tenons alors pas compte des analyses relatives aux objets correspondants.

Dans certains cas, une région mobile peut avoir son comportement analysé en tant qu'objet mobile à part entière et, en même temps, en tant qu'élément d'un groupe de régions mobiles correspondant à un objet ou à un groupe d'objets. Comme l'illustre la figure 5.1, supposons que la tête d'un individu soit la seule partie détectée d'un individu (dont la détection est représentée par la région A), et qu'elle soit à proximité d'un second individu (dont la détection est représentée par la région B). Trois analyses sont alors possibles : la région mobile correspondant à la détection de la tête peut soit être considérée (1) comme un objet mobile à part entière, soit être associée au second individu, cette association étant considérée correspondre (2) à un objet mobile à part entière ou (3) à une foule. Dans cet exemple seule la seconde hypothèse n'est pas valide, puisque elle suppose que l'association des détections des deux individus correspond à un seul individu. Dans ces cas-là, le degré de vraisemblance permet de décider quels scénarios sont valides.

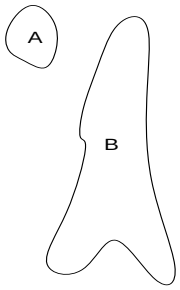

FIG. 5.1 – *La région A correspond à la détection d'un premier individu et la région B correspond à celle d'un second individu.*

Les mesures caractérisant les régions mobiles sont calculées par le module de suivi (décrit dans le chapitre 4). Pour traiter les applications cibles (se référer à la section 1.1), nous utilisons un modèle de région mobile constitué de dix mesures : les cinq points génériques comprenant la position courante, la hauteur, la largeur, la vitesse instantanée, la direction du mouvement et la trajectoire. Ce modèle est représenté figure 5.2.

Définition des propriétés des objets mobiles

Les propriétés symboliques des objets mobiles sont calculées récursivement à partir de sous-propriétés. Au niveau zéro, les propriétés correspondent

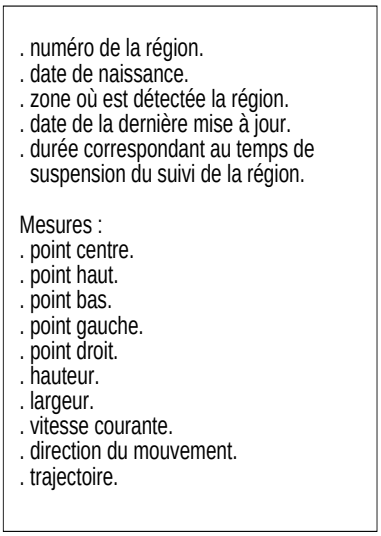

FIG. 5.2 – *Modèle d'une région mobile.*

aux mesures effectuées sur les régions mobiles constituant les objets. Cette définition implique une organisation des propriétés en réseau, comme le montre la figure 5.3.

Nous définissons deux types de propriétés selon que nous nous intéressons à la **valeur** de son calcul, comme *« la hauteur de l'objet mobile »*, ou à son **évolution**, comme *« la hauteur diminue »* ou *« la hauteur reste stable »*. Cette possibilité de qualifier l'évolution des propriétés est essentielle pour reconnaître des activités humaines, car elle est un des moyens les plus utilisés pour décrire ces activités en langage naturel. Par exemple, cette possibilité permet d'exprimer la négation d'une action, appelée parfois non-action, comme *« l'individu ne bouge pas »*. La définition récursive des propriétés permet de décomposer les problèmes relatifs à leur calcul en plusieurs niveaux. Par exemple, une propriété de niveau supérieur (p. ex. *« la hauteur diminue »*) permet d'analyser une évolution d'une sous-propriété (p. ex. *« la hauteur de l'objet mobile »*).

Modèle de propriété

Pour le calcul de ces propriétés, on a défini un modèle de propriété comprenant huit éléments, comme indiqué sur la figure 5.4 : le nom de la propriété, son type (*valeur* ou *évolution*), les objets mobiles impliqués, les sous-propriétés la constituant, la valeur de la propriété, un ensemble de méthodes

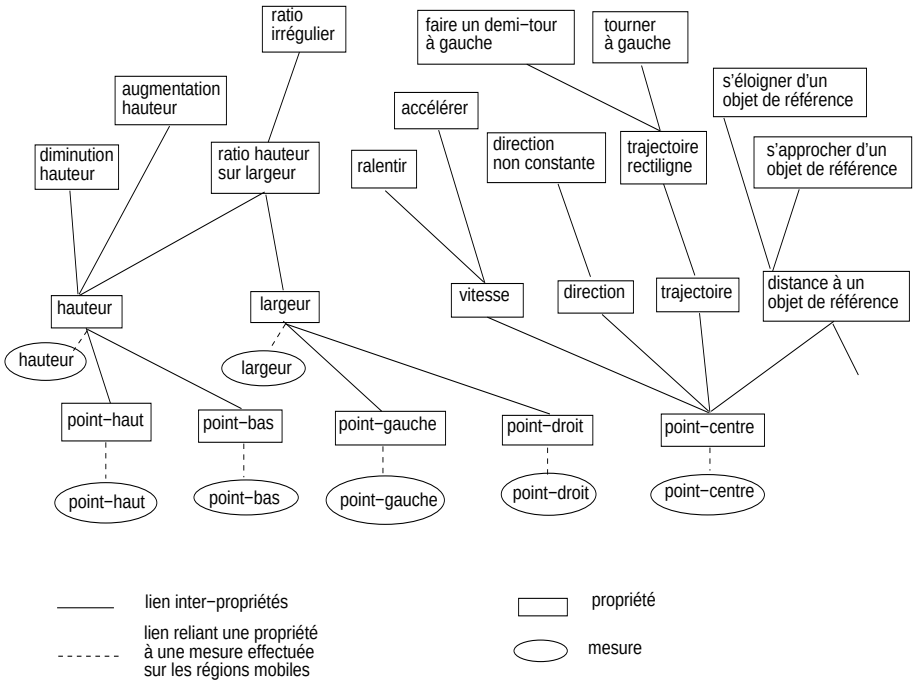

FIG. 5.3 – *Le réseau de propriétés utilisé pour traiter les applications cibles.*

pour calculer cette valeur, son degré de vraisemblance et un ensemble de méthodes pour calculer ce degré de vraisemblance.

Valeur d'une propriété

La valeur de la propriété qualifie la valeur intrinsèque de la propriété ou l'évolution d'une autre propriété. Les méthodes pour calculer cette valeur sont principalement des opérations de filtrage de données aberrantes et de calcul de valeur moyenne. Le choix d'une méthode dépend du type de la propriété. Les propriétés sont généralement calculées sur un court intervalle de temps (typiquement 1 seconde dans les applications abordées), prenant en compte 4 à 5 images. Cet intervalle permet de calculer une valeur moyenne de la propriété et d'éviter que sa valeur soit trop instable. De plus, si une donnée numérique nouvellement calculée est très différente des données précédentes, alors cette grandeur est considérée comme aberrante et n'est pas prise en compte. On opère ainsi un filtre sur les valeurs numériques calculées par les modules de détection et de suivi de régions mobiles. Ces propriétés correspondent donc à l'analyse continue des actions sur un court intervalle

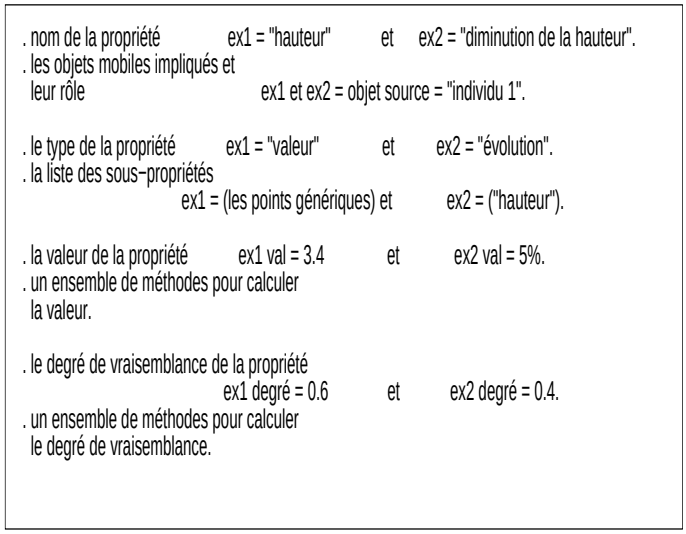

FIG. 5.4 – *Le modèle de propriété avec un exemple de propriété de type* valeur *« hauteur », et un exemple de type* évolution *« diminution de la hauteur ». La valeur de la propriété 2 est le pourcentage de diminution de la propriété 1 par rapport à sa valeur maximale.*

de temps.

Les propriétés les plus pertinentes concernent la trajectoire si la caméra est loin des objets mobiles et concernent l'évolution de la taille, si la caméra est proche des objets. Par exemple, dans plusieurs applications de surveillance de parkings où les individus en mouvement sont loin de la caméra, le seul calcul de leur trajectoire permet bien souvent de reconnaître complètement leurs actions. En ce qui concerne les applications cibles choisies (se référer au chapitre 1), les actions à reconnaître sont simples, comme *« se baisser »*, et ne nécessitent pas des fonctions complexes de filtrage. Pour d'autres types d'application, comme l'analyse de pas de danse dans un ballet (Campbell and Bobick, 1995), il est nécessaire de développer des fonctions calculant par exemple les propriétés cycliques des mesures effectuées sur les régions mobiles. De même, les HMM (Modèles de Markov Cachés) sont des fonctions performantes (Starner and Pentland, 1995), qui permettent de calculer la valeur d'une propriété, prenant en compte sa courte évolution dans le temps.

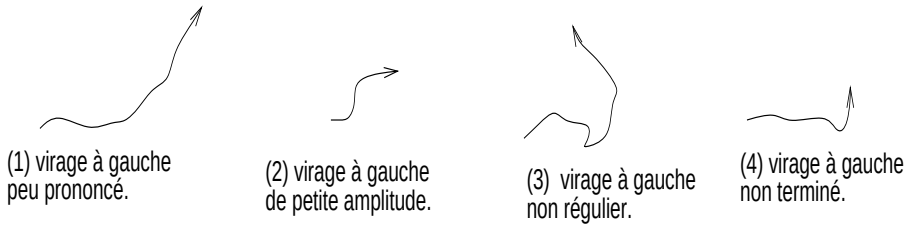

FIG. 5.5 – *Ces quatre trajectoires peuvent toutes correspondre à la trajectoire d'un individu tournant à gauche.*

Généricité des propriétés

Les propriétés des objets mobiles sont définies dans l'objectif d'être génériques (indépendantes de l'application) et d'être utilisées comme briques de base, afin de permettre la reconnaissance de scénarios plus complexes. Malheureusement dans certains cas, il est difficile de trouver de véritables invariants caractérisant une propriété. Dans ces cas là, la valeur de calcul de la propriété, ainsi que les méthodes permettant son calcul, dépendent du contexte de la scène. Par exemple, la figure 5.5 montre différentes trajectoires pouvant toutes correspondre à la propriété *« l'individu tourne à gauche »*. Cette propriété dépend du contexte de la scène : la trajectoire de l'individu dépend de la place disponible pour tourner, ainsi que de la détermination de l'individu qui peut hésiter avant d'agir, ou décider brusquement de tourner. Dans ces cas là, la solution que nous avons adoptée est de définir différentes propriétés possédant des méthodes proches de calcul de la valeur. Les scénarios plus complexes peuvent alors être décrits à l'aide d'une ou de l'ensemble de ces propriétés, selon le type de l'application envisagé.

Approche basée sur l'apparence

L'approche que nous proposons dans cette thèse est basée sur l'apparence des actions (appelée *« appearance-based approach »* en anglais), contrairement aux approches basées sur les lois physiques régissant les interactions entre les objets mobiles et aux approches à base de règles (appelées respectivement *« physics-based »* et *« rule-based approaches »* en anglais). Effectivement, l'approche basée sur l'apparence calcule les propriétés des actions basées sur la perception des régions mobiles, sans essayer de déterminer les raisons des interactions des objets mobiles avec leur environnement (c.-à-d. les moteurs des actions). Dans les applications envisagées, les modèles des objets mobiles et de leur interactions sont difficilement accessibles. L'interprétation

d'actions complexes, faisant généralement référence au domaine d'application, est réalisée au niveau le plus abstrait de la reconnaissance des scénarios. Au niveau des scénarios, c'est l'utilisation des propriétés symboliques qui rend possible un raisonnement abstrait et permet alors d'émettre des hypothèses quant aux intentions des objets mobiles.

Dans d'autres applications, telles que l'analyse des gestes d'un ouvrier saisissant un objet, l'utilisation de lois, comme la physique newtonienne (Mann et al., 1996), et de règles, comme le fait de savoir que la main a un mouvement moteur (Brand et al., 1993), s'avère utile pour interpréter les mesures effectuées. Ces lois ou règles permettent alors de reconnaître des événements dit causaux, comme *« saisir un tournevis »*, pour en déduire directement les raisons des interactions de l'ouvrier avec son environnement, comprenant en particulier des outils et une machine à réparer. Ces approches ne sont utilisables que pour des applications dont le domaine est très spécifique.

5.2.2 Imprécision

Dans le système d'interprétation présenté dans cette thèse, il n'y a pas d'estimation de l'imprécision des mesures utilisées pour la reconnaissance des actions accomplies par les objets mobiles. Ces mesures étant principalement calculées par le module de détection des régions mobiles, le module de reconnaissance d'actions peut alors difficilement estimer leur imprécision. De plus, ces mesures étant particulièrement imprécises (mauvaises conditions de détection des régions mobiles), leur imprécision généralement trop importante est peu utilisable.

5.2.3 Abduction et degré de vraisemblance

Les sous-sections précédentes ont défini le modèle des propriétés et expliqué les méthodes de calcul de leurs valeurs. Dans cette sous-section nous proposons une approche permettant de calculer l'incertitude (c.-à-d. le degré de vraisemblance) de ces valeurs. Cette approche repose sur une phase de diagnostic abductif. Nous commençons alors par montrer l'utilité du raisonnement abductif. Nous décrivons ensuite comment ce raisonnement permet de calculer le degré de vraisemblance d'une propriété. Enfin nous terminons cette sous-section en montrant que l'originalité du processus d'interprétation de séquences d'images réside dans la nécessité d'utiliser un raisonnement abductif.

Nécessité du raisonnement abductif

Les propriétés symboliques relatives à un objet mobile correspondent aux causes (p. ex. *« l'individu se baisse »*) pouvant expliquer les mesures (ou leurs évolutions) effectuées sur les régions mobiles représentant la perception du mouvement de l'objet (p. ex. *« la hauteur de la région diminue »*). L'interprétation de séquences d'images est alors avant tout un raisonnement abductif, les propriétés symboliques représentant les hypothèses (c.-à-d. les causes) réalisées sur les mesures observées (c.-à-d. les effets). Ces hypothèses sont nécessaires, principalement pour deux raisons :

- La nature temporelle des mesures à effectuer : effectivement dans de nombreuses situations, les raisons de l'évolution d'une propriété sont difficiles à déterminer dès le début de l'évolution. De plus, les images étant traitées à la volée, il est nécessaire de traiter tout début d'évolution, car le retour sur les informations passées n'est pas réaliste. Il est alors nécessaire de générer des hypothèses sur toute évolution et d'attendre les mesures suivantes pour déterminer complètement leur signification. Le raisonnement abductif permet alors de calculer ces hypothèses, ainsi que leur degré de confiance.

- La nature continue des mouvements des objets mobiles et les mauvaises conditions de détection des régions mobiles : il est souvent impossible de déterminer une valeur binaire (vrai ou faux) pour une propriété calculée à un instant et à un endroit donné. De même, une action est souvent difficilement décomposable en une séquence de mouvements, entrecoupée de points de repère précis. Il est souvent nécessaire de suivre et d'analyser le mouvement d'un objet mobile sur un certain intervalle de temps avant de pouvoir déterminer la valeur de la propriété. Cette continuité du mouvement est d'autant plus vraie que les objets sont non rigides. Par exemple, les mouvements d'un individu sont moins précis, moins bien décomposés que les mouvements d'un véhicule. La compréhension du comportement d'un individu nécessite ainsi l'analyse de ses mouvements sur un certain intervalle de temps non réductible à un instant. Tout le problème réside alors dans l'obtention de propriétés aux valeurs discrètes et sûres, utilisables à un plus haut niveau d'abstraction, par le module de reconnaissance de scénarios. Ce problème est abordé à la section 6.2.2. Ces propriétés plus fiables sont les hypothèses obtenues par la phase d'abduction.

Le raisonnement abductif permet alors de calculer la vraisemblance d'une propriété à partir des mesures effectuées sur les régions mobiles. Néanmoins,

ce raisonnement peut être optionnel dans certains cas, en particulier lorsqu'il n'existe pas de doute sur les mesures, c.-à-d. dans le calcul de la propriété. Il s'agit alors d'avoir une représentation suffisamment souple pour ne pas imposer ce raisonnement quand il n'est pas nécessaire. Bien souvent, on s'aperçoit de la nécessité du raisonnement abductif lorsqu'on rencontre sur des exemples, un problème spécifique à la propriété traitée. Pour déterminer son utilité, nous employons alors une méthode expérimentale d'essais et d'erreurs. Un axe de recherche consiste à synthétiser cette expertise pour automatiser le développement de la phase de raisonnement abductif.

Degré de vraisemblance d'une propriété

Le raisonnement abductif permet également de diagnostiquer des phénomènes extérieurs susceptibles d'interférer dans le calcul des propriétés symboliques. Par exemple, lorsqu'on s'intéresse à l'évolution de la taille d'un individu, il s'agit de déterminer si la diminution de cette taille est due à une occultation de la tête ou si elle est effectivement due à un mouvement de l'individu, comme *« se baisser »*. Nous appelons alors **désordres** (p. ex. *« la tête de l'individu est occultée »*) les raisons pouvant interférer dans l'évolution des mesures et **symptômes** (p. ex. *« le point le plus haut de l'individu est mal suivi »*), les indices renseignant sur l'occurrence des désordres. La phase de diagnostic consiste alors à calculer les symptômes et à en déduire les désordres par abduction.

Le degré de vraisemblance est calculé à partir du résultat du diagnostic. Ce degré quantifie l'incertitude de la propriété et indique si la valeur de la propriété est une donnée fiable. Si ce degré est suffisamment élevé, alors la propriété peut être utilisée. Les calculs de la phase de diagnostic et du degré de vraisemblance sont décrits dans la section 5.3 ci-dessous.

Cette séparation du calcul de la valeur et de la vraisemblance d'une propriété est caractéristique du système présenté dans cette thèse et n'est pas présente dans la plupart des systèmes de reconnaissance d'actions. Par exemple, les HMM calculent la valeur des propriétés en même temps que leur probabilité d'existence. Ils ne permettent pas de différencier, par exemple, qu'une propriété est faiblement reconnue mais très probable, d'une propriété qui est fortement reconnue mais peu probable.

Reconnaissance d'actions et reconnaissance d'objets statiques

Le fait de comprendre en quoi la reconnaissance d'actions est différente de celle d'objets statiques (c.-à-d. réalisée sur une seule image), permet de mieux

appréhender la nature de ces deux raisonnements. Par exemple, la perception active est clairement différente de la reconnaissance d'objets statiques, car elle autorise l'acquisition planifiée de nouvelles informations. Par contre la reconnaissance d'actions en différé (c.-à-d. à partir de la séquence complète d'images décrivant les actions), s'apparente de par sa nature statique à la reconnaissance d'objets statiques.

La différence entre une reconnaissance statique et dynamique est essentiellement due à trois causes :

- Dans le cas d'une reconnaissance dynamique, toutes les données utiles ne sont pas disponibles au début de la reconnaissance. Une phase de raisonnement abductif est alors nécessaire pour anticiper les données manquantes.

- La reconnaissance statique n'a pas à gérer les problèmes de continuité dans le temps des propriétés caractérisant les objets.

- La détection d'informations pertinentes est facilitée par le mouvement, puisque le mouvement permet d'isoler des objets mobiles du fond statique de la scène. Comme le montre la section 6.3, l'abstraction de ces informations pertinentes et de leurs évolutions facilite l'utilisation de raisonnements symboliques, tels que les logiques. Dans le cas d'une reconnaissance dynamique, il est alors plus facile de mener des raisonnements abstraits.

Cependant, les raisons de la distinction entre les deux processus de reconnaissance font encore l'objet de débats dans la communauté scientifique. En ce qui concerne nos travaux, la nécessité d'un raisonnement abductif dans la reconnaissance d'actions semble une raison essentielle, différenciant le cas dynamique du cas statique. Comme expliqué précédemment, cette nécessité est due à la variation continue et parfois irrégulière dans le temps, des propriétés. Pour cette raison et comme l'indique (Oppenheim, 1992), l'interprétation d'une séquence d'images se restreint difficilement à un problème de classification, ce raisonnement ne remplissant pas les mêmes fonctionnalités que le processus d'abduction. Bien sûr, ce résultat n'implique pas que le raisonnement abductif ne soit pas utile dans certains cas de reconnaissance d'objets statiques, ni même que la classification ne soit pas utile dans des problèmes de reconnaissance d'actions.

L'utilisation du raisonnement abductif conduit à la génération de symboles, correspondant aux propriétés utilisées pour décrire des actions. La manipulation de ces symboles rend alors possible le traitement en langage

naturel des séquences d'images, permettant ainsi de décrire la scène. La reconnaissance d'actions apparaît comme un domaine conduisant à terme au raisonnement symbolique. Un des objectifs de cette thèse est de montrer que le raisonnement symbolique est possible et souhaitable pour interpréter des séquences d'images.

5.2.4 Mise à jour de l'incertitude

Comme le montre la figure 5.6, les propriétés sont reliées les unes aux autres par des liens de dépendance. Cette section étudie comment l'incertitude d'une propriété se propage aux propriétés dépendantes de la propriété donnée. Il existe trois types de liens de dépendance inter propriétés, liés aux trois phases d'enchaînement des étapes du processus d'interprétation. La première phase abstrait les données numériques décrivant les images en données symboliques correspondant à des caractéristiques du comportement des objets mobiles. Elle est appelée *phase d'abstraction* ou phase ascendante et est dirigée par les données. La deuxième phase correspond au contrôle de l'abstraction et permet de diriger l'interprétation en fonction de comportements particuliers. Elle est appelée *phase de contrôle* ou phase descendante et est dirigée par les buts. Cette phase est surtout utilisée pour sélectionner l'attention du système sur des tâches particulières à effectuer. La troisième phase, appelée *phase d'évolution temporelle*, est orthogonale aux deux précédentes. Elle correspond à l'évolution dans le temps des objets mobiles et de leurs propriétés au cours du séquencement des images.

L'existence de ces trois phases induit trois modes de gestion de l'incertitude. Tout d'abord, le mode 1 gère l'incertitude de la phase d'évolution temporelle, qui est particulièrement incertaine. Pour chaque nouvelle image, le système calcule les propriétés relatives aux objets mobiles, vérifie si elles restent cohérentes et si besoin, diagnostique les raisons de leur incohérence. Le fait qu'une propriété soit vraie est ainsi considéré comme une hypothèse et on attribue à la propriété en question un degré de vraisemblance, quantifiant ainsi son incertitude. Deuxièmement (mode 2), la phase d'abstraction propage l'imprécision des données et leur incertitude vers les propriétés les plus abstraites. Enfin (mode 3), la phase de contrôle corrige les incertitudes des propriétés les plus élémentaires à partir du retour des informations. Les phases ascendante et descendante définissent ainsi un réseau de liens de dépendance (les liens d'abstraction et de contrôle) entre les propriétés. Par exemple, la propriété *« le point le plus haut »* d'un individu dépend en premier lieu des points les plus hauts des régions mobiles le constituant (phase ascendante) et dépend également par effet retour de la propriété, *« la taille*

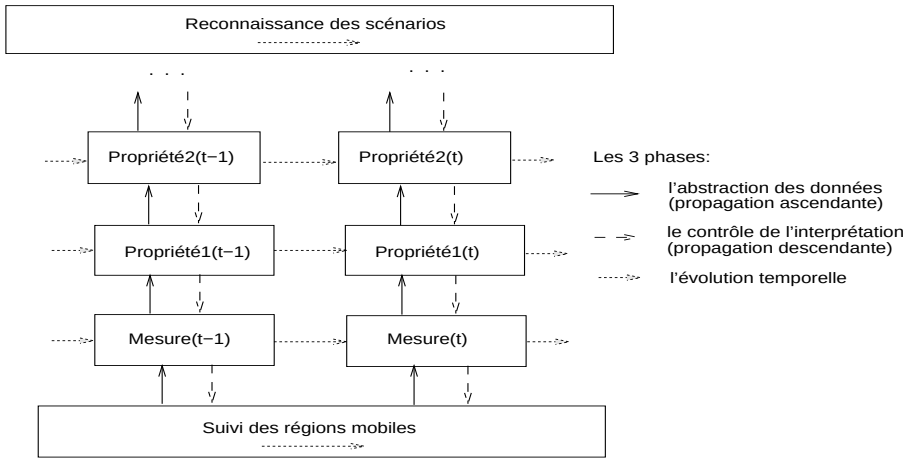

FIG. 5.6 – *Les liens de dépendance inter propriétés correspondent à la phase d'évolution temporelle, suivie des phases de propagation ascendante et de propagation descendante.*

de l'individu » (phase descendante).

La gestion de l'incertitude correspond à la combinaison de ces trois phases, créant ou propageant l'incertitude dans tout le système.

5.3 La phase de diagnostic

Le problème du passage des données numériques en propriétés symboliques réside ainsi dans la phase de diagnostic. Cette section décrit le cadre formel pour mener à bien cette étape.

5.3.1 Choix d'un formalisme

Comme le montre l'état de l'art du chapitre 5.1, le formalisme probabiliste dans l'approche conditionnelle et le formalisme possibiliste dans l'approche logique sont les plus communément utilisés pour mettre en œuvre des raisonnements abductifs. Ces deux formalismes diffèrent principalement sur deux points. D'abord, les probabilités offrent un cadre strict et rigoureux, qui a été depuis longtemps éprouvé et appliqué à tout type d'incertitude, contrairement au cadre possibiliste. Ensuite, les probabilités sont des nombres n'ayant bien souvent que peu de lien direct avec le monde réel, convenant ainsi aux mesures numériques effectuées sur les régions mobiles. Inversement, les possibilités permettent d'attribuer une sémantique plus précise aux incertitudes,

s'adaptant ainsi à la diversité des propriétés symboliques relatives aux objets mobiles et s'exprimant en un langage proche du langage naturel. Cependant, bien que différents dans leurs propriétés secondaires, ces formalismes ne s'opposent pas (Friedman and Halpern, 1995), (Dubois and Prade, 1993). Notre contrainte principale étant d'avoir un formalisme uniforme sur tout le système, nous avons opté pour la logique possibiliste et l'approche logique (Dubois and Prade, 1992a). La motivation principale de ce choix est la plus grande facilité de modélisation de l'incertitude dans le cas d'une application réelle. Pour chaque propriété dont la valeur nécessite une phase de diagnostic, nous utilisons donc la logique possibiliste.

Une fois ce formalisme choisi pour mettre en œuvre les raisonnements abductifs, il reste à choisir le mode de mise à jour et de propagation de l'incertitude le long des liens de dépendance entre les propriétés. La méthode classique consiste à effectuer une mise à jour de l'incertitude pour chaque nouvelle image, suivie de sa propagation dans tout le système. Pour cette raison, nous organisons les propriétés en réseau. Cependant, nous n'utilisons ni la structure d'un ATMS, ni la structure d'un réseau bayesien. Un ATMS s'attache à réviser la croyance dans une propriété (c.-à-d. considérer une propriété comme invalide alors qu'elle était considérée comme vraie auparavant). Or, en ce qui nous concerne, les propriétés ont un degré continu de vraisemblance : il s'agit plutôt de mettre à jour la vraisemblance d'une propriété (c.-à-d. diminuer ou augmenter son degré de vraisemblance). Dans ce sens, un réseau bayesien est plus adapté à notre gestion de l'incertitude des propriétés. Cependant, on reproche généralement à ces réseaux la manière dont sont acquises les probabilités *a priori* et les matrices de probabilités conditionnelles. Ces probabilités apparaissent bien souvent comme des nombres magiques, dont on perçoit difficilement les liens avec le monde réel et dont on comprend peu la génération. Ces remarques sont d'autant plus vraies que la taille du réseau envisagé est importante. Pour utiliser un formalisme plus proche du monde réel, nous utilisons donc la logique possibiliste plutôt que la formule de Bayes. Néanmoins, même dans le cadre de la logique possibiliste, il est nécessaire de choisir des coefficients établis de façon *a priori*. Par exemple, il est nécessaire de pondérer la relation : *« la valeur de la propriété* **v** *peut s'expliquer par l'effet observé* **e** *»*. L'avantage de la logique possibiliste est de faire correspondre de ces coefficients directement avec des expérimentations du système et de faciliter leur ajustement. Ces coefficients sont ainsi plus faciles à initialiser. De plus, les mécanismes de la logique possibiliste sont plus facilement compréhensibles que ceux des probabilités, grâce au type d'opération utilisé. Les probabilités sont manipulées à travers des opérations de multiplication et de division, tandis que les possibilités

sont manipulées à travers des opérations de maximisation et de minimisation. Cette différence du type d'opération permet une stabilisation rapide de la propagation de l'incertitude dans un réseau utilisant le formalisme possibiliste, tandis le formalisme probabiliste favorise une évolution progressive de l'incertitude. Comme il s'agit de mettre à jour le réseau à chaque nouvelle image, le formalisme possibiliste convient mieux.

Pour propager l'incertitude dans le réseau des propriétés nous utilisons alors, sur le modèle des réseaux bayesiens, le formalisme de la logique possibiliste.

5.3.2 Mise en œuvre du diagnostic

Maintenant que le cadre de représentation et le mode de propagation sont choisis, il reste à définir la mise en œuvre du calcul de l'incertitude. Cette sous-section décrit alors la phase de diagnostic, puis propose un mécanisme de propagation de l'incertitude.

La phase de diagnostic d'une propriété consiste à détecter les symptômes (ou les effets) permettant (1) de justifier la valeur de la propriété ou (2) de témoigner d'incohérences (ou de désordres) dues à des erreurs de détection et de suivi des régions mobiles. Ensuite, la phase consiste à déterminer si la justification permet de mieux expliquer les symptômes que les désordres et de modifier en conséquence la valeur de la propriété et son degré de vraisemblance. La justification de la valeur de la propriété est obtenue par défaut, lorsque les désordres identifiés sont peu importants. Il n'est ainsi nécessaire que de représenter et de diagnostiquer les désordres. Pour chaque propriété, on possède $\mathcal{S} = \{ s_j \ / \ j \in < 1, n > \}$, l'ensemble de ses symptômes et $\mathcal{D} = \{ d_i \ / \ i \in < 1, p > \}$, l'ensemble de ses désordres. Soit $\mathcal{R}$, la relation qui associe les désordres aux symptômes ; $(d_i, s_j) \in \mathcal{R}$ indiquant que si le désordre d_i est présent alors le symptôme s_j est observé. Pour chaque désordre d_i, on possède également $\mathcal{S}(d_i)$, l'ensemble des symptômes causés par d_i : $\mathcal{S}(d_i) = \{ s_j \in \mathcal{S} \ / \ (d_i, s_j) \in \mathcal{R} \}$. Toutes ces connaissances sont prédéfinies. La phase de diagnostic consiste à calculer $\mathcal{S}_{\mathcal{OBS}} \subseteq \mathcal{S}$, l'ensemble des symptômes observés et à en déduire $\mathcal{D}_{\mathcal{SOL}} \subseteq \mathcal{D}$, l'ensemble des désordres solutions du diagnostic de $\mathcal{S}_{\mathcal{OBS}}$. Dans (Dubois and Prade, 1992a), on donne plusieurs façons pour définir $\mathcal{D}_{\mathcal{SOL}}$. Un exemple de définition est :

$$\mathcal{D}_{\mathcal{SOL}} = \{ d_i \in \mathcal{D} \ / \ \mathcal{S}(d_i) \subseteq \mathcal{S}_{\mathcal{OBS}} \} \tag{5.2}$$

Ce type de définition n'est pas satisfaisant compte-tenu des spécificités du problème. Les symptômes et les désordres sont interdépendants. Deux

désordres distincts peuvent être révélés par un même symptôme et réciproquement, deux symptômes distincts peuvent correspondre au même désordre. De plus, l'implication $d_i \rightarrow s_j$ n'est pas totale. Pour ces raisons, on considère $\mathcal{S}_{\mathcal{OBS}}$, $\mathcal{S}(d_i)$ et $\mathcal{D}_{\mathcal{SOL}}$ comme des ensembles flous. On note $\mu_{\mathcal{E}}$, la fonction d'appartenance à un ensemble flou $\mathcal{E}$ ($\mu_{\mathcal{E}} : \mathcal{E} \longrightarrow [0,1]$). On définit alors $\mathcal{D}_{\mathcal{SOL}}$ comme étant la couverture pertinente de $\mathcal{S}_{\mathcal{OBS}}$ par $\mathcal{R}$ (Dubois and Prade, 1992a), c.-à-d. l'ensemble des désordres pertinents expliquant $\mathcal{S}_{\mathcal{OBS}}$. Intuitivement, cette définition exprime juste que $\mathcal{R}(\mathcal{D}_{\mathcal{SOL}})$ est proche de $\mathcal{S}_{\mathcal{OBS}}$ mais n'implique pas l'inclusion de l'un des ensembles dans l'autre. La fonction d'appartenance à $\mathcal{D}_{\mathcal{SOL}}$ est donnée par la formule suivante :

$$\mu_{\mathcal{D}_{\mathcal{SOL}}}(d_i) = max_j[\, min(\mu_{\mathcal{S}(d_i)}(s_j),\, \mu_{\mathcal{S}_{\mathcal{OBS}}}(s_j))\,] \tag{5.3}$$

Cette formule consiste à calculer l'importance relative de la taille de $\mathcal{S}(d_i) \cap \mathcal{S}_{\mathcal{OBS}}$ et permet ainsi de calculer $\mathcal{D}_{\mathcal{SOL}}$ à partir de $\mathcal{S}_{\mathcal{OBS}}$ et des $\mathcal{S}(d_i)$.

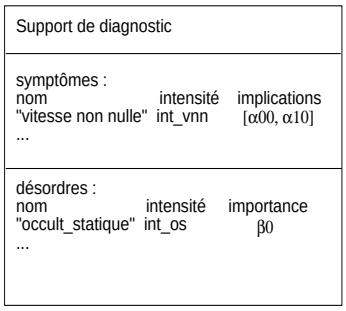

FIG. 5.7 – *Support de la méthode de diagnostic utilisée pour calculer le degré de vraisemblance de la propriété « diminution de taille », décrite ci-après dans la section 5.4.*

On représente le support de la méthode de diagnostic par l'ensemble de ses symptômes et de ses désordres comme le montre la figure 5.7. Un symptôme est caractérisé par son nom (p. ex. *« vitesse trop grande »*), par son intensité (définie ci-après) et par des coefficients d'implication (notés $\alpha_{i,j}$). Ces coefficients sont prédéfinis et pondèrent la relation $(d_i, s_j) \in \mathcal{R}$: la présence de d_i cause l'observation de s_j avec un degré $\alpha_{i,j}$. Ils mesurent le degré d'appartenance de s_j aux $\mathcal{S}(d_i)$: $\mu_{\mathcal{S}(d_i)}(s_j) = \alpha_{i,j}$.

La détection ou l'observation d'un symptôme est le résultat de mesures sur les régions mobiles constituant l'objet mobile. Par exemple, si la vitesse des régions mobiles est supérieure à un seuil alors le symptôme est

dit observé. Ce seuil appartient au contexte de l'application. Si cette vitesse courante est supérieure au seuil, l'intensité du symptôme est alors égale au rapport de la vitesse courante par la vitesse maximale et détermine son degré d'appartenance à $\mathcal{S}_{\mathcal{OBS}}$: $\mu_{\mathcal{S}_{\mathcal{OBS}}}(s_j) = \text{intensité}_{s_j}$. Le calcul de l'intensité des symptômes permet donc de déterminer complètement $\mathcal{D}_{\mathcal{SOL}}$.

Un désordre est représenté par son nom (p. ex. *« mauvais suivi d'objets »*), par son intensité et par un coefficient d'importance, noté β_i. On calcule l'intensité courante du désordre, notée intensité-crt, directement à partir de l'appartenance à $\mathcal{D}_{\mathcal{SOL}}$, puis on met à jour son intensité moyenne à l'instant t, notée intensité(t), en tenant compte de l'historique du désordre :

$$\begin{array}{l} \text{intensité-crt}_{d_i} = \mu_{\mathcal{D}_{\mathcal{SOL}}}(d_i) \\ \text{intensité}_{d_i}(t) = [\text{intensité-crt}_{d_i} + \text{intensité}_{d_i}(t-1)] \ / \ 2 \end{array} \tag{5.4}$$

L'intensité moyenne initiale du désordre est donnée par le contexte de l'application. Comme l'explique le chapitre 2, la base de contexte contient ces incertitudes *a priori*, appelées également incertitudes initiales. L'intensité moyenne est utilisée pour filtrer les gros écarts sur le calcul de l'incertitude et pour obtenir une intensité du désordre plus régulière. Le coefficient β_i mesure l'importance du désordre d_i dans le calcul du degré de vraisemblance de la propriété, défini comme l'opposé de l'incertitude.

Dans cette section nous avons décrit le calcul de l'intensité des désordres, responsables de l'incohérence d'une propriété. La partie suivante explique le calcul et la mise à jour du degré de vraisemblance de la propriété à partir de ces désordres.

5.3.3 Calcul du degré de vraisemblance

La détermination du degré de vraisemblance d'une propriété se passe en deux temps : (1) son calcul à partir des désordres qui lui sont attachés (phase de diagnostic) et (2) sa mise à jour à partir de la propagation du degré de vraisemblance des propriétés dépendantes (phase de propagation ascendante et descendante).

Le calcul du degré de vraisemblance courant, noté vraisemblance-crt, s'obtient en pondérant l'intensité des désordres par leur degré d'importance :

$$\begin{array}{l} \text{vraisemblance-crt} = 1 - max_i[\min(\beta_i, \text{intensité}_{d_i}(t))] \\ \text{vraisemblance(t)} = [\ \text{vraisemblance-crt} + \text{vraisemblance(t-1)}\ ] \ / \ 2 \end{array} \tag{5.5}$$

Le degré de vraisemblance moyen à l'instant t, noté vraisemblance(t), se calcule en combinant le degré de vraisemblance courant avec l'ancienne valeur du degré de vraisemblance ; le degré de vraisemblance initial étant obtenu à l'aide du contexte.

Le deuxième temps concerne sa mise à jour par la propagation du degré de vraisemblance des propriétés, le long des liens de dépendances. Le type de propagation est différente selon la phase d'enchaînement. Pour la phase d'abstraction des données, la propagation du degré de vraisemblance des propriétés en amont, vers les propriétés directement dépendantes en aval (c.-à-d. les propriétés les plus abstraites) est implicite dans le modèle. Elle s'effectue lors du calcul de la valeur de la propriété. Ces calculs utilisent comme données d'entrée les propriétés dépendantes et sont pondérés par le degré de vraisemblance des propriétés. Par exemple, le calcul de la vitesse d'un objet mobile s'effectue à partir de la vitesse des régions mobiles qui le constituent. Si l'une des régions est particulièrement incertaine, elle ne sera pas prise en compte. Comme le degré de vraisemblance de la propriété se diagnostique à partir de ses propriétés en amont, elle dépend directement de leur degré de vraisemblance.

Pour la phase de contrôle, le système utilise la vraisemblance des propriétés directement dépendantes en aval pour corriger le degré de vraisemblance de la propriété courante (c.-à-d. une propriété plus élémentaire). Cette correction dépend du type de la propriété et en particulier de sa pérennité dans le système. En ce qui concerne l'interprétation, plus une propriété est en aval, plus sa durée de vie est longue donc plus les corrections de son degré de vraisemblance doivent être progressives. Par exemple, le degré de vraisemblance de la propriété *« l'individu se dirige vers la voiture »*, doit être corrigé progressivement, car cette propriété se déroule sur un grand intervalle de temps. Pour le lien de dépendance des propriétés en aval vers les propriétés en amont, si le degré de vraisemblance de la propriété en aval, noté *vraisemblance-aval*, augmente et s'il est supérieur à un certain seuil de vraisemblance, alors on augmente d'un pas le degré de vraisemblance de la propriété en amont directement dépendante, noté *vraisemblance-amont*. Ce pas dépend de la pérennité de la propriété en aval par rapport à celle de la propriété en amont. Dans le cas opposé, on diminue le degré de vraisemblance de la propriété en amont, du même pas :

$$
\begin{array}{l}
\text{si (vraisemblance-aval(t)>vraisemblance-aval(t-1))} \\
\text{et (vraisemblance-aval(t)>seuil)} \\
\quad \text{alors vraisemblance-amont(t)} = \min[1, \text{vraisemblance-amont(t)} + \textit{pas}] \\
\text{sinon si (vraisemblance-aval(t)<vraisemblance-aval(t-1))} \\
\text{et (vraisemblance-aval(t)<seuil)} \\
\quad \text{alors vraisemblance-amont(t)} = \max[0, \text{vraisemblance-amont(t)} - \textit{pas}]
\end{array}
\tag{5.6}
$$

5.4 Exemple d'utilisation

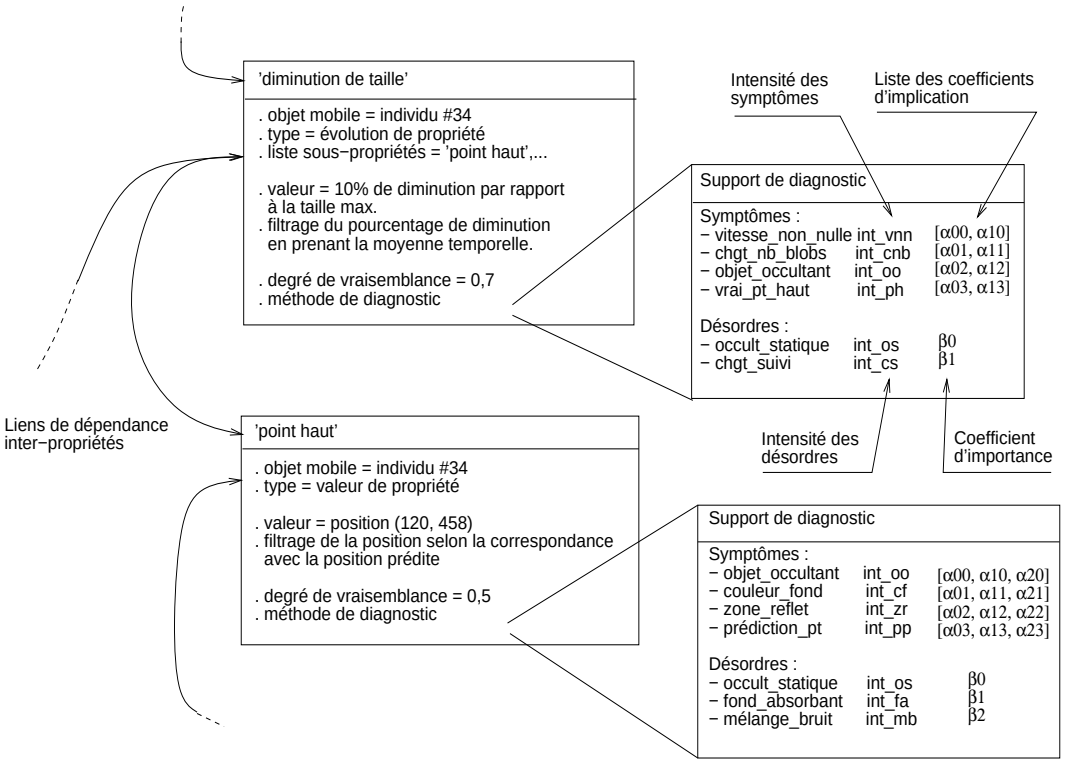

FIG. 5.8 – *Cette figure montre une vue partielle d'un réseau de propriétés. Dans cet exemple, la propriété « diminution de taille », utilise la valeur de la propriété plus élémentaire « point haut ».*

Cette section décrit un réseau de propriétés utilisant le formalisme présenté dans ce chapitre. Sur la figure 5.8, on considère deux propriétés caractérisant un individu en mouvement. La propriété *« point haut »*, permet de

calculer la position du point le plus haut de l'individu, correspondant en général au sommet de sa tête. La valeur de cette propriété est obtenue à partir du point générique *« point haut »*, des régions mobiles constituant l'individu. Pour cela, on compare la nouvelle position du point haut (mesurée sur la nouvelle image) avec sa position prédite à l'aide de la trajectoire de l'objet mobile. Si les deux positions se correspondent, on met à jour la position du point. Si elles ne se correspondent pas, on regarde le degré de vraisemblance de la propriété. Si ce degré est important (peu de désordres diagnostiqués), on met à jour la valeur de la propriété à l'aide de la nouvelle position mesurée. Sinon, si le degré de vraisemblance est faible, on garde l'ancienne valeur de la propriété. Les phénomènes extérieurs (c.-à-d. les désordres) pouvant interférer dans ce calcul sont : une occultation statique partielle, une couleur absorbante du fond de la scène (c.-à-d. un manque de contraste) et un mélange de l'individu à un bruit, comme un reflet. À chacun de ces désordres, on attache un coefficient *a priori* d'importance. Par exemple, le coefficient d'importance β_0 du désordre *« occultation statique »*, est plus important que les autres coefficients d'importance, indiquant que l'influence de ce désordre dans le calcul du degré de vraisemblance de la propriété est la plus grande. Pour chacun de ces désordres, on calcule leur intensité en fonction des symptômes attachés à la propriété : présence d'un objet occultant, couleur du fond de la scène proche de la couleur de l'individu, présence d'un reflet et mauvaise correspondance entre les positions mesurée et prédite du point. Par exemple, les intensités des désordres *« occultation statique »* et *« fond absorbant »* associés à la propriété *« point haut »* décrite sur la figure 5.8, sont calculées à l'aide des formules (5.3), (5.4) :

$$\text{int-crt-os} = \max[\,\min(\alpha_{0,0}, \text{int-oo}), \min(\alpha_{0,1}, \text{int-cf}), \min(\alpha_{0,2}, \text{int-zr}), \min(\alpha_{0,3}, \text{int-pp})\,]$$
$$\text{int-os(t)} = [\text{int-crt-os} + \text{int-os(t-1)}] \,/\, 2$$
$$\text{int-crt-fa} = \max[\,\min(\alpha_{1,0}, \text{int-oo}), \min(\alpha_{1,1}, \text{int-cf}), \min(\alpha_{1,2}, \text{int-zr}), \min(\alpha_{1,3}, \text{int-pp})\,]$$
$$\text{int-fa(t)} = [\text{int-crt-fa} + \text{int-fa(t-1)}] \,/\, 2$$

L'intensité des symptômes utilisés, indique le degré avec lequel ces symptômes sont observés. Par exemple, pour le symptôme *« objet occultant »*, l'intensité de ce symptôme est obtenue par consultation de la base de contexte, en vérifiant la présence d'un objet statique. À chacun de ces symptômes, on attache une liste de coefficients d'implication *a priori*, ces coefficients appartenant aux connaissances du domaine. Par exemple, on utilise le coefficient $\alpha_{0,0}$ pour quantifier la relation : *« l'observation du symptôme* 'objet occultant' *est un indice de la présence du désordre* 'occultation statique' *»*.

La propriété *« diminution de taille »*, décrite également sur la figure 5.8, détermine si la taille de l'individu diminue. Cette propriété a pour type *« évolution »*, et dépend entre autres, de la sous-propriété *« point haut »*. La valeur de la propriété est le pourcentage de diminution de la taille courante, par rapport à la taille maximale qui a été calculée pour l'individu. On calcule cette valeur en filtrant les pourcentages calculés : on prend le pourcentage moyen obtenu sur les cinq dernières images. Les phénomènes extérieurs (désordres) pouvant interférer dans le calcul de la propriété, sont les occultations statiques partielles (la partie supérieure de l'individu) et le changement de la constitution de l'individu au cours de son suivi. Par exemple, le désordre *« changement de suivi »*, rend compte des situations lorsque le suivi de l'individu se mélange au suivi d'un autre objet mobile. Les symptômes permettant de déterminer la présence de ces désordres sont : une vitesse non nulle, le changement du nombre de régions mobiles constituant l'individu et la présence d'un objet occultant. Par exemple, une vitesse non nulle témoigne d'une occultation partielle potentielle de l'individu, lorsque son déplacement le mène derrière un objet statique. Le degré de vraisemblance de la propriété se calcule comme précédemment. La propriété *« diminution de taille »*, dépendant directement de la propriété *« point haut »*, le calcul de la valeur de la propriété *« diminution de taille »* et de son degré de vraisemblance dépendent alors de la valeur de la propriété *« point haut »* et de son degré de vraisemblance. Lors de la phase d'abstraction des propriétés (propagation ascendante), le degré de vraisemblance de la propriété *« point haut »*, est utilisé pour calculer le degré de vraisemblance de la propriété *« diminution de taille »*. Après cette phase d'évolution temporelle (phase de diagnostic), on met à jour le degré de vraisemblance, noté vraisemblance(t) (c.-à-d. l'intensité moyenne de l'ensemble des désordres), à l'aide de la formule (5.5) :

$$\text{vraisemblance-crt} = 1 - \max[\ \min(\beta_0, \text{int-os}), \min(\beta_1, \text{int-fa}), \min(\beta_2, \text{int-mb})\ ]$$
$$\text{vraisemblance(t)} = [\ \text{vraisemblance-crt} + \text{vraisemblance(t-1)}\ ] \ / \ 2$$

Enfin, pendant la phase de contrôle de l'interprétation (propagation descendante), en supposant que la vraisemblance dans la propriété *« diminution de taille »*, augmente et qu'elle soit supérieure à un seuil, on corrige le degré de vraisemblance de la propriété *« point haut »*, à l'aide de la formule (5.6) :

$$\text{vraisemblance(t+1)} = \min[1, \text{vraisemblance(t)} + \text{pas}]$$

L'utilisation de ce formalisme permet d'exprimer les incertitudes sur les propriétés des objets mobiles, mais aussi sur le processus même d'interprétation. L'incertitude de l'interprétation d'un objet mobile est obtenue en regardant le degré de vraisemblance maximal des propriétés les plus pertinentes concernant son comportement. En pratique, les propriétés pertinentes d'un objet mobile sont ses propriétés les plus abstraites et leur degré d'importance est spécifié par l'utilisateur. L'incertitude du processus global d'interprétation de la séquence d'images s'obtient alors à l'aide des incertitudes sur les objets mobiles analysés.

5.5 Conclusion

Ce chapitre propose un formalisme ayant pour but la construction d'un **réseau de propriétés**, permettant d'abstraire les propriétés caractérisant un objet mobile. Les liens du réseau représentent les dépendances inter-propriétés. Ces propriétés sont calculées à partir de mesures numériques et ont une valeur symbolique caractéristique de ces mesures sur un court intervalle de temps. Elles quantifient la valeur de ces mesures, mais aussi leur manière d'évoluer dans le temps. Elles permettent également de masquer les irrégularités des valeurs numériques (p. ex. leur grande variation) et de palier à la présence de valeurs aberrantes ou à l'absence de valeur. Ces propriétés sont des hypothèses servant d'élément de base pour la reconnaissance de scénarios. Le problème essentiel de ce type de réseau est la **gestion de l'incertitude** des propriétés. Le système présenté dans ce chapitre calcule un degré de vraisemblance de ces propriétés et propage leur mise à jour dans le réseau à l'aide de la logique possibiliste. Les degrés de vraisemblance sont calculés par une **phase de diagnostic**. S'ils sont suffisamment élevés, les propriétés correspondantes peuvent être utilisées. Ces degrés sont propagés par une phase ascendante vers les propriétés les plus abstraites, afin de maintenir la cohérence du réseau. On utilise également une phase de propagation descendante pour corriger en retour les valeurs des degrés.

Le **formalisme** proposé utilise le cadre de la logique possibiliste, afin d'adapter le calcul de l'incertitude aux spécificités des propriétés considérées et de prendre en compte la diversité du monde réel. Ce formalisme permet également de traiter de la même manière, l'incertitude sur les mesures (numériques) des régions mobiles et l'incertitude sur les propriétés (symboliques) des objets mobiles.

Ces travaux ont été **intégrés** au module de reconnaissance de scénarios et ont permis de le relier au module de suivi des régions mobiles.

Les travaux futurs consistent à définir des propriétés supplémentaires pour analyser de nouveaux types d'activités humaines, comme l'analyse des gestes d'un individu. Un second axe consiste aussi à automatiser la construction du réseau de propriétés afin de faciliter le développement de nouvelles applications. Ce second axe est plus amplement décrit dans la section 6.5.

Chapitre 6

Reconnaissance de scénarios

Le but de ce chapitre est d'expliquer la reconnaissance de scénarios à partir des propriétés symboliques calculées sur les objets mobiles. Ces scénarios sont décrits dans un formalisme proche du langage naturel et correspondent aux comportements et aux activités d'objets mobiles, jugés intéressants par un opérateur humain. *« L'individu se dirige vers la voiture et tourne autour d'elle »* et *« le groupe d'individus entre sur le quai du métro et se disperse avant de monter dans la rame »* sont des exemples de scénarios que nous souhaitons reconnaître. Le problème de la reconnaissance de scénarios réside alors (1) dans la représentation des scénarios décrits dans un formalisme proche du langage naturel, et (2) dans la définition de méthodes permettant de rattacher ces scénarios aux propriétés symboliques qualifiant la perception des mouvements des objets mobiles.

La reconnaissance de scénarios est l'étape finale du processus global d'interprétation de séquences d'images. L'ensemble des scénarios reconnus constitue l'interprétation de la scène en fonction des besoins d'un utilisateur. Dans ce chapitre, la section 6.1 présente un état de l'art, d'une part sur des travaux décrivant des activités humaines en langage naturel, d'autre part sur des méthodes permettant de reconnaître de telles activités. Dans la section 6.2, nous proposons une approche du problème de reconnaissance de scénarios en définissant la notion de scénario. Puis dans la section 6.3, nous expliquons la nature des principaux scénarios et propriétés utilisés pour analyser des activités humaines. Enfin nous décrivons une implantation de notre approche dans la section 6.4 en proposant des méthodes de reconnaissance de scénarios utilisant en particulier des automates de reconnaissance.

6.1 État de l'art

Cette section présente plusieurs travaux reliant la description d'actions et d'activités en langage naturel à partir de propriétés calculées par des traitements d'images. Cette section présente d'abord des travaux en langage naturel établissant les primitives et relations de base permettant de décrire une action. Ensuite, cette section présente plusieurs systèmes utilisant des descriptions dans un formalisme proche du langage naturel, afin de reconnaître des actions et d'activités.

6.1.1 Description d'actions en langage naturel

Dans (Sablayrolles, 1995), l'auteur donne une sémantique formelle de l'expression du mouvement. De cette étude complète, nous reprenons dans ce rapport uniquement les résultats établissant des spécifications sur la description sémantique d'actions en langage naturel. La reconnaissance d'une action repose sur la construction progressive de référents sémantiques. Cette construction consiste dans une identification de primitives spatio-temporelles, comme le sont les trajectoires, qui sont par la suite reliées entre elles selon des relations de temps, de topologie et de géométrie.

P. Sablayrolles présente alors trois approches utilisées pour décrire les actions et plus particulièrement les verbes de mouvement dans la langue française. La première approche qualifie l'action par rapport à l'espace. Par exemple, il définit la classe de verbes, comme le verbe *« sortir »*, débutant dans une zone de référence et se terminant à l'extérieur. Les relations temporelles et topologiques sont souvent représentées à l'aide de prépositions, telles que *« dans »* ou *« vers »*. Cette approche souligne l'importance du **référentiel spatial** dans la description d'une action. L'auteur décrit trois types de référentiel spatial : déïctiques, intrinsèques ou extrinsèques. On utilise le référentiel déïctique lorsqu'on décrit le mouvement de l'objet mobile par rapport à l'observateur. On utilise le référentiel intrinsèque lorsqu'on décrit le mouvement de l'objet mobile par rapport à lui-même. On utilise le référentiel extrinsèque lorsqu'on décrit l'objet mobile par rapport à un référentiel absolu comme le coin d'une pièce. La deuxième approche étudie les contraintes élémentaires sur l'action définie comme étant un **changement quantifiable** d'une dimension, d'une manière ou d'une intentionalité. Par exemple, le verbe *« fuguer »* met en évidence le changement d'intentions de l'acteur qui prend ici, la décision de s'échapper. Une troisième approche décrit les verbes de mouvement en fonction de la sémantique du lieu de référence, mettant ainsi l'accent sur le **contexte de l'action**. Par exemple, la

description du verbe *« marcher »* n'est pas la même, si l'action se déroule dans une maison ou dans un train. Cette approche explique l'utilisation de symboles pour décrire des actions, plutôt que des primitives numériques. Le langage naturel manipule ainsi plutôt des lieux (portions d'espace possédant un nom et des fonctionnalités) et des postures (descriptions reliées à une intention, p. ex. *« en attente »*) , que des emplacements (portions de surfaces occupées par un objet mobile) et des formes (descriptions géométriques). Une caractéristique essentielle de la description d'une action est ainsi l'utilisation de propriétés imprécises bien que non ambiguës. Les ambiguïtés sont en effet levées par l'utilisation du contexte.

Les travaux présentés dans (Sablayrolles, 1995) concluent en proposant différentes classifications des verbes de mouvement, en fonction de l'approche considérée. L'auteur met ainsi en évidence les principales caractéristiques permettant de décrire une action en langage naturel. Ces caractéristiques servent de support dans la plupart des systèmes essayant de reconnaître des actions décrites en langage naturel. Ces systèmes privilégient une partie de ces caractéristiques selon le type de l'application traité. Par exemple dans (Schirra, 1992), (Schirra and Stopp, 1993), les auteurs définissent une technique, les images mentales, permettant de construire les référents sémantiques d'une action, au cours de sa reconnaissance. Ces images mentales sont définies par un ensemble de contraintes spatio-temporelles. Dans (Olivier et al., 1994), les auteurs mettent l'accent sur l'utilisation du référentiel déïctique, pour reconnaître une action. L'importance de ce référentiel est également soulignée dans (Howarth, 1994). La problématique consiste alors à relier ces caractéristiques de description des actions à un ensemble de propriétés calculables à partir de traitements informatiques.

6.1.2 Reconnaissance d'actions à partir de propriétés symboliques

Dans (Srihari, 1994), l'auteur étudie les travaux essayant de relier des informations numériques et visuelles à des descriptions sémantiques en langage naturel. Elle met en évidence les problèmes de reconnaissance d'actions dus aux difficultés de détection de primitives de traitement d'images permettant de discriminer les actions. Elle conclue que la reconnaissance d'actions est une problématique nouvelle, qui reste pour le moment ouverte. Cependant plusieurs travaux ont abordé cette problématique et cette sous-section présente leurs caractéristiques.

Inférences logiques

Il existe un grand nombre de logiques, permettant de décrire formellement les propriétés du monde. Elles sont en général dédiées à l'étude des propriétés spatiales ou temporelles et certaines permettent même de décrire des propriétés spatio-temporelles.

- **Logiques spatiales** : Dans (Olivier et al., 1994), les auteurs présentent un système qui génère automatiquement une description spatiale d'une pièce donnée. L'utilisateur décrit la pièce et sa position à l'aide de prépositions spatiales. Puis, le système présenté comprend la description de la pièce, la valide, et peut alors générer de nouvelles descriptions en considérant, par exemple, la pièce sous un angle de vue différent. Ces descriptions sont réalisées en langage naturel et sont construites à partir de prépositions spatiales. L'originalité de ces travaux réside dans la prise en compte d'ambiguïtés qualitatives et quantitatives concernant les propriétés spatiales des objets de la pièce. Par exemple, il est possible de préciser le référentiel de la pièce (si l'utilisateur est pris comme référentiel, alors la chaise est vue comme étant devant le bureau) et de quantifier symboliquement la distance entre deux objets, avec des termes tels que *« plus ou moins loin »*.

  De même dans (Beringer et al., 1993), les auteurs présentent un système qui détermine si une déclaration donnée est une description consistante d'une scène donnée. Une déclaration est une conjonction de prépositions; chaque préposition correspondant soit à la propriété d'un objet, comme sa couleur, ou soit à une propriété spatiale, comme *« l'objet x est à droite de l'objet y »*. L'intérêt de ce système est qu'après une période d'apprentissage à partir d'un ensemble de prépositions, le système peut vérifier la consistance de toute nouvelle déclaration.

- **Logiques temporelles** : Parmi les nombreux travaux réalisés en logique temporelle, on peut citer ceux de (Lansky, 1988) où l'auteur a développé une logique temporelle du premier ordre, basée sur la gestion d'événements. Ce système se caractérise par la diversité et la richesse d'expression des contraintes temporelles qu'il manipule. Ces contraintes peuvent aussi bien décrire des relations temporelles que des relations de causalité et de simultanéité, telles que *« deux robots soulèvent en même temps le cube »*. Ces travaux ont en particulier inspirés ceux du projet européen VIEWS, dédié à la surveillance de trafic routier.

De même dans (Pinhanez and Bobick, 1997), les auteurs utilisent un réseau de contraintes basées sur l'algèbre d'intervalles d'Allen afin de gérer des événements, organisés sous la forme d'un réseau. En propageant les contraintes temporelles à travers le réseau, les auteurs peuvent reconnaître une action telle que *« prendre un bol »*. L'originalité de ces travaux est d'avoir relié une logique temporelle avec un système de vision détectant des événements non simulés. Ce système se caractérise également par la prise en compte d'un certain nombre d'erreurs, telles que l'absence de détection d'un événement due à la panne d'un capteur. En tenant compte de ces erreurs, le système reste capable de reconnaître des actions.

- **Logiques spatio-temporelles** : Il existe également des logiques intégrant les notions d'espace, de temps et de mouvements. Par exemple, dans (Galton, 1993), l'auteur propose de rajouter à la logique temporelle d'Allen un ensemble de relations spatiales. Il définit alors un événement comme une transition entre deux relations spatiales. Il peut ainsi exprimer qu'un événement se produit dans un intervalle de temps ou à un instant donné.

Toutes ces logiques définissent un ensemble de contraintes (appelées également règles) manipulant des propositions spatiales, temporelles ou spatio-temporelles comme des briques de base. Ces contraintes permettant une manipulation efficace et sophistiquée de ces propositions aident alors à décrire des relations plus complexes. Cependant pour la plupart, elles ne prennent en compte ni l'imprécision, ni les erreurs de ces propositions, ni la diversité du monde réel. La mise en œuvre de ces logiques afin de traiter une application réelle, est alors rarement abordée. Ces logiques étant proches de descriptions en langage naturel, il semble plus réaliste de les utiliser dans un module interfaçant un système de reconnaissance d'actions avec un module de description de scénarios.

Hiérarchies *a priori*

Une autre approche consiste à décrire des scénarios grâce à des hiérarchies de verbes ou d'événements définies de façon *a priori*, afin de prendre en compte la diversité du monde réel. Ces hiérarchies étant dépendantes du domaine d'application, les systèmes de reconnaissance de scénarios qui leur sont associés sont décrits en fonction de l'application envisagée :

- **Repérage dans une ville et analyse d'un match de football** : Le projet VITRA - VIsual TRAnslator - a pour objectif de transformer des

informations provenant d'images et décrivant des objets mobiles, en des textes en langage naturel (André et al., 1988), (Retz-Schmidt, 1991), (Herzog et al., 1989) et (Herzog, 1995). Dans le projet VITRA, deux systèmes ont été développés. Le premier système, *« City-tour »*, décrit des scènes citadines et peut répondre à un certain nombre de questions concernant les déplacements d'objets mobiles de la scène, principalement des voitures et des piétons. Ce système représente les actions des objets mobiles, telles que *« le piéton se dirige vers l'église »*, à l'aide de leur trajectoire. Le second système, *« Soccer »*, décrit un match de football comme le ferait un commentateur à la radio. Les actions sont décrites au fur et à mesure qu'elles se déroulent (c'est-à-dire en direct) et ne peuvent être reconnues avant d'être complètement achevées. Cette contrainte implique pour un objet mobile qu'aucune trajectoire complète n'est disponible. La description des actions est alors calculée à travers la posture des joueurs correspondant à des clichés instantanés de mouvements, tels que *« dribbler »* et *« passer la balle »*. Les actions sont reconnues à l'aide d'un graphe orienté et étiqueté appelé diagramme de durée. Les nœuds représentent les postures des joueurs et les arêtes représentent les prédicats sur leurs mouvements, comme *« déclencher »*, *« continuer »*, *« réussir »* ou *« arrêter »* (Herzog, 1995). Ce système de description des actions est complété par un module *« Antilima »* (Schirra and Stopp, 1993), qui a pour objectif de représenter l'image mentale d'un auditeur écoutant les commentaires du système, décrivant le match de football. Dans le projet VITRA, les actions sont décrites à partir de hiérarchies *a priori* de verbes de mouvement. Malheureusement, ces hiérarchies sont dépendantes des applications. Dans le système *« City-tour »*, les verbes sont décrits en fonction de la trajectoire des objets mobiles. Dans *« Soccer »*, les verbes sont décrits à partir des postures ponctuelles des objets mobiles. Un objectif de l'équipe de Vitra est d'essayer de définir une hiérarchie unique, pouvant convenir à tout type d'application.

Dans (Sol, 1997), l'auteur continue les travaux réalisés dans le cadre du système *« Soccer »*. Il propose une représentation des actions utilisant une hiérarchie de verbes basée sur la trajectoire des joueurs, permettant ainsi de prédire à plus long terme leurs actions futures et leurs intentions. Cependant, la connexion avec les postures ponctuelles du système *« Soccer »* n'a pas été effectuée.

– **Analyse du trafic routier** : De même, dans le domaine de la surveillance du trafic routier, plusieurs systèmes utilisent des hiérarchies

a priori de verbes. Ces hiérarchies sont constituées à leur base par des verbes élémentaires, comme *« changer de voie »*, et sont constituées à leur sommet par des verbes plus abstraits, comme *« doubler »*. Par exemple, dans (Neumann, 1989), l'auteur a été l'un des premiers à définir ce type de hiérarchie de verbes. Dans le projet européen Esprit VIEWS (Duong et al., 1990b), les auteurs s'inspirent de cette hiérarchie pour définir trois niveaux d'abstraction de verbes : *« changement »*, *« événement »* et *« comportement »*. Ils proposent d'utiliser un automate d'états finis pour reconnaître une succession de verbes de mouvement, sur le modèle d'un analyseur grammatical (Corrall, 1992). Dans (Nagel, 1988) et (Nagel, 1991), l'auteur utilise également une hiérarchie de verbes de mouvements, comportant jusqu'à neuf niveaux d'abstraction de verbes : *« changement »*, *« événement »* ... *« verbe-phrase »*, *« phrase »*, *« épisode »*, *« historique »*. Il propose d'utiliser un diagramme de transitions à chaque niveau d'abstraction, pour représenter la succession de ces verbes. Les nœuds du diagramme représentent les verbes à reconnaître, les flèches reliant ces nœuds représentent les transitions entre les verbes.

Dans (Buxton and Gong, 1995), les auteurs ont poursuivi les travaux réalisés dans le cadre de VIEWS, en décomposant la hiérarchie de verbes à l'aide de deux systèmes : système périphérique et système central. Le système périphérique regroupe les verbes et propriétés élémentaires tels que *« la vitesse »*, *« l'orientation du véhicule »*, *« l'évolution de sa vitesse »* et *« la différence entre sa vitesse et la vitesse d'un véhicule de référence »*. Le système périphérique est représenté par un réseau bayesien qui, au fil du flot des données, établit l'incertitude des verbes et des propriétés élémentaires. Le système central regroupe des verbes plus abstraits tels que *« suivre »*, *« doubler un véhicule de référence »*. Ce système est représenté par un réseau bayesien dynamique qui est construit automatiquement en fonction de la valeur des propriétés élémentaires. Bien que ce second niveau soit construit à la demande, il est prédéfini comme dans les systèmes précédents. L'intérêt du système central est de sélectionner l'attention du système sur la reconnaissance de certains scénarios et d'éviter une reconnaissance exhaustive de tous les scénarios de niveau d'abstraction supérieure.

– **Analyse des actions d'un robot** : Il existe également des systèmes utilisant des hiérarchies *a priori* de verbes, afin de reconnaître des actions réalisées par des robots. Par exemple, dans (Kuniyoshi and Inoue, 1993), les auteurs reconnaissent les actions d'un robot à l'aide

d'un ensemble de verbes de mouvement de base, tels que *« approcher la main »*, *« l'éloigner »* ou *« toucher »*. Ils utilisent un automate d'états finis pour reconnaître un enchaînement de ces verbes.

- **Analyse de comportements humains** : Enfin, il existe également des systèmes qui reconnaissent des actions accomplies par des êtres humains, à l'aide d'une hiérarchie *a priori* d'événements. Par exemple, dans (Royer, 1995) et (Castel et al., 1996), les auteurs utilisent trois niveaux de description des événements : *« changement »*, *« événement »* et *« scénario »* (appelé plan par les auteurs). Un changement indique la nouvelle position d'un objet mobile, un événement correspond à un changement intéressant (reconnu intéressant après une étape de classification) et un scénario représente une suite d'événements, auxquels on associe des contraintes sur les transitions inter-événements et sur le contexte de la scène. Un scénario correspond alors à un graphe temporel d'événements. Il est reconnu à l'aide d'un réseau de Pétri dont le marqueur correspond à l'état courant du plan. Ce système permet de reconnaître des scénarios tels que *« le piéton monte dans le véhicule et démarre »*.

Tous ces systèmes sont basés sur une hiérarchie de verbes de mouvements, définie de façon *a priori*. Ils n'utilisent qu'une notion simple de contrainte temporelle : la succession entre deux actions. Cependant, ces systèmes manipulent des propriétés dépendantes de l'application et caractérisant des verbes de mouvement qui permettent de bien s'adapter au monde réel. Par comparaison avec les inférences logiques, les hiérarchies de verbes sont ainsi mieux adaptées pour traiter des applications réelles, bien qu'ayant une notion de contrainte temporelle plus simple. Le principal inconvénient de ces hiérarchies *a priori* est leur manque de généricité. Pour chaque nouvelle action à reconnaître, il est nécessaire de définir spécifiquement une nouvelle hiérarchie de verbes de mouvement. Ces hiérarchies *a priori* ont également l'inconvénient de figer le nombre de niveaux d'abstraction nécessaires à la description des scénarios. Pour cette raison, H. Nagel a été obligé de définir un grand nombre de niveaux d'abstraction, afin d'avoir plus de souplesse pour décrire des scénarios.

Hiérarchies récursives

Plusieurs systèmes ont résolu le problème de la fixation du nombre de niveaux d'abstraction en définissant récursivement un scénario à partir de sous-scénarios. Par exemple dans (Dousson et al., 1993), les auteurs utilisent

une définition récursive de scénarios, appelée chroniques, afin de reconnaître les activités de robots se déplaçant dans une usine. Un scénario est caractérisé par ses instants de commencement et de terminaison, et par un ensemble de contraintes temporelles portant sur les sous-scénarios le constituant. Cette définition des scénarios est générique car aucune signification est attachée aux contraintes. Cette caractéristique présente l'inconvénient de ne pas pouvoir traiter spécifiquement une contrainte, pour prendre en compte des erreurs de détection, telles que la panne d'un capteur ou l'imprécision d'une détection. De même dans (Chleq and Thonnat, 1996), les auteurs utilisent ce formalisme pour reconnaître les activités d'êtres humains dans des applications de vidéo-surveillance. Initialement, tous les scénarios possibles sont générés et sont reconnus progressivement, en comparant leur description aux nouveaux événements détectés au cours du traitement. Ils reconnaissent par exemple qu'un individu est en attente devant un rayonnage de supermarché.

6.2 Approche proposée pour la reconnaissance

Cette section expose les solutions que nous avons adoptées pour décrire les activités humaines. Elle propose également une définition de la notion de scénario.

6.2.1 Description d'activités humaines

En se basant sur le précédent état de l'art (section 6.1) et sur l'expérience acquise dans le développement des applications cibles, nous avons résolu trois problèmes afin de décrire des activités humaines, pour ensuite pouvoir les reconnaître. Cette sous-section décrit ainsi les solutions choisies relatives à la description des activités humaines.

Le premier problème réside dans le choix du référentiel à adopter pour décrire des activités humaines. Dans le système présenté dans ce mémoire, nous choisissons d'utiliser le référentiel déïctique pour représenter les comportements des objets mobiles. Ce référentiel est naturel pour les applications de vidéosurveillance, car il permet de décrire les actions comme le ferait un opérateur humain regardant l'action se dérouler à l'écran. Nous pouvons alors utiliser directement l'expertise des agents de surveillance. De plus, c'est le référentiel déïctique qui est généralement utilisé dans les applications d'interprétation de séquences d'images. Ce référentiel permet de traiter le plus grand nombre d'applications.

Un second problème consiste à trouver une représentation des référents sémantiques d'une action. Ces référents sémantiques permettent de décrire

toute relation inter-objets, comprenant en particulier les relations de coopération ou de compétition entre les objets mobiles. Dans un scénario pour représenter les référents sémantiques, nous attribuons deux rôles aux objets mobiles : **objet source** ou **objet de référence**. L'objet source correspond à l'objet mobile qui réalise l'action à reconnaître. L'objet de référence correspond à la référence de l'action représentant soit un objet mobile, soit un objet statique de l'environnement de la scène. Cet objet statique peut également désigner une zone de la scène. On définit alors six types de scénarios selon la combinaison de ces rôles :

- Les scénarios avec un objet de référence appartenant au **contexte** et **proche** de l'objet source. Par exemple, dans le scénario *« l'individu s'asseoit sur le banc »*, *« l'individu »* est l'objet source et *« le banc »* est l'objet de référence.
- Les scénarios avec un objet de référence appartenant au **contexte** et **éloigné** de l'objet source (p. ex. *« l'individu se dirige vers la porte »*).
- Les scénarios avec un objet de référence **mobile** et **proche** de l'objet source (p. ex. *« l'individu touche l'autre individu »*).
- Les scénarios avec un objet de référence **mobile** et **éloigné** de l'objet source (p. ex. *« l'individu suit l'autre individu »*).
- Les scénarios **sans objet de référence** et avec un objet source correspondant à **un** objet réel (p. ex. *« l'individu avance puis revient en arrière »*).
- Les scénarios **sans objet de référence** et avec un objet source correspondant à **un groupe** d'objets réels (p. ex. *« le groupe d'individus s'éparpille puis se rassemble »*).

Un troisième problème concerne le traitement des groupes d'individus, dont il est souvent difficile de déterminer et de séparer les différents éléments. Dans ce cas un groupe d'individus est représenté à l'aide d'un seul objet mobile. Les scénarios relatifs à un objet mobile de type groupe permettent ainsi de représenter le comportement global du groupe. Dans le cas où il y a une ambiguïté dans la nature d'un objet mobile groupe (par exemple, si le groupe peut correspondre en réalité à un seul individu), nous analysons d'abord sur cet objet mobile groupe aussi bien les scénarios mettant en jeu un groupe que les scénarios mettant en jeu un seul individu. Dans un second temps, lorsque l'analyse des scénarios est suffisamment avancée, nous comparons le degré de

vraisemblance des différents scénarios. Selon ces degrés, nous décidons alors de la nature réelle de ces objets mobiles. Cette méthode de reconnaissance revient à réaliser une double analyse de scénarios en cas d'ambiguïté.

Nos choix concernant la description des activités humaines s'appuient sur les travaux de P. Sablayrolles qui a étudié exhaustivement tous les verbes de mouvement de la langue française. Pour cette raison, nous pensons que cette représentation permet de décrire les activités d'objets mobiles intervenant dans la plupart des systèmes d'interprétation de séquences d'images.

6.2.2 Définition de la notion de scénario

Distinction entre propriétés et scénarios

Cette sous-section a pour but de définir la notion de scénario et de la caractériser. Un scénario correspond à la perception d'activités accomplies par des objets mobiles à partir d'images 2D. La notion de scénario se différencie de celle de propriété selon trois points de vue. Premièrement, les scénarios et les propriétés n'ont pas la même granularité temporelle. Les propriétés sont calculées sur de très courts intervalles de temps et sont considérées comme instantanées. Leur utilisation a pour objectif d'obtenir des informations suffisamment fiables. Les scénarios sont définis sur des intervalles dont la durée n'est pas réductible à un instant et pouvant être aussi grande que souhaitée. Les propriétés sont ainsi des briques de base servant à construire les scénarios. Deuxièmement, les propriétés sont moins dépendantes d'une application que les scénarios. Cette distinction scénario-propriété facilite le développement de nouvelles applications. Effectivement, les propriétés sont définies de façon à pouvoir les réutiliser pour la reconnaissance de nouvelles activités, tandis que les scénarios sont définis dans le but de correspondre à la diversité du monde réel et sont ainsi dédiés à la reconnaissance d'activités spécifiques. Troisièmement, un scénario manipule essentiellement des informations abstraites tandis qu'une propriété réalise le passage d'informations numériques à des informations symboliques.

Définition récursive des scénarios

On définit les scénarios récursivement comme une combinaison de sous-scénarios. Au niveau zéro, un scénario est défini à partir des propriétés des objets mobiles. Contrairement aux hiérarchies statiques, la définition récursive des scénarios permet de déterminer et d'ajuster plus facilement le niveau d'abstraction des scénarios à de nouvelles spécifications. Par exemple Nagel

dans (Nagel, 1991), souligne le problème des étapes optionnelles dans le scénario, *« se garer »*. Lorsqu'un individu cherche à garer sa voiture, il peut échouer à vouloir se garer dans une première place de stationnement trop étroite et donc chercher ensuite une autre place. Dans le cas d'une hiérarchie statique, Nagel est obligé de définir un niveau supplémentaire dans la hiérarchie pour prendre en compte l'activité, *« se garer sur un emplacement donné »*, par rapport à l'activité plus générale, *« se garer »*. L'utilisation d'une définition récursive permet alors de ne pas figer le nombre de niveaux utilisés.

Deux types de combinaison

Un scénario a ainsi comme principal objectif de permettre l'adaptation à la diversité des activités rencontrées dans des applications réelles. Pour réaliser cette caractéristique, il nous est apparu nécessaire de distinguer deux types de combinaison en sous-scénarios : selon que la combinaison représente une contrainte temporelle ou atemporelle. Par exemple, *« l'individu 1 suit l'individu 2 »* correspond à une combinaison **atemporelle** de sous-scénarios. Pour que ce scénario soit reconnu, il est nécessaire que les deux individus vérifient trois propriétés. Il faut qu'ils aient une vitesse et une trajectoire similaires et qu'ils soient à une distance respectable l'un de l'autre. Ces propriétés, correspondant à des scénarios de niveau de récursion zéro, restent invariantes dans le temps. Par opposition, le scénario *« roder dans un parking »* combine une séquence **temporelle** de sous-scénarios. Ce scénario, dans l'application cible de surveillance de parkings, signifie que l'individu se dirige vers une voiture, reste un certain laps de temps près de cette voiture, puis change de direction et se dirige vers une autre voiture. Ce scénario est alors reconnu lorsque l'on a successivement reconnu les trois sous-scénarios, correspondant aux étapes de la reconnaissance du scénario. On distingue ces deux types de combinaison de sous-scénarios car la nature de leur reconnaissance est intrinsèquement différente. Dans le cas d'une combinaison atemporelle un scénario peut être reconnu rapidement, même avec un faible degré de vraisemblance, tandis que dans le cas d'une combinaison temporelle il est nécessaire d'attendre la fin de la séquence de sous-scénarios.

6.3 Nature des propriétés et scénarios utilisés

Cette section discute la nature des principaux scénarios et propriétés que nous avons utilisés pour analyser des activités impliquant des objets

mobiles, par rapport aux différentes approches présentées dans l'état de l'art à la section 6.1.

Ces scénarios et propriétés concernent essentiellement des propriétés et de scénarios temporels et spatiaux, car ils sont les plus utiles pour traiter les applications cibles choisies, décrites dans le chapitre 1. Cependant, dans le cas où la qualité des images serait meilleure (caméra plus proche des objets mobiles, bonne résolution des images), des propriétés et scénarios relatifs à l'aspect des objets mobiles, comme leur forme et leur couleur, seraient intéressantes pour analyser plus finement les activités impliquant les objets mobiles. Pour l'instant en ce qui concerne l'aspect des objets mobiles, nous nous sommes limités aux propriétés et scénarios relatifs à leurs dimensions (c.-à-d. hauteur et largeur), mais les méthodes de calcul et de manipulation des propriétés et des scénarios sont généralisables à des propriétés et des scénarios relatifs à l'aspect des objets mobiles plus complexes.

6.3.1 Propriétés et scénarios temporels

Cette sous-section aborde la notion de temps dans le cadre de la reconnaissance de scénarios. Nous commençons par décrire les principales informations relatives au temps que nous utilisons dans notre approche. Ensuite nous présentons les différents raisonnements temporels menés pour analyser des activités impliquant des objets mobiles, puis nous exposons la forme de raisonnement que nous avons adopté.

Informations temporelles utilisées

Nous manipulons le temps selon **deux niveaux de granularité** : sur un court intervalle de temps souvent réductible à un instant (jusqu'à 4 ou 5 images) ou sur un long intervalle de temps (supérieur à 4 images). À chacun de ces niveaux, nous utilisons la notion de temps soit pour conserver les informations appartenant au passé, soit pour prédire le futur :

- Les raisonnements réalisés sur un court intervalle de temps concernent le suivi des régions mobiles et le calcul des propriétés de base. Les informations appartenant au passé sont conservées grâce à la trajectoire et les informations sur le futur consistent à prédire la nouvelle position des régions mobiles et parfois leur mouvement. La vitesse, l'évolution de la couleur et les déformations des contours sont des exemples de propriétés de base évoluant dans le temps.

– Les raisonnements temporels menés sur des intervalles de temps non réductibles à des instants, concernent la reconnaissances de scénarios. Les informations passées et futures relatives à un scénario, sont respectivement représentées par les états antérieurs et postérieurs à l'état courant du scénario. Les informations passées sont utilisées pour déterminer l'état courant et participent directement au processus de reconnaissance de scénarios. Prédire une action dans le futur lointain consiste à déterminer les états postérieurs éloignés de l'état courant.

Tout développement de module de reconnaissance de scénarios pose les questions suivantes : à quel niveau de granularité temporelle doit-on raisonner (sur des instants ou sur des intervalles)? Dans quels buts utilise-t'on une logique temporelle? Quelles conditions faut-il remplir pour utiliser une telle logique? Cette sous-section se propose de répondre en partie à ces questions.

Raisons d'utilisation d'une logique temporelle

On utilise une logique temporelle sophistiquée pour réaliser trois objectifs principaux : raisonner sur le futur, déterminer les intentions des objets mobiles et vérifier la cohérence du système d'interprétation. En effet, l'utilisation d'une logique temporelle permet d'obtenir une phase de prédiction efficace. La prédiction est un raisonnement de déduction : la connaissance des états passés et présents permet, à l'aide de plans prédéfinis, de déduire les états futurs possibles. Après une phase de prédiction, il est alors possible d'ordonnancer les ensembles d'états possibles et de raisonner sur le futur. Par exemple, ceci permet de prévoir des situations particulières comme des collisions entre différents objets mobiles (Tsang and Howarth, 1991).

Deuxièmement, la phase de prédiction permet d'analyser les intentions des objets mobiles. Un problème classique en reconnaissance d'actions consiste à savoir si l'on cherche à déterminer les causes responsables des actions des objets mobiles, ou s'il est possible de déterminer leurs intentions. Cette opposition entre intention et cause, est à rapprocher de la différence entre le type d'approche envisagé utilisant l'apparence des actions ou utilisant des modèles physiques régissant les mouvements des objets mobiles (se référer au chapitre 5.2). Qu'il s'agisse d'intention ou de cause, l'objectif de la reconnaissance d'actions n'a pas pour but de déterminer les raisons profondes qui motivent les objets mobiles, mais seulement d'avoir des indices sur ce qu'ils comptent entreprendre. Par exemple dans (Sol, 1997), l'auteur analyse des séquences d'un match de football et essaie de prédire l'intention des joueurs, comme *« passer la balle »* ou *« marquer un but »*, en étudiant l'évolution

de la disposition des joueurs dans le temps. Bien qu'étant l'objectif final de nombreuses applications d'analyse de comportements, cet objectif d'analyse des intentions est très ambitieux et n'est pas encore vraiment traité. En effet avant de pouvoir atteindre cet objectif, il est nécessaire en premier lieu de connaître le passé et le présent avec suffisamment de certitude.

Troisièmement, l'étape de prédiction permet de gérer la cohérence du système en comparant l'état présent et l'état prédit au niveau des états antérieurs. Par exemple dans VIEWS (Duong et al., 1990b), les auteurs gèrent la cohérence tout au long du processus d'analyse : au niveau du suivi des régions mobiles, de l'identification des objets mobiles et de la reconnaissance des scénarios. Ils effectuent alors de nombreux retours en arrière et remises en cause d'analyses antérieures. Dans le système proposé dans ce mémoire, lorsque nous devons faire un choix relatif à une analyse de comportements ambiguës (p. ex. choix d'une méthode de reconnaissance), le principe utilisé est de geler l'analyse en cours ou de poursuivre simultanément les différentes options, si la confiance dans les propriétés calculées ne permet pas d'effectuer ce choix avec suffisamment de certitude. Ce principe n'impliquant pas de retours en arrière, permet d'éviter une certaine perte de temps dans le traitement. Nous préférons limiter la comparaison entre l'état prédit et l'état présent à son utilisation pour la correction du degré de vraisemblance des actions reconnues. Si ces deux états ne se correspondent pas, nous diminuons le degré de vraisemblance des actions reconnues à l'état présent. À part cette correction du degré de vraisemblance, nous n'avons pas rencontré de situations qui nécessitaient l'utilisation d'une logique temporelle. L'intérêt d'utiliser une logique temporelle est ainsi directement lié aux objectifs des applications abordées.

En ce qui concerne le choix de la logique temporelle à adopter, il n'existe pas de contraintes fortes relatives à la reconnaissance de scénarios. Dans ce type de processus, les notions temporelles utilisées étant généralement simples, les logiques temporelles classiques peuvent être utilisées. Une étude de ces logiques temporelles est présentée dans (Chleq, 1995).

Conditions d'utilisation d'une logique temporelle

La principale condition d'utilisation d'une logique temporelle, est la nécessité que les propriétés temporelles soient obtenues avec suffisamment de confiance. En particulier, il est possible d'utiliser une logique temporelle, lorsque l'on peut raisonner directement sur des instants temporels. Le choix d'une granularité temporelle dépend alors du type de l'application envisagée. Premièrement, la notion d'instant temporel dépend de la fréquence du

nombre d'images par seconde. Par exemple, un instant temporel mesuré dans une séquence d'images à 4 images par seconde peut devenir un intervalle temporel dans une séquence d'images à 25 images par seconde. Deuxièmement, le choix d'une granularité dépend de la précision des mesures effectuées sur les objets mobiles et par conséquent du type de capteur utilisé. Par exemple, si ces mesures sont discrètes (c.-à-d. vraies ou fausses en un point du temps et à une position donnée), alors il est possible d'utiliser des raisonnements temporels sophistiqués. Par exemple dans (Dousson et al., 1993), les auteurs utilisent des barrières optiques comme capteurs, permettant ainsi de déterminer précisément la position de robots mobiles se déplaçant dans une usine. En propageant des contraintes temporelles sur leurs positions, les auteurs analysent ainsi les différentes activités des robots. De même dans (Tsang and Howarth, 1991), les auteurs utilisent une logique temporelle à travers un TMM (*« Time Map Manager »*, en anglais) pour contrôler le trafic d'avions au sol dans les aéroports, afin d'éviter les collisions. L'emploi d'une logique temporelle sophistiquée est possible dans ce type d'application, car les objets mobiles sont correctement détectés et la précision requise est peu importante. Par exemple dans ces derniers travaux, les auteurs utilisent dans leur raisonnement temporel des propriétés relatives à la détermination de la piste où est détecté un avion donné.

Il existe différentes situations où une logique temporelle est difficile à mettre en œuvre. C'est principalement le cas lorsque l'on essaye d'analyser des activités humaines plus complexes avec de mauvaises conditions de détection. Il est alors difficile de déterminer le début et la fin d'une propriété et de connaître son degré de vraisemblance. De même, une activité est dans ce cas difficilement décomposable en une séquence d'événements. Le calcul de propriétés symboliques à partir de mesures effectuées sur de courts intervalles de temps permet alors d'utiliser ces propriétés comme si elles étaient vraies à un instant donné. Pour cela dans notre approche, nous considérons qu'une propriété est vraie dès que son degré de vraisemblance est supérieur à un seuil. Ces propriétés seraient alors manipulables par des logiques temporelles et permettraient de raisonner à la fois sur des instants et sur des intervalles. Malheureusement en pratique ces propriétés sont peu fiables en ce qui concerne nos applications cibles. La possibilité d'utiliser une logique temporelle est ainsi directement liée aux conditions d'utilisation des applications abordées.

Raisonnement temporel utilisé

Comme indiqué dans l'état de l'art (section 6.1), il existe en effet certains systèmes de reconnaissance d'actions utilisant une logique temporelle dans des applications comparables à nos applications cibles. Par exemple dans (Pinhanez and Bobick, 1997), les auteurs utilisent une logique dérivée de l'algèbre d'intervalles d'Allen pour reconnaître l'action, *« prendre le bol »*. Ils ont alors besoin d'exprimer des contraintes temporelles non élémentaires, comme *« l'intervalle de temps de validité de la propriété 1 chevauche celui de la propriété 2 »*, pour des raisons de granularité temporelle. Effectivement, la granularité des états décrivant l'action étant importante, il est nécessaire de préciser à l'intérieur même d'un état, comment les propriétés se combinent. Il est alors possible de se passer de ces contraintes temporelles non élémentaires, en considérant une granularité des états plus fine. Même dans ces travaux comparables aux nôtres, l'utilisation d'une logique temporelle est pratique, mais n'est pas nécessaire puisqu'un changement de granularité temporelle permet de réaliser un raisonnement temporel comparable.

Pour cette raison nous ne nous sommes intéressés qu'à des opérations simples de manipulation du temps. Nous calculons le temps que dure une action et nous exprimons les contraintes sur l'enchaînement des actions. Par exemple, nous indiquons qu'une action 1 doit commencer après qu'une action 2 soit finie. Le fait que nous puissions nous contenter d'opérations simples de manipulation du temps pour décrire des scénarios réels, est permis par la possibilité de représenter de deux façons distinctes les deux niveaux de granularité temporelle. En effet, nous représentons la notion d'instant à l'aide de propriétés et la notion d'intervalle à l'aide de scénarios. Cette double représentation permet de s'adapter à la diversité du monde réel. L'intérêt de notre approche est alors de pouvoir distinguer ces deux granularités temporelles et de pouvoir les intégrer dans le même formalisme, afin de traiter des applications réelles. Les opérations de base et l'utilisation de différentes granularités temporelles nous suffisent pour mener des raisonnements temporels dans le cadre de nos applications cibles.

Cependant comme le montrent les sous-sections précédentes, l'intérêt et la possibilité d'utiliser une logique temporelle sont directement liés aux objectifs et aux conditions d'utilisation des applications abordées. L'utilisation d'une logique temporelle fait alors partie de nos perspectives de recherche.

6.3.2 Notions spatiales

Propriétés et scénarios spatiaux

Un problème classique pour tout système d'interprétation consiste à décider si les propriétés spatiales 2D sont suffisantes pour mener à terme le raisonnement spatial ou s'il est nécessaire de considérer les propriétés 3D. Si on prend exemple sur un être humain, il utilise le plus souvent des relations topologiques grossières, comme l'appartenance d'un objet mobile à une zone particulière (p. ex. *« l'individu est sur le banc »*), plutôt que des propriétés métriques précises. Les propriétés 2D et le contexte de la scène permettent en général de calculer ces relations.

Cependant, les propriétés 3D sont nécessaires dans certains cas. En particulier, elles sont plus précises et permettent de calculer les relations spatiales avec plus de confiance. Par exemple, si la profondeur de la position des objets mobiles dans la scène peut varier avec une grande amplitude, alors les propriétés 3D sont nécessaires pour calculer la hauteur des objets mobiles. Les propriétés 3D permettent également de passer d'une vue 2D de la scène à une représentation absolue de l'espace (indépendant de l'angle de vue de la caméra). En général, on visualise cette représentation absolue à l'aide d'une vue de dessus de la scène, permettant de montrer les scénarios reconnus avec moins d'ambiguïté. Cette vue de dessus est ainsi utilisée pour montrer de façon conviviale les performances du système d'interprétation lors de démonstrations. La représentation 3D de la scène est également nécessaire dans le cas où le système d'interprétation utilise plusieurs capteurs. Cette représentation sert alors de support pour fusionner les informations en provenance des différents capteurs.

Au niveau des scénarios, la notion d'espace intervient uniquement à travers l'utilisation des propriétés spatiales, puisque l'on ne considère généralement qu'une seule granularité spatiale. Pour certaines applications prenant en compte une grande étendue spatiale, on peut cependant être amené à considérer plusieurs granularités. Par exemple, pour un système gérant un réseau de caméras sur l'étendue globale du métro, on peut avoir à combiner des scénarios se déroulant dans différentes stations. Ce problème n'a pas été abordé dans le cadre de nos travaux.

En ce qui concerne le système présenté dans ce mémoire, nous avions commencé à utiliser une méthode de calibration de la scène 2D, permettant de transformer des coordonnées 2D en coordonnées 3D et inversement. Cependant, nous nous sommes limités au calcul des propriétés 2D. En effet, la

mise en œuvre de cette méthode de calibration n'est pas justifiée par le gain du raisonnement dans un espace 3D. De plus, la qualité de l'acquisition des séquences d'images ne permet pas de calculer avec suffisamment de précision les propriétés 3D. Les principales propriétés spatiales que nous utilisons sont la position des objets mobiles, leurs dimensions (c.-à-d. hauteur et largeur) et la distance les séparant d'un objet d'intérêt. Pour un groupe d'objets mobiles, nous calculons en plus le dispersement des objets au sein du groupe. L'utilisation effective des propriétés 3D des objets mobiles fait partie de nos perspectives de recherche.

Propriétés et scénarios spatio-temporels

Une fois les propriétés temporelles et spatiales calculées, nous en déduisons les propriétés spatio-temporelles des objets mobiles. Pour cela, nous utilisons leur trajectoire et les scénarios qui leur sont associés, comme support afin de relier les propriétés temporelles aux propriétés spatiales. Ne pas calculer directement les propriétés spatio-temporelles permet d'adapter le calcul de ces propriétés au contexte de la scène. Par exemple, si un objet mobile a été perdu pendant un temps important, alors nous avons besoin de calculer une vitesse plus globale, prenant en compte les déplacements de l'objet mobile sur une grande échelle de temps. Les principales propriétés spatio-temporelles que nous calculons sont la trajectoire des objets mobiles, leur vitesse, l'évolution de leurs dimensions et l'évolution de la distance les séparant d'un objet d'intérêt. Elles correspondent aux évolutions des propriétés spatiales choisies. Pour un groupe d'objets mobiles, nous calculons en plus l'évolution du dispersement des objets au sein du groupe.

Les propriétés et scénarios spatiaux, temporels et spatio-temporels sont ainsi caractérisés par leur valeur et leur degré d'incertitude, qui sont calculés par des fonctions spécifiques. Ces fonctions permettent d'appréhender la diversité du monde réel, mais en contre-partie nécessitent un développement particulier. Cette contrainte rend difficile la réutilisation d'anciennes applications en vue du développement de nouvelles. Ce problème est discuté à la fin de ce chapitre.

6.3.3 Contexte

Dans la communauté de reconnaissance de scénarios, l'utilisation du contexte fait l'unanimité. Par exemple dans (Sablayrolles, 1995), l'auteur explique que c'est une des trois sources essentielles d'informations, utilisées

pour décrire en langage naturel une action accomplie par un objet mobile. De même dans (Howarth, 1994), l'auteur explique l'importance du contexte dans la reconnaissance de scénarios. Cependant, comme ce processus est dépendant de l'application, il est difficile de décrire de façon générale les utilisations du contexte. L'information contextuelle correspond en général à une information supplémentaire, comparable parfois à un oracle, permettant de reconnaître un scénario.

En ce qui concerne le système décrit dans ce mémoire, nous utilisons les informations contextuelles de la même façon que les propriétés caractérisant les objets mobiles. Nous définissons alors pour ces informations une valeur et un degré de vraisemblance. La différence avec les propriétés est l'appartenance des informations contextuelles à la base de contexte et le fait que ces informations soient en général générées de façon *a priori*. Dans l'application cible de surveillance de parking, le contexte est déterminant pour gérer par exemple, la disparition des individus. Si un individu disparaît derrière une voiture pendant un long moment, l'individu est alors considéré comme intéressant par le système, c.-à-d. suspect pour un opérateur humain. Parcontre s'il disparaît derrière un arbre, aucune alarme n'est déclenchée.

Le point dur et limitatif de l'utilisation du contexte réside dans l'automatisation de son acquisition. Ce problème est discuté dans le chapitre 2.

6.4 Réalisation du module de reconnaissance

6.4.1 Modèle de scénario

D'après la définition de la sous-section 6.2.2, un scénario est défini récursivement comme une combinaison de sous-scénarios. Comme le montre la figure 6.1, un scénario est représenté à l'aide d'un modèle composé de sept attributs : le nom du scénario, les objets mobiles concernés et leur rôles, le type de combinaison, les sous-scénarios utilisés, la valeur de reconnaissance du scénario, un ensemble de méthodes pour calculer cette valeur, le degré de vraisemblance du scénario et un ensemble de méthodes pour calculer ce degré de vraisemblance.

Le modèle de scénario est construit sur le même principe que le modèle de propriété. La valeur de reconnaissance quantifie le degré de reconnaissance du scénario relativement à une séquence d'images donnée. Si cette valeur est suffisamment élevée, alors le scénario est considéré comme reconnu. Le degré de vraisemblance du scénario indique si sa valeur de reconnaissance est une donnée fiable. Si ce degré est suffisamment élevé, alors la valeur de reconnaissance peut être utilisée. Selon le type de combinaison des sous-scénarios,

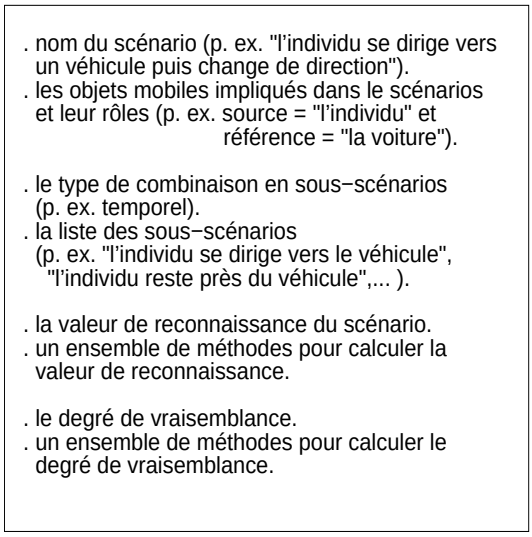

FIG. 6.1 – *Le modèle de scénario*

nous calculons de deux façons différentes la valeur de reconnaissance et le degré de vraisemblance du scénario.

6.4.2 Scénarios atemporels

Pour un scénario atemporel, la combinaison de ses sous-scénarios représente une contrainte portant sur les valeurs de reconnaissance des sous-scénarios. La valeur de reconnaissance du scénario atemporel quantifie alors la vérification de cette contrainte. La méthode de calcul de cette vérification est spécifique au scénario. Par exemple, la reconnaissance du scénario *« l'individu 1 suit l'individu 2 »* consiste à vérifier que les deux individus ont une trajectoire et une vitesse similaires et à vérifier également que les individus sont à une distance respectable l'un derrière l'autre. La méthode de reconnaissance appelle alors trois fonctions spécifiques pour calculer les mesures évaluant la similarité des trajectoires et des vitesses, ainsi que la distance entre les deux individus. De façon plus générale, ces fonctions spécifiques dépendent des propriétés concernées (p. ex. la trajectoire) et des opérations manipulant ces propriétés (p. ex. test d'égalité). La valeur de reconnaissance du scénario est alors obtenue en combinant ces trois mesures. Pour éviter une trop grande fluctuation de cette valeur, nous utilisons la moyenne de la valeur de reconnaissance calculée sur un court intervalle de temps. Pour calculer le degré de vraisemblance du scénario, nous procédons de la même manière que

pour les propriétés : nous utilisons une phase de diagnostic abductif. Dans le cas des scénarios, les phénomènes extérieurs pouvant interférer dans le calcul de la valeur de reconnaissance sont dus principalement à la confusion entre des comportements proches. Les symptômes utilisés pour détecter ces phénomènes extérieurs sont essentiellement des informations contextuelles. Par exemple, pour le scénario *« l'individu s'accroupit »*, un phénomène extérieur est le scénario concurrent *« l'individu s'asseoit »*. Un symptôme évitant de confondre ces deux comportements est la présence d'un siège à proximité de l'individu. Cette information appartient au contexte de la scène (se référer à la section 2.3).

Nous utilisons plus particulièrement deux scénarios atemporels que nous avons défini par défaut dans le module de reconnaissance de scénarios : *« l'objet mobile se comporte comme un individu »* et *« l'objet mobile se comporte comme un bruit »*. Nous caractérisons le comportement d'un individu par la régularité de ses propriétés (c.-à-d. leurs évolutions progressives et continues) et par son ratio *« hauteur / largeur »*. Pour caractériser un bruit, nous utilisons différents types d'irrégularité. Par exemple, l'alternance de la disparition et de la réapparition au même endroit d'une région mobile est typique d'un bruit tel qu'un reflet.

6.4.3 Scénarios temporels

Les scénarios temporels sont reconnus à l'aide d'un automate d'états finis. Cet automate est constitué d'un état d'initialisation et d'un ensemble d'états correspondant aux sous-scénarios à reconnaître. La valeur de reconnaissance du scénario est l'état courant de l'automate. Le scénario est reconnu lorsque tous ses sous-scénarios sont successivement reconnus. Lorsque l'on se trouve dans un état 1 correspondant à un sous-scénario 1, trois situations peuvent se produire :

- si le sous-scénario 1 est suffisamment reconnu (sa valeur et son degré de vraisemblance sont supérieurs à des seuils) et si le sous-scénario 2 commence à être reconnu, alors on passe à l'état 2 qui devient le nouvel état courant.

- Si le sous-scénario 1 n'est plus suffisamment reconnu et que le sous-scénario 2 n'a pas commencé à être reconnu, alors on retourne à l'état d'initialisation qui devient l'état courant.

– Si le sous-scénario 1 reste suffisamment reconnu, mais que le sous-scénario 2 n'a pas encore commencé à être reconnu, alors on reste dans l'état 1.

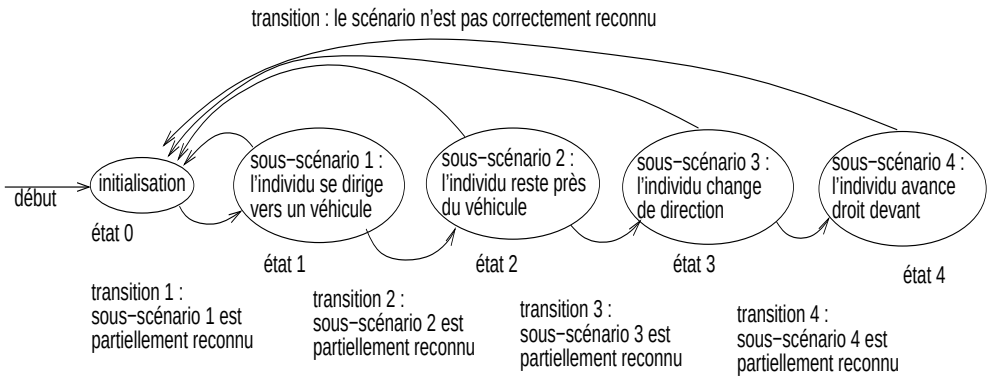

FIG. 6.2 – *Cette figure montre l'automate servant à reconnaître le scénario « l'individu se dirige vers un véhicule puis change de direction » avec ses cinq états de reconnaissance. Si la reconnaissance d'un des sous-scénarios est un échec, la reconnaissance du scénario principal retourne à l'état d'initialisation.*

Au niveau d'un état, on utilise alors trois seuils. Le premier indique quand la reconnaissance du scénario reste à l'état courant (le sous-scénario courant est suffisamment reconnu), le deuxième quand la reconnaissance retourne à l'état d'initialisation (le sous-scénario courant n'est pas suffisamment reconnu) et le troisième indique quand la reconnaissance passe à l'état suivant. Ces seuils sont spécifiques à un état donné et nécessitent d'être ajustés pendant une phase d'apprentissage, pour reconnaître efficacement le scénario étudié. Dans le but de faciliter la phase d'implantation du système, nous nous sommes limités ici à l'utilisation d'automates élémentaires. Dans ce type d'automate, soit la reconnaissance du scénario est un échec et l'état courant redevient l'état initial, soit le scénario continue à être reconnu et l'état courant passe à l'état suivant. Ces automates comportent alors plusieurs états, mais un seul enchaînement principal des états. Parmi nos perspectives, nous prévoyons d'utiliser des automates plus complexes, possédant plusieurs états suivants possibles.

Le degré de vraisemblance du scénario est obtenu en faisant la somme pondérée des degrés de vraisemblance des sous-scénarios déjà reconnus. Les degrés de vraisemblance sont ainsi propagés des scénarios les plus élémen-

taires vers les scénarios les plus abstraits. La somme des degrés des sous-scénarios est normalisée afin que tous les degrés de vraisemblance du système soient comparables entre eux. Par exemple, comme le montre la figure 6.2, l'automate servant à reconnaître le scénario *« l'individu se dirige vers un véhicule puis change de direction »* (ce scénario est utilisé pour reconnaître le cas d'un individu rodant près de véhicules), est constitué d'un état d'initialisation et de quatre autres états correspondant à quatre sous-scénarios. Ces sous-scénarios sont des scénarios élémentaires, correspondant à des propriétés caractérisant l'individu, et définies au chapitre 5. Le scénario principal est reconnu lorsque tous les sous-scénarios ont été reconnus successivement et si le degré de vraisemblance est suffisamment important. Si le premier sous-scénario est reconnu avec un degré de vraisemblance de 0.8 et que son coefficient de pondération vaut 0.25, alors le degré du scénario principal devient : 0.8 * 0.25 = 0.2, les autres sous-scénarios ayant un degré de vraisemblance nul.

6.4.4 Discussion

Les scénarios atemporels ont été jusqu'à présent très peu traités, tandis que différentes approches ont abordé le problème des scénarios temporels. Ce paragraphe compare alors l'approche proposée de reconnaissance de scénarios, aux approches présentées dans l'état de l'art dans la section 6.1, et permet de justifier l'intérêt de l'utilisation d'automates.

- Premièrement, la notion d'automate est naturelle pour décrire des scénarios, puisque l'usage d'automates est très répandu pour analyser grammaticalement des phrases en langage naturel décrivant des scènes.

- De plus, les automates sont à la base de la plupart des méthodes de reconnaissance de scénarios temporels. Par exemple H. Nagel dans (Nagel, 1991), utilise en fait un automate, qu'il appelle diagramme de transition. Les chemins de parcours de ce diagramme correspondent à des automates élémentaires, comparables à ceux que nous utilisons dans notre approche. La méthode de H. Nagel est sur ce point similaire à la nôtre.

- D'autres systèmes, comme dans (Castel et al., 1996), utilisent des réseaux de Pétri. Cette utilisation se caractérise par la séparation de la gestion de l'enchaînement des états à reconnaître, de la gestion de contraintes non temporelles associées aux objets mobiles. Cependant, nous pensons qu'il est plus simple d'utiliser le même formalisme pour

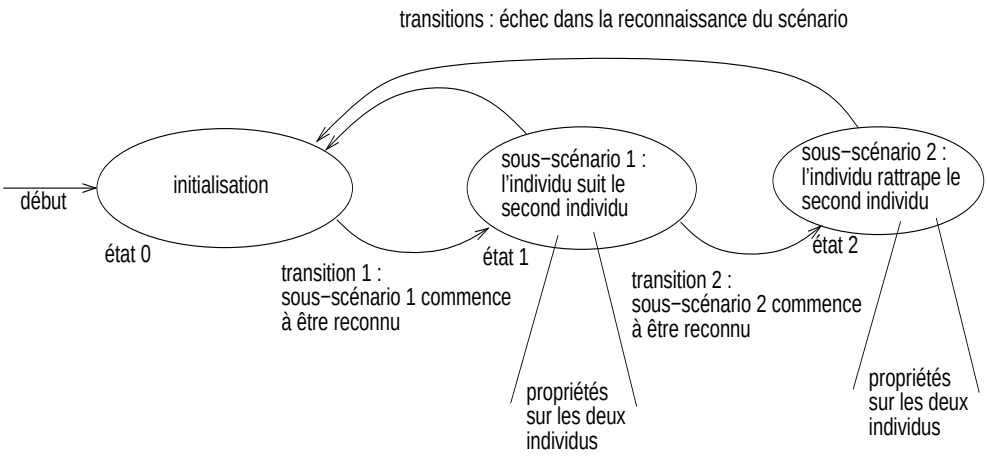

FIG. 6.3 – *L'automate du scénario « l'individu suit un second individu, puis le rattrape », est constitué des états correspondant à la reconnaissance de deux sous-scénarios atemporels.*

représenter les contraintes temporelles et atemporelles, à l'aide de la notion de scénario. Ce formalisme commun permet plus de souplesse pour combiner les deux types de contrainte.

– Les modèles de Markov cachés (appelés HMM et décrits dans la section 5.1) utilisent également un automate pour reconnaître des scénarios temporels. Ces scénarios sont en général de faible granularité temporelle et n'ayant pas de définition récursive, ils ne permettent pas la reconnaissance de scénarios plus abstraits. A. Pentland, conscient de ce problème concernant son application de reconnaissance du langage de signes, utilise un automate englobant ces HMM, afin d'analyser grammaticalement une longue séquence de gestes. Un second inconvénient des HMM est qu'ils ne permettent pas d'établir le degré de vraisemblance de la reconnaissance des scénarios. En effet, les HMM calculent la probabilité qu'un scénario soit reconnu, mélangeant les notions de valeur de reconnaissance et de degré de vraisemblance. Par exemple dans (Starner and Pentland, 1995), ils permettent de donner la probabilité qu'une séquence d'images donnée corresponde à un des signes du langage des sourds-muets. Un troisième inconvénient réside dans la génération des distributions de probabilités. On génère ces distributions à l'aide d'une phase d'apprentissage, nécessitant un nombre conséquent d'exemples de séquences d'images, illustrant le scénario à reconnaître.

Cependant, il est souvent difficile d'obtenir des séquences d'images, qui soient des exemples de scénarios intéressants (c.-à-d. correspondant à des comportements anormaux). Par exemple, nous avons très peu de scènes filmées d'agression dans les métros et leur simulation demande trop de moyens à mettre en œuvre. Ce problème est d'autant plus vrai que la granularité temporelle de ces scénarios est importante. Il est alors plus facile de caractériser ces scénarios en utilisant des propriétés symboliques invariantes, plutôt que d'essayer de calculer des distributions de probabilité les décrivant. Un automate manipulant des propriétés symboliques est ainsi mieux adapté qu'un automate probabiliste, tel que les HMM.

– Enfin, le modèle que nous proposons se singularise par sa définition récursive en sous-scénarios. Cette définition récursive permet la combinaison de différents types de scénario et d'ajuster de façon souple la durée des scénarios à de nouvelles spécifications. Par exemple, pour l'application de surveillance du métro, nous avons besoin d'utiliser le scénario temporel *« l'individu suit un second individu, puis le rattrape »*, défini à partir de deux sous-scénarios atemporels : *« l'individu suit un second individu »* et *« l'individu rattrape un second individu »*. Ces deux sous-scénarios sont eux-mêmes reconnus à partir des propriétés caractérisant les deux individus, définies au chapitre 5. La figure 6.3 montre l'automate de reconnaissance du scénario temporel, combinant la reconnaissance des deux sous-scénarios.

Malgré ces points positifs, les méthodes de reconnaissance de scénarios, comprenant les automates et les diagnostics ont comme inconvénient majeur la difficulté de leur mise en œuvre. Par exemple, la phase de diagnostic n'est pas utile lorsqu'aucun phénomène extérieur n'interfère dans la reconnaissance du scénario. Dans ce cas-là, nous n'utilisons pas de phase de diagnostic et nous calculons le degré de vraisemblance du scénario, directement à partir de celui de ses sous-scénarios. C'est le cas en général des scénarios temporels, où nous nous limitons à propager les degrés de vraisemblance. La difficulté de mise en œuvre de ces méthodes est également due à la dépendance des scénarios au domaine d'application. En tant que travaux futurs, nous prévoyons alors d'automatiser le développement de ces méthodes à partir de descriptions génériques des scénarios. Ce point est discuté ci-après, dans la section 6.5.

6.4.5 Algorithme du processus de reconnaissance

Cette sous-section a pour but de décrire l'algorithme du processus de reconnaissance des scénarios, et d'expliquer comment le contrôle de ce processus est réalisé, afin de permettre un traitement en temps réel des séquences d'images.

Le module de reconnaissance a pour but de reconnaître les scénarios relatifs aux objets mobiles de la scène. Il opère selon quatre étapes principales :

- Tout d'abord, le module de reconnaissance considère toute région mobile suivie, et tout groupe de régions mobiles proches les unes des autres, comme constituant des objets mobiles. Il calcule ensuite toutes les propriétés des objets mobiles, ainsi que leur degré de vraisemblance. Ce module analyse également un ensemble de scénarios de base, relativement à tout objet mobile. Ces scénarios étant assez génériques, ils sont prédéfinis dans le système d'interprétation. Par exemple le scénario *« l'individu change de direction »*, est analysé pour tout objet mobile.

- Une fois la première étape terminée, le module de reconnaissance calcule le degré d'intérêt de tous les objets mobiles. Ce degré d'intérêt tient compte de la durée de vie des objets mobiles, ainsi que du nombre et du degré de vraisemblance des scénarios impliquant les objets mobiles. Dans ce calcul, nous donnons un poids plus important aux scénarios les plus intéressants, comprenant en particulier les scénarios les plus abstraits.

- Pour un petit nombre d'objets, considérés comme les objets les plus intéressants, le module de reconnaissance analyse des scénarios supplémentaires, qui sont en général plus spécifiques et qui possèdent une granularité temporelle plus importante. Par exemple pour ces objets, ce module essayera de reconnaître le scénario *« l'individu se dirige vers une voiture puis change de direction »*; ce scénario servant à détecter les individus rodant dans un parking.

- Enfin, le module de reconnaissance déclenche une alarme lorsqu'un scénario intéressant est reconnu avec suffisamment de vraisemblance. Les scénarios intéressants sont ceux spécifiés par l'utilisateur. Le scénario permettant de détecter qu'un individu rode dans un parking est un exemple de scénario intéressant.

Cette décomposition en étapes du processus de reconnaissance a pour objectif d'obtenir un traitement en temps réel. La notion de processus temps réel signifie dans ce mémoire, le traitement de la séquence d'images à la volée. Elle signifie également l'adaptation du traitement au temps imparti. Notre objectif estalors la conception d'un traitement minimal, et si le temps le permet, la possibilité d'analyser plus finement les comportements des objets mobiles. Ces spécificités sont toutes imposées par le type d'application cible envisagé, décrit dans la section 1.1.

Dans ce but, nous avons défini deux niveaux de traitement, l'un minimal et l'autre supplémentaire. Le niveau **minimal** traite les images dans leur ensemble pour détecter et suivre toutes les régions mobiles et pour calculer les propriétés nécessaires des objets mobiles, afin ensuite d'analyser leurs comportements si besoin. Ce niveau correspond aux deux premières étapes de l'algorithme de reconnaissance. Le niveau **supplémentaire** sélectionne l'attention du système sur les objets mobiles les plus intéressants et sur leurs comportements. Il réalise des traitements supplémentaires, spécialement pour ces objets mobiles, afin d'affiner l'analyse de leurs comportements. Ce niveau correspond aux deux dernières étapes de l'algorithme de reconnaissance.

Pour déterminer le temps alloué à chacun de ces niveaux, il est nécessaire de choisir entre l'exhaustivité et la précision de l'interprétation : si on attache plus d'importance à traiter un nombre important d'objets mobiles, ou si on préfère analyser plus finement les comportements des objets mobiles les plus intéressants. Ce choix revient à déterminer le nombre d'objets mobiles d'intérêt pour lesquels on compte poursuivre la reconnaissance de scénarios et la fréquence de ces reconnaissances. Ce nombre dépend également de la fréquence des images et du temps de traitement des différentes tâches du système d'interprétation.

6.5 Conclusion

6.5.1 Améliorations à court terme

Notre objectif est de développer un système d'interprétation pouvant s'adapter et traiter différentes applications de vidéo-surveillance. Un problème essentiel réside alors dans la génération semi-automatique des méthodes de reconnaissance de scénarios, comprenant les diagnostics abductifs et les automates de reconnaissance. Le mécanisme pour générer un automate de reconnaissance est un problème classique. D'abord on construit les états et les transitions de l'automate, à partir de la liste de sous-scénarios.

Ensuite, il s'agit de déterminer les seuils décrits à la section 6.4, indiquant dans quelles situations la reconnaissance du scénario reste dans l'état courant de l'automate, ou passe dans un état suivant. Le problème de génération d'un automate réside ainsi dans l'ajustement des seuils de reconnaissance, problème qui peut être résolu en pré-traitement lors de phases d'expérimentation.

Par contre, le problème de génération de diagnostics abductifs ne possède pas de solution simple. L'étape de diagnostic consiste à établir le degré de vraisemblance d'un scénario donné, sachant que plusieurs phénomènes extérieurs peuvent expliquer des situations similaires. Par exemple, considérons les deux scénarios suivants : scénario 1, *« l'individu s'asseoit »* et scénario 2, *« l'individu s'accroupit »*. Il existe alors deux cas, selon lesquels la reconnaissance du scénario 1 diffère. Dans le cas où seul le scénario 1 peut être reconnu, il suffit pour reconnaître ce scénario de vérifier seulement que la taille correspondant à la détection de l'individu diminue. Dans le cas où le scénario 2 peut également se produire, alors la méthode de reconnaissance du scénario 1 doit être modifiée, pour tenir compte de la possibilité d'occurence du scénario 2. Il est alors nécessaire de vérifier par exemple la présence d'un siège, pour reconnaître effectivement le scénario 1. Pour concevoir une méthode de reconnaissance d'un scénario, il est nécessaire de connaître la description du scénario courant, mais également de tenir compte de la description de tous les phénomènes extérieurs pouvant interférer dans la reconnaissance du scénario. Il n'est pas alors possible de prédéfinir une bibliothèque de diagnostics, dû à la nécessité de connaître par avance toutes les descriptions des scénarios et phénomènes extérieurs impliqués dans la scène traitée, donc dépendants de l'application. Pour cette raison, nous pensons que le meilleur mécanisme pour générer les méthodes de reconnaissance de scénarios, et plus particulièrement leur phase de diagnostic, est d'utiliser des descriptions de scénarios prédéfinies et génériques, ces descriptions étant contenues dans une bibliothèque associée au module de reconnaissance de scénarios. Ce mécanisme consiste alors à sélectionner l'ensemble des scénarios possibles *a priori*, puis à générer chaque méthode de reconnaissance en tenant compte de tout cet ensemble de scénarios.

Cependant, avant de développer ce mécanisme et d'automatiser le développement de différentes applications, nous pensons qu'il est nécessaire d'avoir un système d'interprétation robuste et ayant été testé pour un grand nombre de scénarios. Pour cette raison, nous n'avons pas pour le moment généré automatiquement des méthodes de reconnaissance de scénarios à partir de descriptions prédéfinies de scénarios. Ces améliorations constituent un problème difficile et font parties des axes futurs de recherche.

6.5.2 Contributions et perspectives

Ce chapitre propose un **formalisme** ayant pour objectif la reconnaissance de scénarios, qui en calculant les propriétés des objets mobiles, permet à l'aide de scénarios de déterminer leurs comportements et activités. Ce formalisme possède quatre caractéristiques le différenciant des approches déjà proposées dans le domaine :

- Ce formalisme sépare le calcul des propriétés de la reconnaissance des scénarios, permettant la séparation des entités dépendantes du domaine d'application (les scénarios), des entités indépendantes (les propriétés).

- Il décrit les scénarios à l'aide de référents sémantiques, permettant ainsi d'utiliser une description des activés des objets mobiles dans un langage proche du langage naturel. Dans un scénario, un objet mobile peut ainsi se voir attribuer différents rôles. Selon la nature de ces rôles, nous avons construit six types de scénarios.

- Il propose un modèle de scénario. Un scénario est alors défini par la combinaison de sous-scénarios et de propriétés relatifs aux objets mobiles intervenant dans le scénario. Nous avons utilisé deux types de combinaisons (atemporelle et temporelle), afin d'ajuster au mieux le processus de reconnaissance à la diversité du monde réel. De plus, le formalisme de reconnaissance définit récursivement les scénarios en sous-scénarios pour attribuer de façon précise le niveau d'abstraction du scénario.

- Il reconnaît les scénarios temporels à l'aide d'automates, comme le font la plupart des approches de reconnaissance de scénarios; mais de plus il reconnaît les scénarios atemporels à l'aide de fonctions de vérification de contraintes sur leurs sous-scénarios.

- Il traite uniformément les propriétés et les scénarios en leur attribuant un degré de vraisemblance pour gérer leur incertitude de façon homogène à l'aide d'une phase de diagnostic.

Le formalisme de reconnaissance permet ainsi de mener un raisonnement symbolique au niveau des scénarios relatif aux activités des objets mobiles, tout en tenant compte de l'incertitude et de l'évolution irrégulière des propriétés perçues.

Ces travaux ont conduit au développement et à l'intégration au système d'interprétation d'un **module de reconnaissance de scénarios** (écrit en

langage C++). Des exemples d'utilisation de ce module sont donnés ci-après dans le chapitre 7.

Un inconvénient majeur de ce formalisme est le manque de généricité des méthodes de reconnaissance des scénarios, limitant la réutilisation du système d'interprétation et le développement de nouvelles applications. Un axe de recherche consiste à développer les améliorations présentées dans la sous-section précédente, afin de générer automatiquement les méthodes de reconnaissance. Ces améliorations proposent d'adapter les méthodes de reconnaissance au contexte de l'application. Un prolongement de ces améliorations consiste à utiliser des techniques d'apprentissage.

Chapitre 7

Système d'interprétation

Ce chapitre étudie les différents problèmes liés à l'architecture et à l'implantation du système d'interprétation de séquences d'images. Ce chapitre prolonge l'analyse du problème réalisée au chapitre 1. Les problèmes liés à l'architecture sont multiples dus tout d'abord, à la taille importante des systèmes d'interprétation. Ces systèmes doivent gérer de nombreuses et diverses fonctionnalités appartenant à plusieurs domaines. Deuxièmement, une telle architecture doit vérifier les spécificités contraignantes décrites dans le chapitre 1, telles qu'être générique et favoriser un traitement en temps réel.

Nous commençons ce chapitre en proposant une architecture pour le système d'interprétation développé dans le cadre de cette thèse et nous décrivons précisément ses caractéristiques. Ensuite, nous décrivons dans la section 7.2 deux exemples d'utilisation de ce système d'interprétation. Enfin, la section 7.3 conclue ce chapitre en décrivant les performances et les limites du système.

7.1 Architecture du système proposé

7.1.1 Description de l'architecture

Comme le montre la figure 7.1 et comme nous l'avons précédemment indiqué, l'architecture proposée est structurée selon l'enchaînement classique des tâches de détection, de suivi et de reconnaissance des scénarios. Cependant, cette architecture diffère en plusieurs points par rapport aux différentes approches présentées dans l'état de l'art du chapitre 1 :

– Nous donnons un rôle central à la base de contexte, qui sert ainsi de support de raisonnement tout au long du processus d'interpréta-

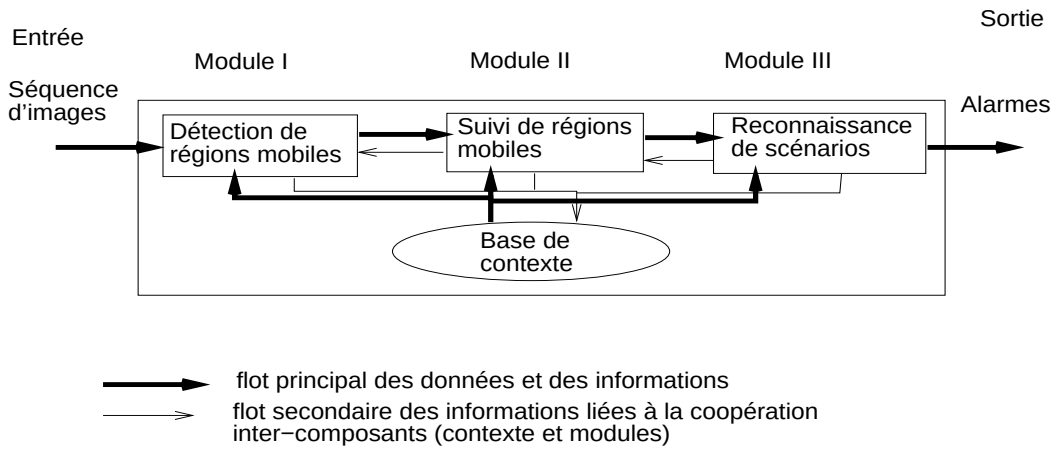

FIG. 7.1 – *Le système d'interprétation de séquences d'images est constitué d'une base de contexte et de trois modules.*

tion. Inversement, si les informations contextuelles sont nombreuses, le nombre de connaissances *a priori* utilisées est faible. Par exemple, le modèle *a priori* d'individu que nous utilisons est élémentaire. Il se limite à caractériser la valeur du ratio hauteur sur largeur de l'individu. Cependant comme expliqué au chapitre 6, nous comptons utiliser des bibliothèques contenant par exemple, des descriptions *a priori* et génériques de scénarios pouvant se produire dans les applications considérées.

– Nous n'avons pas de tâche spécifique pour identifier les objets mobiles. Les données d'entrée étant imprécises et incertaines, il est difficile de reconnaître les objets de la scène à partir des régions mobiles. Par exemple sur certaines séquences bruitées de parking, il est difficile de différencier un groupe d'individus, d'un individu seul poussant un caddie en utilisant des méthodes telles que des méthodes de classification. Nous préférons alors identifier les objets de la scène en nous basant sur l'analyse de leurs comportements. Par exemple, si on reconnaît pour un objet mobile donné le comportement *« tourner autour d'un véhicule »*, et que cet objet vérifie le modèle élémentaire *a priori* caractérisant un individu, alors nous en déduisons que cet objet mobile est probablement un individu.

– Cette architecture possède un module de suivi à part entière. Ce point est particulièrement important, car le problème de suivi des régions

mobiles reste un des problèmes clés du processus d'interprétation. Nous distinguons ce module du reste du système, afin de lui donner toute son importance. Le module de suivi n'est alors ni un module de vision bas niveau, ni un module d'interprétation. Par contre, occupant une position centrale, il peut coopérer et bénéficier des résultats obtenus par les modules de détection et de reconnaissance de scénarios.

- Le module de reconnaissance de scénarios est constitué d'un seul bloc, malgré la distinction que nous faisons entre les propriétés des objets mobiles et les scénarios relatifs à leurs comportements (ce point est discuté au chapitre 6). En effet, bien que le modèle de ces deux notions soit différent, leur implantation est identique afin de faciliter leur mise en œuvre. Ceci permet par exemple, de gérer l'incertitude des propriétés et des scénarios de façon homogène avec le même formalisme.

7.1.2 Caractéristiques de l'architecture proposée

Dans cette sous-section, nous énumérons plusieurs caractéristiques du système proposé. Cependant notre objectif n'étant pas d'obtenir un système opérationnel à grande échelle, ces caractéristiques ne sont encore que des prémisses de fonctionnalité. Pour une réalisation plus complète, ces fonctionnalités nécessitent une étude systématique en vue du développement d'une application particulière. Nous nous sommes alors limités à implanter ces caractéristiques de base et à permettre l'extension du système vers des fonctionnalités plus abouties. Le système d'interprétation prend ainsi en compte cinq fonctionnalités :

- Le système proposé est modulaire et extensible. Il est modulaire de part sa structure. Il est extensible du fait de l'indépendance relative de chaque module. Il est alors plus facile d'actualiser un module, afin de prendre en compte de nouvelles applications. La base de contexte renforce cette utilisation modulaire du système. En effet les modules devant coopérer, nous centralisons au niveau du contexte les informations partagées par les différents modules. Par ailleurs, la base de contexte contient les informations dépendantes de l'application. De ce fait, nous pensons que le système proposé est relativement générique et qu'il permet le développement de nouvelles applications sans surcoût prohibitif.

- Ce système favorise un traitement dirigé par les buts. Le module de reconnaissance des scénarios signale les objets mobiles les plus impor-

tants, ainsi que les régions d'intérêt (se référer à la section 6.4). Ayant connaissance des objets mobiles les plus importants, le module de suivi peut alors favoriser ces objets afin de ne pas perdre leur piste. Ayant connaissance des régions d'intérêt, le module de détection n'est plus alors obligé de traiter entièrement les images. Le mode dirigé par les buts permet ainsi d'améliorer les résultats du processus global d'interprétation et de réduire le temps de traitement du système. Ce point est important pour le processus d'interprétation et a fait l'objet de travaux spécifiques de différentes équipes (Howarth and Buxton, 1993).

- Le système permet un traitement en temps réel des applications envisagées. Le temps de traitement du système est en partie optimisé, du fait de la monotonie de son raisonnement. Il n'y a pas de remise en cause (« *backtrack* » en anglais) des résultats obtenus. Au lieu de vérifier la cohérence du raisonnement par intervalles réguliers, le système calcule en continu la vraisemblance de tous les résultats obtenus. Cette gestion de l'incertitude permet de plus, de fournir un résultat exact (même s'il est peu informatif) à tout moment du déroulement du traitement. Le système peut également adapter son traitement en fonction du temps qui lui est imparti, en choisissant prioritairement les objets mobiles à traiter et les scénarios à analyser.

- Enfin, le système proposé permet aux modules de communiquer entre eux, grâce à la position centrale de la base de contexte. Cette coopération permet à chaque module de tirer avantage des résultats obtenus au cours du processus d'interprétation. Ce point est décrit plus amplement dans la section suivante.

7.1.3 Coopération inter-modules

Les relations entre les modules du système d'interprétation proposé et la base de contexte sont décrites sur la figure 7.1. Les traits en gras correspondent au flot principal du traitement des données et des informations. Premièrement ce flot correspond à la phase d'abstraction des données, en commençant à partir du traitement des images (module I) et en finissant par le déclenchement des alarmes (module III). Deuxièmement, ce flot correspond à l'utilisation par les modules des informations contenues dans la base de contexte. De façon complémentaire, nous utilisons un flot d'informations correspondant aux échanges de communication inter-modules. Ce flot est représenté sur la figure par des traits fins. Ce flot complémentaire correspond

ainsi aux :

- Échanges du module de reconnaissance vers le module de suivi. Le module de reconnaissance maintient la liste des scénarios reconnus relatifs aux objets mobiles. Premièrement, ces scénarios permettent au module de suivi de connaître les objets mobiles les plus importants. Les objets importants sont ceux qui correspondent à des objets réels de la scène (p. ex. les individus) et dont les scénarios sont considérés comme importants par l'utilisateur (p. ex. *« l'individu tourne autour d'un véhicule »*). Le module de suivi peut alors sélectionner son attention sur ces objets et s'attacher à ne pas perdre leur piste. Deuxièmement, les scénarios reconnus permettent au module de suivi de déterminer les régions mobiles correspondant à du bruit et de pouvoir arrêter de les suivre. Troisièmement, ce retour d'information permet au module de suivi d'évaluer les performances de son traitement. Par exemple, le module de suivi utilise les degrés de vraisemblance des scénarios reconnus, afin de rendre plus robuste la phase d'association des régions mobiles (se référer au chapitre 4).

- Échanges du module de suivi vers le module de détection. En connaissant les objets mobiles intéressants, le module de suivi peut signaler au module de détection la localisation des régions d'intérêt (R.O.I.) sur les prochaines images. En sélectionnant son attention sur ces régions, le module de détection évite de traiter entièrement les images.

- Échanges des modules du système d'interprétation vers la base de contexte. Ces échanges consistent essentiellement pour les modules, à maintenir à jour les listes de régions mobiles, d'objets mobiles et des scénarios relativement à leur localisation spatiale. L'indexation de ces entités sur la décomposition de l'espace permet d'utiliser la base de contexte comme support de raisonnement. Les bénéfices tirés de la centralisation de ces listes dans la base de contexte sont expliqués en détail au chapitre 2.

Dans le cas d'une utilisation systématique du contexte, l'objectif est de centraliser tous ces échanges à travers la base de contexte. Le retour d'information des modules de plus haut niveau (modules II et III), en direction des modules de plus bas niveau (modules I et II) n'est pas alors nécessaire, toutes ces informations étant contenues dans la base. Cependant, nous laissons la présence de ces liens puisque ces informations ne sont pas actuellement toutes disponibles dans la base de contexte.

Par ailleurs, la centralisation de ces échanges permet de garder un contrôle sur les tâches allouées à chaque module. Par exemple, en réduisant le nombre de régions d'intérêt (R.O.I.), nous limitons le temps de traitement du module de détection. La base de contexte reste alors un élément central permettant aux modules de coopérer.

7.2 Exemples d'utilisation

Les séquences d'images choisies comme données d'entrée du système ont été prises dans le cadre du projet européen Esprit HPCN PASSWORDS. Nous utilisons ces séquences pour illustrer et tester le fonctionnement de notre système d'interprétation.

7.2.1 Scène se déroulant sur un parking

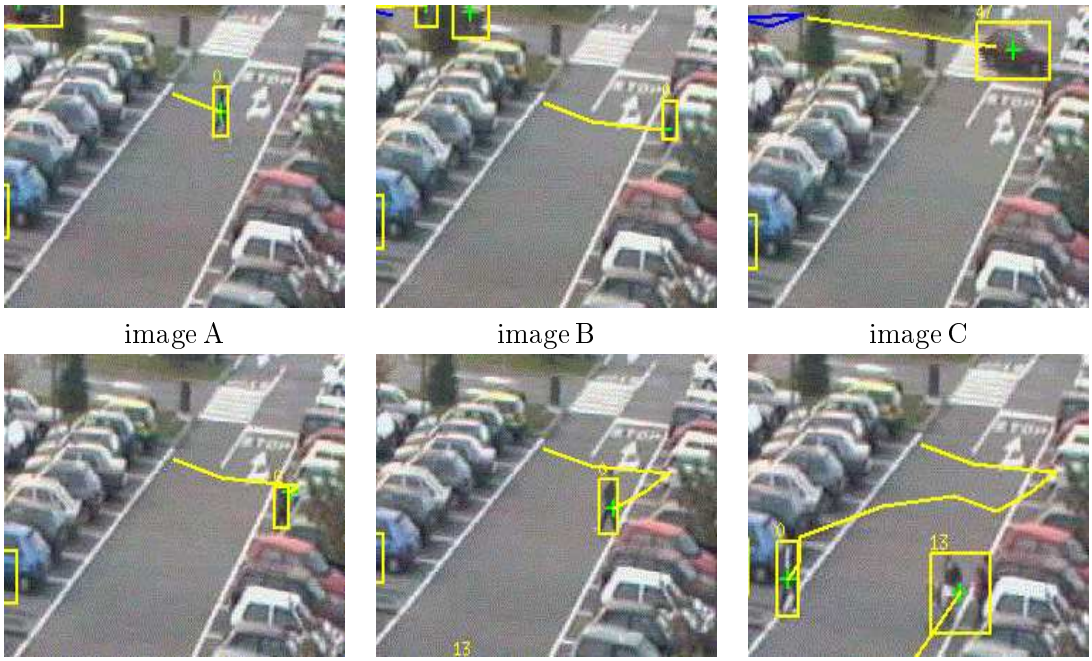

image A image B image C

image D image E image F

FIG. 7.2 – *Cette séquence d'images illustre le scénario « l'individu se dirige vers un véhicule, puis change de direction », utilisé pour détecter un individu rodant dans un parking.*

La figure 7.2 représente six échantillons d'une séquence d'images comprenant 1200 images prises à une cadence de 5 images par seconde, soit une durée totale de 4 minutes. Les images sont des images couleur 512*512, mais chaque échantillon ne correspond qu'à une portion d'image (environ 1/5 de l'image d'origine prise au centre) pour faciliter la lisibilité du document. Dans cette séquence d'images on voit un individu évoluant sur le parking d'un supermarché. Il se dirige vers une rangée de voitures, puis disparaît et change de direction lors de sa réapparition et finalement se redirige vers la rangée opposée de voitures. Cette séquence est typique d'un individu rodant dans un parking et a été réalisée par un acteur. Dans le but de reconnaître cette activité, nous avons défini le scénario principal *« l'individu se dirige vers un véhicule puis change de direction »*. Bien entendu, la reconnaissance de ce scénario n'est qu'un indice et ne permet pas de juger les intentions de l'individu. L'automate de reconnaissance de ce scénario est décrit sur la figure 6.2.

- Sur l'image A, le mouvement de l'individu est perçu à travers la détection d'une seule région mobile, dont la boîte englobante est représentée par un rectangle. La trace du suivi de l'individu représentant sa trajectoire est dessinée par des traits blancs. L'individu est correctement suivi depuis le moment où il a été détecté, correspondant au début de la séquence d'images. L'individu étant de couleur sombre sur un fond sombre (la chaussée), nous avons dû régler au préalable le seuil de contraste de l'intensité pour avoir un niveau de détection suffisant.

  Depuis que l'individu a été détecté, le système d'interprétation calcule les propriétés relatives à ses déplacements, telles que sa vitesse et sa trajectoire. L'individu étant suivi avec suffisamment de confiance (pendant un intervalle de temps suffisamment long et sans suspension de son suivi), il devient alors un objet d'intérêt pour le système d'interprétation, qui choisit l'individu comme objet source pour l'analyse de comportements. L'objectif de l'opérateur humain étant de surveiller les véhicules, les zones correspondant aux deux rangées de véhicules ont été sélectionnées à l'initialisation comme objets de référence pour l'analyse de comportements. Le système d'interprétation essaie alors de reconnaître différents scénarios, dont en particulier le scénario *« l'individu se dirige vers un véhicule puis change de direction »*, ayant pour objet source *« l'individu »* et pour objet de référence *« la rangée de voitures »* la plus à droite sur l'image.

- Sur l'image B, on peut voir que la trajectoire de l'individu est rectiligne,

montrant qu'il est en train de progresser en direction de la rangée de véhicules. Le sous-scénario 1, *« l'individu se dirige vers un véhicule »*, est alors reconnu à l'aide de la valeur de deux propriétés : *« la rectitude de la trajectoire »* de l'individu et *« la diminution de la distance séparant l'individu de l'objet de référence »*. Le calcul de la propriété *« la rectitude de la trajectoire »* est possible grâce à l'utilisation d'une méthode (décrite au chapitre 4), qui lisse la trajectoire de la région mobile associée à l'individu. Cette région étant détectée de façon irrégulière (le bas de l'individu se confond par moment avec la chaussée), la trajectoire mesurée est en réalité en dents de scie. La seconde propriété est calculée à partir de l'évolution de la distance moyenne séparant les régions représentant l'individu et l'objet de référence. Le degré de vraisemblance du sous-scénario 1 a atteint la valeur de 0.9. Le coefficient de pondération de chaque sous-scénario (défini à la section 6.4) étant ici de 0.25, le degré de vraisemblance du scénario principal devient *0.9 * 0.25 = 0.23*.

- Sur l'image C, l'individu a atteint l'objet de référence (la rangée de voitures), mais n'est plus visible. Il est occulté par les véhicules et son suivi est suspendu pendant une cinquantaine d'images. Le système d'interprétation considère en première approximation que l'individu reste à proximité des véhicules, tant qu'aucune autre information ne contredit cette hypothèse. Le sous-scénario 2 *« l'individu reste près d'un véhicule »* est ainsi reconnu avec un degré de vraisemblance suffisant, et l'automate passe de l'état 1 à l'état 2. Pour évaluer le degré de vraisemblance du sous-scénario 2, le système d'interprétation diagnostique si cette reconnaissance n'est pas due à deux comportements concurrents *« l'individu démarre et part avec un véhicule »* et *« l'individu évolue occulté par les véhicules »*. Aucun symptôme correspondant à ces comportements n'étant observé (p. ex. départ d'une voiture ou réapparition de l'individu de l'autre côté de la rangée de voitures), le degré de vraisemblance du sous-scénario 2 est suffisamment élevé. Le degré de ce sous-scénario ayant atteint la valeur de 0.6, le degré de vraisemblance du scénario principal devient *0.23 + 0.6 * 0.25 = 0.38*.

- Sur l'image D, l'individu réapparaît à proximité de l'endroit de sa disparition. Le degré de vraisemblance du sous-scénario 2 augmente et sa valeur de reconnaissance reste valide. L'automate reste alors dans l'état 2. Le degré de ce sous-scénario atteignant maintenant la valeur de 0.8, le degré de vraisemblance du scénario principal devient *0.23 +*

*0.8 * 0.25 = 0.43.*

- Sur l'image E, on voit l'individu changer de direction et se diriger vers la rangée opposée de véhicules. Le sous-scénario 2 est toujours reconnu mais avec une valeur de reconnaissance plus faible, car l'individu s'éloigne. Par contre, le sous-scénario 3 *« l'individu change de direction »* devient valide avec un degré de vraisemblance suffisamment élevé. L'automate passe alors dans l'état 3. Le degré du sous-scénario 3 ayant atteint la valeur de 0.7, le degré de vraisemblance du scénario principal devient *0.43 + 0.7 * 0.25 = 0.6.*

- Sur l'image F, l'individu maintient sa direction vers l'autre rangée de véhicules. À ce moment-là, les propriétés relatives à l'individu et à l'objet de référence correspondent au modèle du sous-scénario 4 *« l'individu se dirige droit devant »* : la trajectoire de l'individu est rectiligne et la distance séparant les deux objets augmente. Le sous-scénario 4 est alors reconnu avec un degré de vraisemblance élevé. L'automate passe dans l'état 4. Le degré de ce sous-scénario ayant atteint la valeur de 0.9, le degré de vraisemblance du scénario principal devient *0.6 + 0.9 * 0.25 = 0.83.* Ce degré étant suffisamment élevé, le scénario principal est également reconnu. Le système d'interprétation déclenche alors une alarme (c.-à-d. il envoie un message sur la console de l'opérateur humain), puisque ce scénario a été spécifié à l'initialisation comme intéressant pour l'application.

L'exemple de la séquence d'images de la figure 7.2 montre que le module de reconnaissance est effectivement capable de reconnaître le scénario *« l'individu se dirige vers un véhicule puis change de direction »*, avec un bon degré de vraisemblance. Ce scénario utilise des propriétés invariantes, telles que *« la rectitude de la trajectoire »*, permettant de décrire ce type d'activité. Il est ainsi conçu pour prendre en compte toutes les séquences illustrant cette activité. Dans ce but, les seuils de reconnaissance des sous-scénarios et de transitions entre les états de l'automate sont fixés avec une marge d'erreur suffisante, pour prendre en compte la diversité des séquences d'images illustrant l'activité en question. Cependant, ce scénario n'a pas été testé sur d'autres séquences d'images étant donné les difficultés pour obtenir des exemples de séquences illustrant cette activité. Pour valider pleinement ce scénario, il serait nécessaire de le tester sur au moins une dizaine de séquences illustrant l'activité dans diverses situations. Néanmoins, comme le scénario ne manipule que des informations symboliques donc variant peu, nous jugeons ce résultat comme satisfaisant.

7.2.2 Scène se déroulant dans un métro

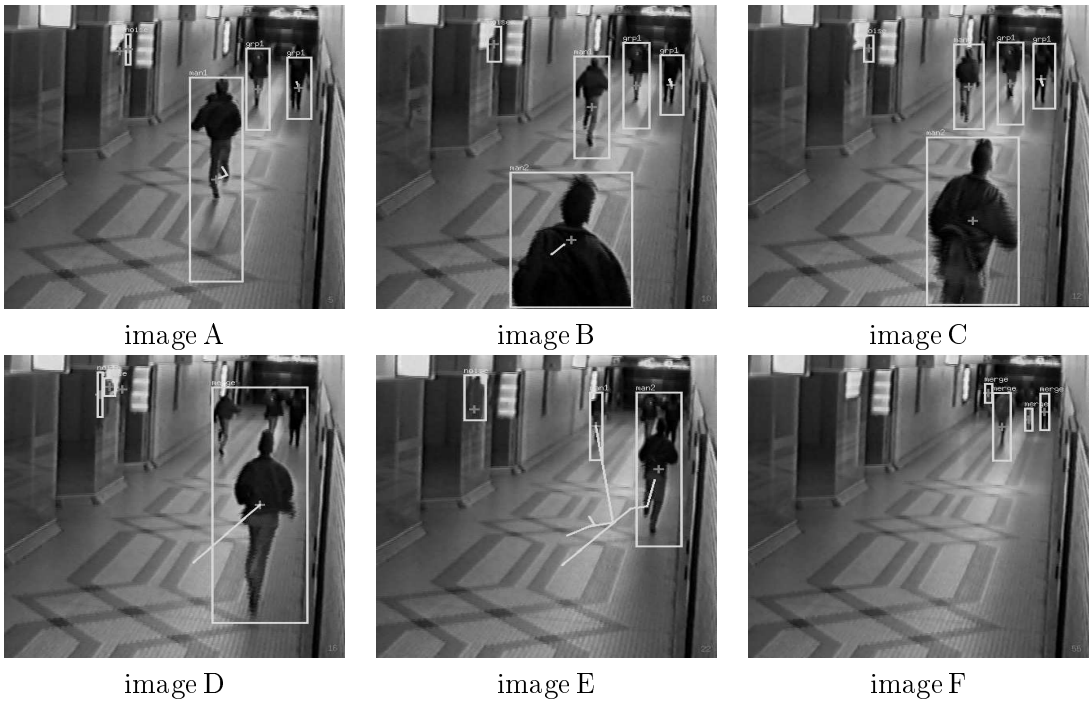

image A image B image C

image D image E image F

FIG. 7.3 – *Cette séquence d'images illustre le scénario « l'individu 1 suit l'individu 2, puis le rattrape », utilisé pour détecter des situations dangereuses dans un métro.*

La figure 7.3 représente six échantillons d'une séquence d'images comprenant 1000 images couleur 512*512, prises à une cadence de 5 images par seconde, soit une durée totale de 3 minutes 20. Dans cette séquence d'images, on peut voir deux individus courir l'un derrière l'autre (le second rattrape le premier individu) dans le couloir d'un métro. Notre objectif est de reconnaître le scénario *« l'individu suit un second individu puis le rattrape »*. Cet objectif est particulièrement difficile en raison de la présence de nombreux reflets sur le carrelage du métro, de la position basse de la caméra et de la présence d'un groupe de deux passants évoluant également dans le couloir. Un premier problème est alors de suivre tous ces objets mobiles jusqu'à ce qu'ils sortent de la scène. Un second problème consiste à reconnaître le précédent scénario relativement aux deux individus en train de courir. L'automate

de reconnaissance de ce scénario est décrit sur la figure 6.3 :

- Au centre de l'image A, le premier individu appelé individu 1, entre dans la scène tout en courant. Dès qu'il est détecté, le système d'interprétation calcule les propriétés de base relativement à son comportement, telles que *« sa vitesse »*. On peut noter également sur cette image, que l'individu 1 est mal détecté ; la région mobile correspondant à sa détection fusionne avec celle d'un reflet localisé à ses pieds. De plus, on peut voir deux passants appelés groupe de passants, progressant vers le fond de la scène.

- Sur l'image B, un second individu appelé individu 2, entre dans la scène en courant derrière l'individu 1. Comme l'individu 1 a été correctement suivi depuis le début, la propriété *« l'individu 1 a une vitesse élevée »* est reconnue. Cette propriété correspondant à un individu en train de courir fait partie des propriétés spécifiées à l'initialisation comme intéressantes par l'opérateur humain. Le système d'interprétation considère alors l'individu 1 comme un objet d'intérêt et le choisit comme objet de référence dans l'analyse de comportements.

- Sur l'image C, l'individu 2 est toujours correctement suivi. La propriété *« l'individu 2 a une vitesse élevée »* est valide (la vitesse de l'individu 2 est supérieure à un seuil) et son degré de vraisemblance est suffisamment élevé : cette propriété est donc reconnue. L'individu 2 devient un second objet d'intérêt pour le système d'interprétation. Il est choisi comme objet source pour l'analyse de comportements. Ensuite, le système d'interprétation commence à analyser différents scénarios en considérant l'individu 2 comme objet source et l'individu 1 comme objet de référence. En particulier, il analyse le scénario principal *« l'individu 2 suit l'individu 1 puis le rattrape »* ainsi que le sous-scénario 1 *« l'individu 2 suit l'individu 1 »*. À cet instant-là, les propriétés relatives à l'individu 1 et l'individu 2 correspondent au modèle du sous-scénario 1 : ils ont une vitesse et trajectoire similaires et ils sont à une distance respectable l'un de l'autre. Ces propriétés ayant des degrés de vraisemblance suffisants, le sous-scénario 1 est alors reconnu. Par ailleurs, le système d'interprétation initialise l'automate du scénario principal. Le sous-scénario 1 étant reconnu, l'automate passe directement à l'état 1.

- Sur l'image D, les régions mobiles correspondant aux détections de l'individu 1, de l'individu 2 et du groupe de passants se mélangent et leurs

pistes respectives sont temporairement suspendues. Le module de suivi utilise alors une cible composée pour suivre l'ensemble des pistes mélangées et attendre que les individus soient de nouveau correctement détectés (se référer au chapitre 4). Cependant le système d'interprétation continue toujours à analyser les scénarios et les sous-scénarios relatifs aux individus. À ce niveau-là, les degrés de vraisemblance du scénario principal et du sous-scénario 1 diminuent mais restent suffisamment élevés. L'automate de reconnaissance reste à l'état 1.

– Sur l'image E, l'individu 1 est de nouveau correctement détecté. Comme sa vitesse et sa trajectoire étaient précédemment calculées avec suffisamment de précision, le module de suivi arrive à rattacher la région mobile correspondant à cette nouvelle détection à l'ancienne piste de l'individu 1. Sa trajectoire est ainsi correctement prolongée de la position d'entrée de l'individu 1 dans la scène jusqu'à sa position courante. Par contre la région mobile correspondant à la détection de l'individu 2 est toujours superposée à celle du groupe de passants. En première approximation, les propriétés relatives à l'individu 2 sont calculées à partir de cette région mobile correspondant à la détection de l'ensemble constitué de l'individu 2 et du groupe de passants. Par chance, cette région mobile correspond assez bien à la détection de l'individu 2. À ce moment-là, les propriétés relatives à l'individu 1 et l'individu 2 correspondent au modèle du sous-scénario 2 *« l'individu 2 rattrape l'individu 1 »* : ils ont une vitesse et trajectoire similaires et sont proches l'un de l'autre. La valeur de la reconnaissance du sous-scénario 1 diminue tandis que celle du sous-scénario 2 augmente. Le degré de vraisemblance du sous-scénario 2 étant suffisamment élevé, l'automate du scénario principal passe alors à l'état 2. Néanmoins, le degré de vraisemblance du scénario principal n'est pas suffisamment élevé pour que ce scénario soit reconnu.

– Sur l'image F, tous les individus sont enfin de nouveau correctement détectés. Cependant, les pistes des individus ayant été perdues sur un intervalle de temps trop important, le module de suivi est incapable de faire les correspondances entre les régions mobiles nouvellement détectées et les anciennes pistes. Néanmoins en première approximation, le système d'interprétation estime la position des individus 1 et 2 au niveau de ces régions mobiles nouvellement détectées. Le sous-scénario 2 reste alors reconnu mais pas avec un degré de vraisemblance suffisant pour que le scénario principal soit reconnu. L'automate de reconnais-

sance reste ainsi à l'état 2 et aucune alarme n'est générée. Par la suite, les protagonistes sortent de la scène sans que le scénario principal soit reconnu.

L'exemple de cette séquence d'images montre les limites du système d'interprétation. Premièrement, le module de suivi a perdu à plusieurs reprises la piste des individus qu'il n'a pas pu réparer par la suite. Ces pertes de suivi sont dues à la présence de trop nombreux reflets et à plusieurs occultations dynamiques sur des intervalles de temps importants. Deuxièmement, le module de reconnaissance de scénarios est bloqué dès que les performances du module de suivi ne sont pas satisfaisantes. Pour continuer la reconnaissance du scénario principal *« l'individu 1 suit l'individu 2 puis le rattrape »* malgré les erreurs du module de suivi, nous avons été obligés de renforcer artificiellement les degrés d'importance des individus 1 et 2. Normalement, la piste de ces individus étant perdue, leur analyse aurait dû être suspendue. Néanmoins, afin de tester le module de reconnaissance, nous avons ajouté une règle à l'initialisation empêchant que leur degré d'importance puisse diminuer. Cet exemple montre ainsi les difficultés à mener à terme la reconnaissance de scénarios complexes, concernant une application aux conditions d'utilisation contraignantes.

Cet exemple montre également l'intérêt de continuer l'analyse des comportements relatifs à des objets mobiles, dont le suivi a été temporairement perdu. Nous avons traité une telle situation à l'occasion de cette séquence d'images, en essayant d'estimer la position des objets mobiles dont on avait perdu la piste. Cependant nous n'avons pas conçu de méthode générique permettant de reprendre la reconnaissance d'un scénario quelconque, relatif à des objets mobiles dont on aurait récupéré le suivi. Ces travaux font parties de nos perspectives de recherche.

7.3 Conclusion et performances obtenues

Dans ce chapitre, nous avons proposé une architecture dédiée à l'interprétation de séquences d'images. Cette architecture a comme principales caractéristiques :

- Elle est modulaire. Elle est constituée de quatre composants principaux (trois modules indépendants et une base de contexte) et donne une place centrale à la base de contexte.

- L'architecture proposée permet un mode de traitement dirigé par les buts. Le système d'interprétation peut ainsi sélectionner son attention sur des tâches spécifiques.
- Elle favorise la coopération entre les modules et l'utilisation systématique de la base de contexte par ces modules.
- Elle favorise un traitement en temps réel du processus d'interprétation.

Ces travaux ont abouti au développement d'un système d'interprétation opérationnel. Nous avons développé (en langage C++) un module de suivi de régions mobiles et un module de reconnaissance de scénarios conformément aux modèles proposés dans ce mémoire. Nous avons intégré au système d'interprétation un module de détection de régions mobiles (écrit en langage C), développé dans le cadre du projet européen Esprit PASSWORDS. Nous avons développé un logiciel graphique permettant l'acquisition et la construction de la base de contexte selon le modèle proposé dans le chapitre 2. La base de contexte est ainsi implantée sous la forme d'un fichier texte. Cependant, le chargement automatique de ce fichier par le système d'interprétation reste à être implanté. Dans l'état actuel du système, le chargement du contexte est réalisé manuellement à l'initialisation.

Le traitement total d'une image 512x512 couleur prend 2,5 secondes sur un Sun Sparc 10. Ce temps est essentiellement dû au traitement du module de détection de régions mobiles, le temps de traitement du reste du système étant en moyenne de 0,2 seconde. La cadence d'acquisition des images étant de 4 à 5 images par seconde, ce système ne réalise pas encore un traitement en temps réel. Cependant, nous n'avons pas cherché à optimiser le système d'interprétation pendant sa phase d'implantation.

Nous avons testé notre système à l'aide des séquences d'images prises dans le cadre du projet européen Esprit PASSWORDS. Deux séquences de scènes de métro, cinq de parking et une de supermarché ont servi ainsi d'exemple de test. Ces séquences comprennent entre 200 et 1200 images entrelacées, durant ainsi de 40 secondes à 4 minutes. Sur ces séquences nous détectons de un à quinze objets mobiles. La durée moyenne du suivi d'un objet mobile est d'une cinquantaine d'images, c.-à-d. 12 secondes. Nous avons pu résoudre une dizaine de situations d'occultation statiques et dynamiques (partielles et totales).

Nous avons pu reconnaître automatiquement des scénarios non élémentaires et générer ainsi des alarmes, sur seulement trois séquences d'images (deux de parking et une de métro). Les scénarios reconnus sont : *« l'individu*

avance puis rebrousse chemin dans le parking », *« l'individu 1 court derrière l'individu 2 dans le couloir du métro et le rattrape »* et *« l'individu se dirige vers un véhicule puis après l'avoir atteint se dirige vers d'autres véhicules »*. Ce dernier scénario sert d'indice afin de détecter un individu rodant dans un parking. Les autres séquences d'images n'ont pas abouti à la reconnaissance de scénarios non élémentaires pour l'une des deux raisons suivantes :

- Le manque relatif de robustesse du système. Les séquences d'images sont prises dans des conditions réelles (p. ex. mauvaise résolution de la caméra, présence de nombreux reflets) et ont de ce fait généré un trop grand nombre d'erreurs de détection. Le système d'interprétation n'a pas alors pu suivre les objets mobiles sur les séquences complètes. Il eut été possible de reconnaître des scénarios non élémentaires sur certaines de ces séquences, en modifiant le système d'interprétation spécifiquement pour ces séquences. Cependant, ce travail est coûteux en temps de développement et n'est pas l'objectif principal de nos travaux.

- La plupart des séquences d'images que nous avons à notre disposition ne possèdent que des comportements élémentaires (p. ex. *« l'individu descend sur la voie du métro »* et *« l'individu s'arrête longtemps devant un rayonnage de supermarché »*). Ces séquences d'images ne peuvent pas alors servir d'exemple test pour le système d'interprétation.

Dans un premier temps, nous comptons tester notre système d'interprétation de façon systématique afin de valider toutes ses fonctionnalités, en particulier celles concernant la reconnaissance de scénarios non élémentaires. Dans un deuxième temps, nous envisageons de terminer le développement des fonctionnalités décrites dans la sous-section 7.1.2. De cette manière, nous comptons obtenir un système suffisamment générique (permettant de s'adapter à une nouvelle application avec une phase minime de réglages) et opérant à cadence réelle d'utilisation.

Conclusion

Contributions

Nous avons exposé dans ce mémoire nos contributions au processus d'interprétation de séquences temporelles d'images. Ces contributions peuvent se regrouper en cinq points concernant le processus global d'interprétation et ses différents composants :

- **Modèle du processus d'interprétation** : nous avons proposé un modèle composé de trois tâches principales : (1) la détection des régions mobiles, (2) le suivi des régions mobiles et (3) la reconnaissance des scénarios. Ce modèle se caractérise également par l'utilisation systématique d'informations contextuelles, par la gestion de l'incertitude tout au long du traitement et par la coopération des tâches du processus d'interprétation.

- **Contexte** : nous avons proposé un formalisme général permettant de définir et de délimiter les informations contextuelles. Nous avons développé une représentation du contexte permettant son utilisation rationnelle et systématique. La réalisation de cette représentation a nécessité la définition d'une vingtaine d'attributs, afin de prendre en compte la diversité des éléments du contexte utilisés dans le cas du processus d'interprétation. Nous avons également développé un logiciel graphique (écrit en langage C et à l'aide de la boîte à outils MOTIF) permettant d'acquérir et de construire une base de contexte.

- **Suivi de régions mobiles** : nous avons conçu un algorithme de suivi dont l'objectif est de suivre des objets mobiles à partir des régions correspondant à la perception de leurs mouvements. Cet algorithme permet en particulier de suivre des objets non rigides tels que des individus. Il peut gérer différents types de problèmes (p. ex. des erreurs de

détection et des occultations) et temporiser la résolution des ambiguïtés d'association entre des régions mobiles déjà suivies et des régions nouvellement détectées. Nous avons également implanté un module de suivi de régions mobiles (écrit en langage C++) correspondant à l'algorithme proposé.

- **Passage du numérique au symbolique** : nous avons conçu un modèle de propriété caractérisant de manière symbolique un objet mobile à partir des mesures numériques effectuées sur les régions de l'image, correspondant à la perception des mouvements de l'objet. L'implantation de ce modèle a nécessité la définition d'une quinzaine de propriétés que nous avons organisées en réseaux. Nous avons également développé un modèle de méthodes de diagnostic abductif permettant de calculer et de mettre à jour l'incertitude de ces propriétés tout au long du processus d'interprétation. Ce réseau de propriétés a été intégré au module de reconnaissance de scénarios et a permis de le relier au module de suivi des régions mobiles.

- **Reconnaissance de scénarios** : nous avons développé un formalisme permettant de décrire des scénarios relatifs à des activités humaines. Nous avons également conçu un modèle de scénario afin de représenter des contraintes temporelles et atemporelles sur la perception des comportements des objets mobiles. La réalisation de ce modèle a nécessité le développement d'une part de méthodes permettant de reconnaître un scénario donné et d'autre part de méthodes évaluant l'incertitude de ces reconnaissances. En particulier, nous avons développé une méthode à base d'automates d'états finis permettant de reconnaître des scénarios temporels. Nous avons réalisé un module de reconnaissance de scénarios (écrit en langage C++) se conformant au modèle proposé.

- **Système d'interprétation** : nous avons proposé une architecture d'un système d'interprétation correspondant au modèle proposé et vérifiant les spécifications nécessaires à la réalisation d'applications d'interprétation de séquences d'images. Nous avons ainsi développé un système complet d'interprétation (écrit en langages C et C++) intégrant les modules précédemment décrits. Ce système a pour objectifs d'être modulaire, générique et de permettre un traitement proche du temps réel. Il a été testé sur plusieurs séquences d'images de scènes de métro et de parking.

Tout au long de nos travaux, nous nous sommes attachés à défendre trois aspects du processus d'interprétation. Tout d'abord, nous avons montré que

l'originalité du processus d'interprétation réside dans la nécessité d'utiliser une phase de raisonnement abductif. Deuxièmement, nous pensons que le raisonnement symbolique peut faire partie intégrante du processus d'interprétation (à l'aide en particulier du passage du numérique au symbolique), mais qu'il peut également renforcer ce processus. En effet, le raisonnement symbolique permet de raisonner sur le futur, de contrôler les modules de détection et de suivi, ainsi que d'améliorer leur robustesse. De plus, ce raisonnement rend possible l'élaboration de descriptions de scènes plus abstraites, allant jusqu'à des descriptions en langage naturel. Troisièmement, nous considérons que l'interprétation de séquences d'images constitue une nouvelle discipline. Elle se distingue, tout en tirant profit, de travaux accomplis dans des domaines divers, tels que la détermination du mouvement, la reconnaissance d'objets statiques, la description d'activités en langage naturel et la modélisation du contexte. La problématique de l'interprétation de séquences d'images n'est pas de développer de nouveaux traitements spécifiques, tels que les Modèles de Markov Cachés (HMM), mais plutôt de trouver une combinaison adéquate de traitements ayant déjà fait leurs preuves dans les domaines cités précédemment.

Perspectives

Plusieurs séries de tests ont montré cependant un relatif manque de robustesse du système d'interprétation proposé (p. ex. des pertes du suivi d'objets mobiles). Les limitations du système sont principalement dues à la qualité des séquences d'images utilisées, acquises dans des conditions réelles d'utilisation et contenant de nombreuses imperfections, telles que des reflets. Elles sont également dues à la complexité des scénarios que nous souhaitons reconnaître. Les axes futurs de recherche consistent premièrement à augmenter la robustesse du système proposé et deuxièmement à diversifier la classe d'applications cibles. Ces axes de recherche reviennent à renforcer les performances de chaque composant du système :

- **Contexte** : l'utilisation d'informations contextuelles s'avère essentielle pour nombre d'applications. Cependant l'acquisition de la base de contexte reste fastidieuse, malgré l'aide apportée par le logiciel d'acquisition. De plus, certaines applications nécessitent l'utilisation de capteurs modifiant le contexte pendant le déroulement du processus d'interprétation (p. ex. caméras mobiles et caméras pan-tilt-zoom). Nous avons ainsi prévu d'étendre la base de contexte pour permettre d'acquérir et de mettre à jour automatiquement les informations contextuelles

pendant le déroulement du processus d'interprétation. Cette automatisation peut par exemple, être réalisée à l'aide de méthodes statistiques ou de méthodes d'apprentissage symbolique.

- **Détection de mouvement** : le module de détection des régions mobiles peut évoluer du fait de l'utilisation de nouveaux capteurs, tels que des caméras infrarouges, de nouveaux algorithmes de détection, tels que la détermination du flot optique, ou de l'utilisation de plusieurs capteurs, par exemple pour avoir plusieurs angles de vue d'une même scène. Dans cette situation, ce module serait en mesure de fournir des propriétés supplémentaires sur les données d'entrée des autres modules du système d'interprétation. Un premier axe de recherche consiste alors à adapter ces autres modules pour tirer profit de l'évolution du module de détection des régions mobiles (p. ex. fusion des propriétés relatives à un même objet mobile).

  Au cours de ce mémoire, nous avons également montré l'importance de la coopération entre le module de détection et les autres modules du système d'interprétation. Un second axe de recherche est de réaliser cette coopération, à l'aide en particulier de techniques de pilotage de programmes de traitement d'images.

- **Suivi des régions mobiles** : le suivi des régions mobiles est un des points clés du processus d'interprétation. Une série de tests a montré que ce module ne permettait pas dans certaines situations, de réparer les erreurs de détection ou les pertes de suivi. Un axe de recherche consiste donc à améliorer la robustesse de l'algorithme de suivi.

  Dans certaines applications, de nombreux objets mobiles sont rigides et possèdent un modèle de leur forme permettant d'améliorer les performances de l'algorithme de suivi. Un second axe de recherche est alors d'adapter l'algorithme de suivi au type de l'objet mobile suivi.

- **Reconnaissance de scénarios** : un de nos objectifs est de pouvoir utiliser le système d'interprétation dans différentes applications avec un minimum de changement. Cependant, les méthodes de reconnaissance de scénarios comprenant en particulier les méthodes de diagnostic abductif et de reconnaissance par automates dépendent du contexte et des objectifs de l'application. Par conséquent, le changement d'application peut induire des changements importants dans les méthodes de reconnaissance de scénarios et rendre nécessaire la définition de nouvelles méthodes. Malheureusement, il est difficilement envisageable de

pouvoir disposer de bibliothèques prédéfinies de telles méthodes. Par contre, il est plus facile de construire des bibliothèques prédéfinies de descriptions génériques de scénarios. Un axe de recherche consiste alors à générer automatiquement les méthodes de reconnaissance de scénarios à partir de bibliothèques de descriptions génériques de scénarios.

Dans ce mémoire, nous avons également montré que des techniques de raisonnements symboliques telles que des logiques temporelles permettaient de reconnaître des scénarios plus complexes. Un deuxième axe de recherche est de déterminer les conditions, les limitations et les bénéfices attendus de ces techniques.

Ces axes de recherche ont pour objectifs de faciliter la mise en œuvre du système d'interprétation afin de traiter de nouvelles scènes et d'aborder de nouveaux domaines d'applications. Pour l'instant, nous avons abordé le problème d'interprétation essentiellement dans le cadre de la vidéosurveillance. À plus long terme, nous envisageons d'étendre le système proposé et de pouvoir traiter les différents types d'applications décrits dans le chapitre 1, tels que l'analyse de gestes.

Bibliography

André, E., Herzog, G., and Rist, T. (1988). On the simultaneous interpretation of real world image sequences and their natural language description: the system SOCCER. In *proc. of ECAI*, pages 449–454, Munich, Germany.

Azarbayejani, A., Waren, C., and Pentland, A. (1996). Real-time 3D tracking of the human body. In *Proc. of IMAGE'COM 96*, Bordeaux.

Bar-Shalom, Y. and Fortmann, T. (1988). *Tracking and data association.* Academic press, London.

Barron, J., Fleet, D., and Beauchemin, S. (1994). Systems and experiment. Performance of optical flow techniques. *Int'l Journal of Computer Vision*, 12:1:43–77.

Bascle, B., Bouthemy, P., Deriche, R., and Meyer, F. (1994). Tracking complex primitives in an image sequence. In *proc. of the ICCV'94, Jerusalem.*

Baumberg, A. and Hogg, D. (1995). An adaptive eigenshape model. In *proc. of the British Machine Vision Conference (BMVC)*, Birmingham.

Becker, D. and Pentland, A. (1997). Using a virtual environment to teach cancer patients, t'ai chi, relaxation and self-imagery. In *proc. of the ACM Siggraph Symposium on Interactive 3D Graphics*, Providence, RI, USA.

Beringer, A., Hölldobler, S., and Kurfess, F. (1993). Spatial reasoning and connectionist inference. In *proc. of the Int'l Joint Conf. on Artificial Intelligence (IJCAI).*

Bobick, A. and Davis, J. (1996). Real-time recognition of activity using temporal templates. In *proc. of the Workshop on Applications of Computer Vision.*

Bobick, A. and Pinharez, C. (1995). Using approximate models as source of contextual information for vision processing. In *proc. of the IEEE workshop on Context-Based Reasoning (ICCV'95)*.

Bogaert, M., Chleq, N., Cornez, P., Regazzoni, C., Teschioni, A., and Thonnat, M. (1996). The PASSWORDS project. In *proc. of the Int'l Conf. on Image Processing (ICIP)*, Lausanne (Suisse).

Bouthémy, P. (1988). Modèles et méthodes pour l'analyse du mouvement dans une séquence d'images. *Technique et Science Informatique*, 7:6:527–546.

Brand, M., Birnbaum, L., and Cooper, P. (1993). Sensible scenes: Visual understanding of complex scenes through causal analysis. *AAAI*.

Brand, M., Oliver, N., and Pentland, A. (1997). Coupled hidden markov models for complex action recognition. In *proc. of CVPR*, Puerto Rico, USA.

Brémond, F. and Thonnat, M. (1996a). A context representation for surveillance systems. In *Proc. of the Workshop on Conceptual Descriptions from Images at the European Conference on Computer Vision (ECCV)*, Cambridge.

Brémond, F. and Thonnat, M. (1996b). Interprétation de séquences d'images et incertitude. In *Proc. of the Rencontres sur la Logique Floue et ses Applications (LFA)*, Nancy.

Brémond, F. and Thonnat, M. (1997a). Analysis of human activities described by image sequences. In *Proc. of the 10th international FLAIRS Conference*, Florida.

Brémond, F. and Thonnat, M. (1997b). Issues in representing context illustrated by scene interpretation applications. In *proc. of the Int'l and Interdisciplinary Conf. on Modeling and Using Context (CONTEXT-97)*, Rio de Janeiro.

Brémond, F. and Thonnat, M. (1997c). Object tracking and scenario recognition for video-surveillance. In *the poster sessions of the 15th Int'l Joint Conference on Artificial Intelligence (IJCAI)*, Nagoya (Japan).

Brémond, F. and Thonnat, M. (1997d). Recognition of scenarios describing human activities. In *Proc. of the International Workshop on Dynamic Scene Recognition from Sensor Data*, ONERA (Toulouse).

Brémond, F. and Thonnat, M. (1997e). Tracking multiple non-rigid objects in a cluttered scene. In *proc. of the 10th Scandinavian Conference on Image Analysis (SCIA)*, Lappeenranta (Finland).

Buxton, H. and Gong, S. (1995). Visual surveillance in a dynamic and uncertain world. *Artificial Intelligence*, 78(1-2):431–459.

Campbell, L. and Bobick, A. (1995). Recognition of human body motion using phase space constraints. In *proc. of the fifth International Conference on Computer Vision*, pages 624–630, Cambridge MA.

Castel, C., Chaudron, L., and Tessier, C. (1996). What is going on? A high level interpretation of sequences of images. In *Proc. of the ECCV'96 workshop on Conceptual Descriptions from Images*, University of Cambridge.

Cayrac, D., Dubois, D., and Prade, H. (1995). Practical model-based diagnosis with qualitative possibilistic uncertainty. In *Uncertainty in Artificial Intelligence*.

Cerf, V. L. and Pintado, M. (1997). An adaptive model of camera-driven urban intersections observation. In *Proc. of the International Workshop on Dynamic Scene Recognition from Sensor Data*, Onera-Cert, Toulouse.

Charniak, E. (1991). Bayesian networks without tears. *AI magazine*, 12(4).

Charpillet, F., Haton, J., and Marquis, P. (1992). Abduction, raisonnement hypothéthique et maintien de cohérence. In *PRC-GDR IA, marseille.*

Chleq, N. (1995). *Contribution à l'étude du raisonnement temporel. Résolution avec contraintes et application à l'abduction en raisonnement temporel.* PhD thesis, l'École Nationale des Ponts et Chaussées.

Chleq, N. and Thonnat, M. (1996). Realtime image sequence interpretation for surveillance applications. In *Proc. of the IEEE International Conference on Image Processing, ICIP*, pages 801–804, Lausanne, Switzerland.

Choi, S., Seo, Y., Kim, H., and Hong, K. (1997). Where are the ball and players? Soccer game analysis with color-based tracking and image mosaik. In *ICIAP'97.* to appear.

Clement, V., Giradoun, G., Houzelle, S., and Sandakly, F. (1993). Interpretation of remotely sensed images in a context of multisensor fusion

using a multispecialist architecture. *IEEE Transaction on Geoscience and Remote Sensing*, 31(4):779–791.

Corrall, D. (1992). Deliverable 3: Visual monitoring and surveillance of wide-area outdoor scenes. Technical report, Esprit Project 2152: VIEWS.

Cox, I. and Hingorani, S. (1996). An efficient implementation of reid's Multiple Hypothesis Tracking algorithm and its evaluation for the purpose of visual tracking. In *IEEE Transactions on pattern analysis and machine intelligence*, volume 18.

Dawson-Howe, K. (1996). Active surveillance using dynamic background substraction. Technical Report TCD-CS-96-06, Dept. of Computer Science, Trinity College, Dublin, Ireland.

Dekneuvel, E., Ghallab, M., and Thibault, J. (1992). Hypotheses management for scene interpretation in a multisensory perception machine. In *Proc. of the ECAI, Vienna, Autriche*, pages 795–799.

Djian, D., Probert, P., and Rives, P. (1996). Reconnaissance de modèles géométriques simples à l'aide de réseaux bayésiens. In *Proc. of the RFIA*, volume 1, pages 396 – 404.

Dousson, C., Gabarit, P., and Ghallab, M. (1993). Situation recognition representation and algorithms. In *Proc. of the IJCAI-93, chambéry*, volume 1, pages 166–172.

Du, L., Sullivan, G., and Baker, K. (1993). Quantitative analysis of the viewpoint consistency constraint in model-based vision. In *Proc. of the International Conference on Computer Vision 93, Berlin, Germany*, pages 632–639.

Dubois, D., Lang, J., and Prade, H. (1990). Handling uncertain knowledge in an ATMS using possibilistic logic. In *Proc. of the Workshop on Truth Maintenance Systems, Springer-Verlag.*

Dubois, D. and Prade, H. (1992a). Fuzzy relation equations and abductive reasoning. Technical report, IRIT Toulouse.

Dubois, D. and Prade, H. (1992b). Possibilistic abduction. In *Proc. of the International conference IPMU, Spain*, pages 157 – 160.

Dubois, D. and Prade, H. (1993). Fuzzy sets and probability: misunderstanding, bridges and gaps. In *Proc. of the IEEE International Conference on Fuzzy Systems, San francisco*, pages 1059 – 1068.

Duong, V., Buxton, H., Howarth, R., Toal, P., Gong, S., King, S., Thoméré, J., and Hyde, J. (1990a). D203: Spatio-temporal reasoning. Technical report, Esprit Project 2152: VIEWS.

Duong, V., Howard, R., Hill, G., Toal, P., King, S., Gong, S., Thoméré, J., and Hyde, J. (1990b). D201: The representation of event, behaviour and scene. Technical report, Esprit Project 2152: VIEWS.

François, E. (1991). *Interprétation qualitative du mouvement à partir d'une séquence d'images*. PhD thesis, université de Rennes I.

Friedman, N. and Halpern, J. (1995). Plausibility measures: a user's guide. In *UAI, Quebec*.

Friedman, N. and Halpern, J. (1996). A qualitative markov assumption and its implications for belief change. http://robotics.standford.edu/users/nir, to appear.

Galton, A. (1993). Towards an integrated logic of space, time and motion. In *International Joint Conference on Artificial Intelligence (IJCAI), Chambery, France*.

Ghallab, M., Grandjean, P., Lacroix, S., and Thibault, J. (1992). Représentations et raisonnement pour une machine de perception multi-sensorielle. In *Proc. of the PRC-GDR IA, Marseille*, pages 121–167.

Gong, S. and Buxton, H. (1993). From contextual knowledge to computational constraints. In Illingworth, J., editor, *Proc. of the British Machine Vision Conference (BMVC)*, volume 1, pages 229–238.

Grandjean, P. (1991). *Perception multisensorielle et interprétation de scènes*. PhD thesis, LAAS - Université Paul Sabatier de Toulouse.

Herzog, G. (1995). From visual input to verbal output in the visual translator. Projet VITRA 124, Universität des Saarlandes, Saarbrücken, Germany.

Herzog, G., Sung, C., André, E., Enkelmann, W., Nagel, H., Rist, T., Wahlster, W., and Zimmermann, G. (1989). Incremental natural language description of dynamic imagery. Projet VITRA 58, Universität des Saarlandes, Saarbrücken, Germany.

Howarth, R. (1994). *Spatial representation, reasoning and control for a surveillance system*. PhD thesis, Queen Mary and Westfield College.

Howarth, R. (1995). Interpreting a dynamic and uncertain world: high-level vision. *Artificial Intelligence*, 9(1):37–63.

Howarth, R. and Buxton, H. (1992a). Analogical representation of space and time. In *Alveys conference*, volume 10, pages 467–478.

Howarth, R. and Buxton, H. (1992b). Analogical representation of spatial events for understanding traffic behaviour. In *Proc. of the ECAI*, pages 785–789.

Howarth, R. and Buxton, H. (1993). Selective attention in dynamic vision. In *Proc. of the IJCAI*, pages 1579–1584.

Hutber, D. (1995). *Suivi multi-capteurs de cibles multiples en vision par ordinateur, appliqué à un véhicule dans un environnement routier*. PhD thesis, I.N.R.I.A., Sophia Antipolis. in english.

Huttenlocker, P. and Rucklidge, W. (1992). Tracking non-rigid objects in complex scenes. In *proc. of Int'l Conf. on Computer Vision (ICCV)*, Berlin.

Intille, S. and Bobick, A. (1995). Closed-world tracking. In *Proc. of the 5th Int'l Conference on Computer Vision (ICCV)*, Cambridge, MA.

Johnson, N. and Hogg, D. (1996). Learning the distribution of object trajectories for event recognition. *Image and Vision Computing*, 14(8):609–615.

Karmann and Brandt, V. (1989). Detection and tracking of moving objects by adaptative background extraction. In *proc. of ICIP*, pages 1051–1058.

Koller, D., Daniilidis, K., and Nagel, H. (1993). Model-based object tracking in monocular image sequences of road traffic scenes. *International Journal of Computer Vision*, 10(3):257–281.

Koller, D., Weber, J., and Malik, J. (1994). Towards realtime visual ased tracking in cluttered traffic scenes. In *Proc. of the European Conf. on Computer Vision (ECCV)*, Sweden.

Kollnig, H., Nagel, H., and Otte, M. (1994). Association of motion verbs with vehicle movements extracted from dense optical flow fields. In *Proc. of the ECCV 94, Stockholm, Sweden.*

Kuniyoshi, Y. and Inoue, H. (1993). Qualitative Recognition of Ongoing Human Action Sequences. In *Proc. of the Int'l Joint Conference on Artificial Intelligence (IJCAI)*, volume 2, pages 1600–1609, Chambéry.

Lansky, A. (1988). Localized event-based reasoning for multi-agent. *Computational Intelligence*, 4(4):319–340.

Mann, R., Jepson, A., and Siskind, J. (1996). Computational perception of scene dynamics. In *proc. of the 4th European Conference on Computer Vision (ECCV)*, Cambridge, UK.

McCarthy, J. (1993). Notes on formalizing context. In *Proc. of the IJCAI, Chambery (France)*, pages 555–560.

Meyer, F. and Bouthemy, P. (1992). Region-based tracking in an image sequence. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 476–484.

Milhaud, N. and Médioni, G. (1994). Learning, recognition and navigation from a sequence of infrared images. In *Proc. of the Int'l Conf. on Pattern Recognition (ICPR)*, volume 1, pages 822–825, Jerusalem.

Miura, J. and Shirai, Y. (1993). An uncertainty model of stereo vision and its application to vision motion planning of robot. In *Proc. of the IJCAI*, page 1618.

Mohnhaupt, M. and Neumann, B. (1990). Understanding object motion: Recognition, learning and spatiotemporal reasoning. Research Report FBI-HH-B-145/90, University of Hamburg.

Motamed, C. and Vannoorenberghe, P. (1997). Video surveillance using behavioural knowledge. In *Proc. of the International Workshop on Dynamic Scene Recognition from Sensor Data*, Onera-Cert, Toulouse.

Nade, T. (1995). Module d'Acquisition et de Représentation d'Environnements Statiques (MARES). Master's thesis, DESS ISI.

Nagel, H. (1991). The representation of situations and their recognition from image sequences. In *Proc. of the RFIA Lyon Villeurbanne*, pages 1221–1229.

Nagel, H. H. (1988). From image sequences towards conceptual descriptions. *Image and Vision Computing*, 6(2):59–74.

Neumann, B. (1984). Natural language description of time-varying scenes. Technical report, FBI-HH-B-105/84 Fachbereich Informatik der Universität Hamburg, FRG.

Neumann, B. (1989). *Semantic structures : advances in natural language processing*, chapter 5, pages 167–206. David L. Waltz.

Nevatia, R. and Médioni, G. (1996). Computer vision research at the university of southern california. *Int'l Journal of Computer Vision*.

Olivier, P., Maeda, T., and Tsujii, J. (1994). Automatic depiction of spatial descriptions. In *Proc. of the AAAI Seattle, Washington*, pages 1405–1410.

Oppenheim, V. (1992). *Symbolic and knowledge-based signal processing*. Prentice Hall.

Ossola, J. (1996). *Coopération de systèmes à base de connaissances pour l'analyse et la reconnaissance d'objets naturels complexes : application au classement de galaxies ou de zooplanctons.* PhD thesis, Université de Nice - Sophia Antipolis.

P. Remagnino, J. Matas, J. I. and Kittler, J. (1993). A scene interpretation module for an active vision system. In *SPIE*, volume 2056, pages 98–107.

Pacholczyk, D. and Pacholczyk, J. (1996). Traitement symbolique des informations incertaines. In *Proc. of the RFIA, Rennes*, volume 2, pages 625–634.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufman, San Mateo, CA.

Pentland, A. (1995). Machine understanding of human action. In *proc. of the 7th Int'l Forum on Frontier of Telecommunication Technology*, Tokyo.

Piechowiak, S., Rodriguez, J., and Millot, P. (1994). Implémentation d'une méthode de diagnostic selon les premiers principes integrant le temps. In *proc. of the 14ème Journées Internationales Intelligence Artificiel, Systèmes Experts et Langage Naturel*, Paris.

Pinhanez, C. and Bobick, A. (1997). Human action detection using pnf propagation of temporal constraints. Technical report, M.I.T., Media Laboratory.

Polana, R. and Nelson, R. (1994). Low level recognition of human motion. In *Proc. of the IEEE workshop on motion of non-rigid and articulated objects*, Austin, Texas.

Prokopowicz, P., Swain, M., and Kahn, R. (1994). Task and environment-sensitive tracking. In *Proc. of the Workshop on Visual Behaviors*, pages 73–78.

Rao, B., Durrant-Whyte, H., and Sheen, J. (1993). A fully decentralized multi-sensor system for tracking and surveillance. *The International Journal of Robotics Research*, 12:20–44.

Retz-Schmidt, G. (1991). Recognizing intentions, interactions, and causes of plan failures. Projet VITRA 77, Universität des Saarlandes, Saarbrücken, Germany.

Ricquebourg, Y. (1993). Segmentation et suivi d'objets mobiles par modèles structurels adaptatifs. Master's thesis, INSA.

Robert, L. (1993). *Perception stéréoscopique de courbes et de surfaces tridimensionnelles. Applications à la robotique mobile.* PhD thesis, École polytechnique.

Rohr, K. (1994). Towards model-based recognition of human movements in images sequences. *Computer Vision and Graphism Image Processing (CVGIP)*, 59:94–115.

Royer, V. (1995). Hierarchical correspondance between physical situations and action models. Technical report, O.N.E.R.A. DES/SIA. available at the Perception project.

Sablayrolles, P. (1995). *Sémantique formelle de l'expression du mouvement. De la sémantique lexicale au calcul de la structure du discours en français.* PhD thesis, Thèse IRIT - Université Paul Sabatier Toulouse.

Sato, A., Mase, K., Tomono, A., and Ishii, K. (1993). Pedestrian counting system robust against illumination changes. In *Proc. of Visual Communications and Image Processing'93*, Massachusetts.

Schirra, J. (1992). Connecting visual and verbal space. In *proc. of the 4th workshop on time, space, movement and spatio-temporal reasoning*, Bonas, France.

Schirra, J. and Stopp, E. (1993). Antlima a listener model with mental images. *IJCAI, Chambéry, France*, 1:175–180.

Sellam, S. and Boulmakoul, A. (1994). Intelligent intersection: Artificial intelligence and computer vision techniques for automatic incident detection. *Artif. Intell. Applic. to Traffic Engng*, pages 189–200.

Shekhar, C., Moisan, S., and Thonnat, M. (1994). Towards an Intelligent Problem-Solving Environment for Signal Processing. *Mathematics and Computers in Simulation*, 36:347–359.

Sol, D. (1997). *Observateur : structures pour représenter, interpréter et décrire des scènes dynamiques*. PhD thesis, Université de Savoie, Chambéry, France.

Srihari, R. (1994). Computational models for integrating linguistic and visual information: A survey. *AIR*, 8(5-6):349–369.

Starner, T. and Pentland, A. (1995). Visual recognition of american sign language using hidden markov models. In *proc. of th Intl. Workshop on Automatic Face- and Gesture-Recognition*, Zurich.

Strat, T. (1993). Employing contextual information in computer vision. In *DARPA93*, pages 217–229.

Strat, T. and Fischler, M. (1990). A context-based recognition system for natural scenes and complex domains. In *DARPA90*, pages 456–472.

Thonnat, M., Clement, V., and van den Elst, J. (1994). Supervision of perception tasks for autonomous systems: the OCAPI approach. *Journal of Information Science and Technology*, 3(2):140–163.

Toal, A. and Buxton, H. (1992). Spatio-temporal reasoning within a traffic surveillance system. *Proc. of the European Conference on Computer Vision (ECCV)*.

Tsang, E. and Howarth, R. (1991). Scheduling in both space and time. In *Proc. of the 11th Workshop on Expert Systems and their applications, Avignon*, pages 361–373.

Tsuji, S. and Li, S. (1993). Making cognitive map of outdoor environment. In *Proc. of the IJCAI*, pages 1632–1638.

Turner, R. (1995). Context-sensitive, adaptive reasoning for intelligent AUV control: Orca project update. In *proc. of the 9th Int'l Symposium on Unmanned Untethered Submersible Technology (AUV'95)*, New Hampshire.

Wang, H. and Brady, M. (1995). Real-time corner detection algorithm for motion estimation. *Image and Vision Computing*, 13:695–703.

Welch, G. and Bishop, G. (1995). An introduction to the Kalman filter. Technical report, University North Carolina.

Wilson, A. and Bobick, A. (1997). Recognition and interpretation of parametric gesture. Technical Report 421, M.I.T., Media Laboratory Perceptual Computing Section.

Woodfill, J. and Zabih, R. (1991). An algorithm for real-time tracking of non-rigid objects. In *proc of AAAI*, Stanford.

Zhang, Z. (1993). Token tracking in a cluttered scene. Research report 2072, I.N.R.I.A., Sophia Antipolis.