Chapitre 9

Espace-temps, langage et vision

Une observation est une perception, mais une perception préparée à l'avance.
Karl Popper, *La connaissance objective.*

9.1 Un cadre empirique pour le raisonnement spatial

Nous allons aborder ici un problème pour lequel le raisonnement qualitatif sur le mouvement et la représentation de connaissances spatiales joue un rôle central, qui est celui des rapports entre langage et vision et plus spécifiquement ici les aspects liés à l'observation et à la description de scènes par des capteurs numériques. Cela permet aussi de tester certains aspects des travaux développés jusqu'ici en les plaçant dans un cadre ayant une importance pratique. L'observation de scènes est un problème complexe qui a des ramifications dans de nombreux domaines, depuis la vision (reconnaissance de formes, suivi de mouvements, etc) jusqu'à des aspects de l'intelligence artificielle dont nous ne nous préoccupons pas ici (la reconnaissance de plans ou la modélisation de systèmes multi-agents par exemple), et il n'est donc pas question de l'aborder dans toute sa généralité. Nous allons simplement montrer comment on peut articuler sur ce problème le langage que nous avons étudié pour parler du mouvement et raisonner sur des mouvements topologiques. Nous pouvons replacer cette étude dans un projet plus global de surveillance de scènes qui a été initié par les travaux d'Hervé Pensec (Pensec, 1996) et qui est présenté dans (Borillo et Pensec, 1995; Borillo et Pensec, 1996).

Depuis quelques années, la communauté de recherche en vision a jeté des ponts vers les communautés en I.A. qui se préoccupent de raisonnement spatio-temporel et de langage naturel pour enrichir les systèmes de vision généralement concernés par des tâches de "bas-niveau", sans considération pour la représentation de ce qui est vu ou des intentions des agents impliqués, c'est à dire sans représentation de haut niveau d'informations contextuelles utiles à l'interprétation d'images. La réflexion est engagée maintenant sur les liens entre les concepts mis en jeu dans l'interprétation d'images, leur ancrage dans le langage naturel et le raisonnement symbolique sur des données spatiales, temporelles, et spatio-temporelles, depuis le numéro spécial de AI Review (McKevitt, 1994), et de nombreux workshops organisés en parallèle de grandes conférences, tant en I.A. ("Representation and Processes between Natural Language and Vision" en 1996 à l'ECAI) que dans le domaine de la vision artificielle ("Conceptual descriptions from images" à l'ECCV'96). De nombreux travaux ont été consacrés à ces

problèmes depuis bien plus longtemps [1], mais la convergence d'opinion sur la nécessité de l'intégration de domaines jusque là assez distincts est relativement récente.
Cet intérêt est dû à la nécessité de raisonner sur des données symboliques, d'une part dans des application où la masse de données implique des mécanismes de tri et de description manipulables et compréhensibles par l'humain, voir par exemple (Pinhanez et Bobick, 1996), et pour reprendre le titre de (Wahlster, 1987), "un mot vaut 1000 images" pour décrire une situation à un opérateur humain ; les représentations symboliques et/ou en langage naturel ont d'ailleurs aussi un intérêt en synthèse d'images (Aurisset, 1997; Kalita et Lee, 1997). D'autre part ce besoin se fait sentir dans des systèmes d'observations où le système doit reconnaître des actions complexes, mettant éventuellement en jeu les intentions et les plans des agents observés, par exemple dans des projets variés (Fernyhough *et al.*, 1998; Intille et Bobick, 1998; Brémond et Thonnat, 1997; Castel *et al.*, 1996; Nagel, 1994; Nagel et Kollnig, 1996) ou bien doit pouvoir faire du raisonnement spatial ou temporel de haut niveau, cf. l'argumentation de (Neumann et Schroder, 1996).

Le mouvement occupe une place très importante dans cette perspective. La détection du mouvement est un problème central en vision pour la reconnaissance et le suivi d'objets. Pour la description de haut niveau il est également central dans la mesure où les comportements observés sont d'abord des mouvements, à partir desquels les actions et les intentions des agents sont étudiés et détectés. Nous allons surtout ici tenter de montrer l'intérêt du raisonnement spatio-temporel qualitatif pour la description de situations dynamiques, pour synthétiser l'information présente dans une scène observée ou raisonner symboliquement sur cette information. Le langage du mouvement est donc particulièrement crucial pour la description de telles situations.

9.2 Quelques approches de la description de scènes

La cas de l'observation Un cas typique d'application qui met en jeu des informations de haut niveau, du raisonnement spatio-temporel et du raisonnement sur des intentions d'agents est le cas de systèmes de surveillance, où l'objectif est de détecter des comportements et de reconnaître des actions particulières, comme un vol de voiture sur un parking (Castel *et al.*, 1996), des actions de conducteurs sur des routes (Fernyhough *et al.*, 1998; Nagel, 1994), ou bien des actions de jeu dans des matchs sportifs (Herzog et Wazinski, 1994; Intille et Bobick, 1998). Il y a alors alliance de plusieurs modules : un module destiné à la reconnaissance et au suivi d'objets dans la scène qui vont servir d'entrée au niveau symbolique, un module de traduction entre les représentations numériques et les représentations symboliques manipulées par le dernier module qui fait du raisonnement spatio-temporel, généralement pour faire de la reconnaissance de plans, soit à l'aide d'un graphe d'états qui détermine le scénario représenté le plus plausible pour la situation observée (Castel *et al.*, 1996; Fernyhough *et al.*, 1998), soit par des algorithmes spécifiques qui associent des scénarios possibles aux événements observés (Pensec, 1996; Maaß, 1994).
Nous allons plutôt insister sur les quelques approches qui font intervenir explicitement des expressions en langage naturel pour manipuler l'information spatio-temporelle pertinente dans les scènes observées. On verra aussi les liens entre ces expressions et le raisonnement spatio-temporel.

1. On peut faire remonter les premiers travaux à (Badler, 1975), cf. aussi dans les années 80, le projet Vitra (Wahlster, 1987).

9.2.1 Le projet VITRA (VIsual TRAnslator)

Le projet Vitra est un projet de l'Université de Saarbrück qui a débuté en 1985 (Wahlster, 1987; Herzog et Wazinski, 1994). Son objectif était la génération de descriptions langagières de scènes spatiales obtenues par des systèmes de vision artificielle. Plus spécifiquement le projet était focalisé sur les relations spatiales existant entre objets présents dans une séquence de scènes (2D ou 3D). Deux sous-projets sont représentatifs de l'approche : Citytour et Soccer. Le premier était une aide pour des touristes consultant un plan de ville, avec des trajectoires d'objets (véhicules, piétons) sur ce plan. Le système peut répondre à des requêtes faites a posteriori sur l'évolution de la scène, avec des expressions en langage naturel utilisant des prépositions spatiales ou des verbes de mouvement. Le deuxième projet avait pour objectif la description automatique de matchs de football, allant jusqu'à la reconnaissance d'actions complexes et d'intentions des joueurs à partir des données spatiales, le tout pour un commentaire en temps réel du match.

La méthodologie générale de ce projet ambitieux était la suivante : l'entrée du système est constituée de cartes où les objets avaient été reconnus et traités par des systèmes classiques de vision bas-niveau, associant une identité à chaque objet de manière à pouvoir leur associer certaines propriétés non purement géométriques (comme le fait d'avoir une orientation intrinsèque, un type particulier, etc...). De là le système calcule des relations soit de façon absolue (par exemple, un objet est-il à l'intérieur d'une région ou non) soit par des fonctions d'applicabilité ad hoc qui rendent globalement compte d'effets contextuels propres aux applications choisies pour certaines expressions telles que "l'objet x est devant l'objet y". Le mouvement d'objet peut être reconstruit sur la même base, avec des représentations différentes pour chaque projet : dans Citytour, un mouvement est une liste de couples $\langle point, instant \rangle$ localisant le "centre" d'un objet, pour Soccer une suite de relations de localisation (l'objet est dans une zone, puis une autre). et les verbes de mouvement sont représentés par un ensemble de contraintes portant sur ces types de trajectoires, suivant l'étude présentée dans (Hays, 1989). Dans le cadre du projet Soccer, ces représentations sont combinées avec des représentations d'actions possibles et de plans qui doivent permettre de déterminer les intentions des agents présents dans la scène, puis de produire une description linguistique de la scène.
Pour donner une idée de la représentation des éléments lexicaux, on peut prendre l'exemple de la préposition "at" (*à*). La fonction d'applicabilité de cette préposition reliant un objet à localiser x et un objet de référence y, de coordonnées respectives x_1, x_2 et y_1, y_2 correspondant à la position de leur centre sur l'image, est donnée par la formule suivante où d_{ref} est une distance de référence réglée manuellement pour chaque situation afin de tenir compte d'effets contextuels :

$$at(x, y) = \exp\left(-\frac{(x_1 - y_1)^2 + (x_2 - y_2)^2}{d_{ref}^2}\right).$$

Plus le degré d'applicabilité est élevé plus l'expression "x est à y" est admissible pour décrire la scène. Des prépositions plus complexes sont modélisées par des fonctions spline, par exemple le degré d'applicabilité de "behind" (*derrière*) est illustré schématiquement figure 9.1. Dans le cas de réponses à des requêtes fermées (la relation R est elle vraie entre x et y ?), le système calcule simplement le degré d'applicabilité, qui doit être au-dessus d'une valeur seuil donnée pour que la relation soit exprimable. Dans le cas de descriptions de matchs de football, le système dispose d'un algorithme de reconnaissance de plans pour inférer les intentions possibles des joueurs à partir de l'observation de leur comportement spatial.

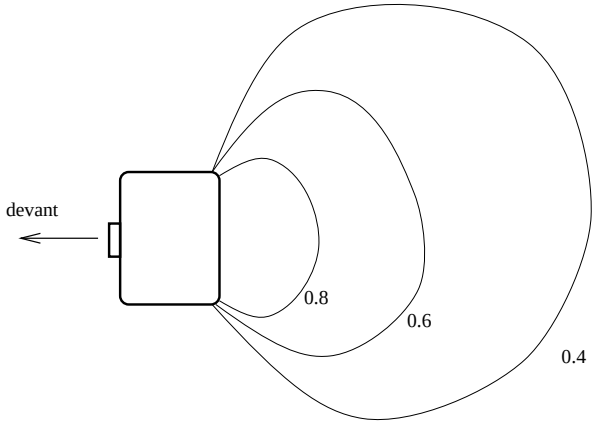

FIG. 9.1 - *Degrés d'applicabilité pour "derrière"*

Portée et limites de Vitra Ce projet a été une étude assez vaste sur les méthodes à mettre en oeuvre pour lier langage et vision d'une façon systématique et sur l'information spatiale mise en jeu dans le langage et la vision. Le plus gros problème resté non résolu de façon satisfaisante est celui du choix des descriptions adéquates quand plusieurs possibilités semblent possibles. Dans chaque application de Vitra le problème est résolu en amalgamant différents facteurs pragmatiques en une fonction de décision, mais le poids de chaque facteur n'était jamais explicite et les quelques exemples qui illustraient le mécanisme sont trop simplifiés pour être généralisables. La fonction graduelle d'applicabilité, même si on peut lui trouver des justifications linguistiques ne peut prétendre à aucune généralité, est probablement réglée différemment pour chaque relation et dans chaque application. L'ontologie même du mouvement est différente entre Citytour et Soccer, car il n'y a aucune séparation entre la représentation de la sémantique des éléments lexicaux et la façon dont les choix de descriptions étaient implémentés pour des tâches différentes. C'est une des raisons qui nous a fait exprimer sous forme logique les contraintes qui portent sur la sémantique des éléments lexicaux utilisés au chapitre 8 par exemple, de façon à les rendre génériques. Même pour un domaine particulier comme celui des descriptions d'itinéraires, cela permet de rendre explicite ce que l'implémentation devra respecter pour être conforme au sens des expressions utilisées.

9.2.2 Le projet Wire

Dans la même veine que Vitra, mais conscient du manque de généralité du projet par rapport à la sémantique de l'espace et du mouvement, le projet plus modeste Wire (Borillo et Pensec, 1996) s'est efforcé d'apporter quelques réponses au problème de la description de scènes observées par des capteurs. L'objectif est, comme dans Soccer, de reconnaître les intentions des agents présents dans les scènes observées, et de décrire ces plans en langage naturel, en utilisant des modèles de l'espace et du temps qualitatifs. Le type de scènes observées est celui de régions géographiques dans lesquelles évoluent des véhicules circulant sur un réseau de voies de communication, les agents dont le système doit reconnaître les actions étant des unités militaires. Certains types d'information symbolique sont utilisés pour représenter des prédicats de langage, essentiellement les relations méréo-topologiques classiques et les relations du calcul d'événements pour le temps. Par exemple, la classification des verbes de

mouvement de (Laur, 1991) est utilisée pour caractériser des types de mouvement qualitatifs pertinents pour la description de scènes. A la différence de Vitra, la représentation des expressions linguistiques est donc fondée sur des concepts réellement qualitatifs et donc plus proches du sens commun qu'ils doivent modéliser. Malheureusement ces concepts qualitatifs sont particuliers au format des données observées. Les scènes étant des images planes (vues de dessus), les objets sont représentés par des polygones au niveau numérique, et au niveau symbolique, ce sont les relations qualitatives *entre les polygones 2D* qui sont calculées, et non pas des relations entre les objets dans un espace réel en 3D. Cela fait que la représentation des prédicats de langage est dépendante du format des données numériques et cette représentation de la sémantique des expressions est spécifique à l'application visée. Par exemple, la préposition *sur* mettant en relation un mobile x et une portion de voie de communication y est représenté par PP(x, y) parce qu'en vue de dessus, le polygone déterminé par un mobile ne peut être qu'inclus dans le polygone déterminé par la portion de route ; mais on ne veut pas avoir comme interprétation que l'objet fait partie de la route. L'intérêt de manipuler la sémantique avec des primitives de sens commun est de s'abstraire des particularités de chaque application pour la représentation du sens (au contraire de Vitra qui a une représentation sémantique différente pour chaque sous-projet), et de pouvoir ainsi réutiliser des études linguistiques génériques. C'est donc sur l'étape (indispensable) de traduction entre la représentation numérique et les concepts symboliques que porte les particularités de l'application visée, mais pas sur le modèle symbolique lui-même. Sans cela on ne peut en effet se servir des propriétés inférentielles de ces théories puisque l'on a pas respecté leur interprétation, et en effet, (Pensec, 1996) doit utiliser un module de raisonnement spécifique qui ne prend pas en compte des propriétés des théories qualitatives utilisées. Un autre avantage du découplage entre la manipulation des informations symboliques et la façon dont elles sont calculées à partir des données numériques est de pouvoir utiliser de façon modulaire des procédures de calcul efficace pour déterminer ces informations symboliques. Si on a besoin de calculer si la relation $sur(x, y)$ est vérifiée dans une scène, on peut considérer que cette préposition est une primitive (et on peut la calculer de façons différentes selon que l'on a une scène 2D vue sous un angle ou un autre, ou 3D, etc.) et avoir une représentation explicite des inférences possibles avec les autres expressions représentées dans le système (par exemple les contraintes sur les itinéraires vues au chapitre précédent). Le projet Wire se propose de faire de la reconnaissance de plans dans le contexte que l'on a mentionné, et l'ambition du projet a pu justifier quelques raccourcis au niveau de la représentation pour le rendre faisable. Notre ambition, plus limitée, est de voir comment on peut articuler les informations symboliques et numériques en faisant du raisonnement, en gardant des représentations génériques des expressions du langage naturel et des informations qualitatives spatio-temporelles, en ne faisant porter les particularités d'une éventuelle application informatique que sur le calcul des primitives symboliques.

Une autre particularité de Wire qui le distingue nettement de Vitra est le constat que l'on ne peut décrire une scène par des prédicats de langage sans avoir défini au préalable des intentions pour l'observateur. Un gros problème de Vitra était en effet de choisir parmi plusieurs descriptions possibles d'une scène (parce que plusieurs relations avec des objets différents ou non ont un degré d'acceptabilité important) celles qui est pragmatiquement la plus adéquate. Plusieurs types d'intentions sont mentionnés pour préciser ce choix de descriptions, bien qu'il manque les définitions explicites des concepts de pertinence, qui joue par exemple pour des requêtes comme "qu'y a-t-il eu d'anormal dans la scène entre t et t' ?" (Borillo et Pensec, 1996).

9.2.3 Bilan

Nous pouvons retenir sur ces études des rapports langage et vision, qu'assez peu d'auteurs se sont posé des question sur la nature des concepts spatiaux mis en jeu dans l'expression du mouvement, sauf (Fernyhough *et al.*, 1998) qui a incorporé quelques notions qualitatives, et (Pensec, 1996) qui a pris en compte certaines études linguistiques. On peut alors distinguer les approches de la vision de haut-niveau utilisant le langage comme médium d'analyse selon quelques critères qui n'ont sans doute rien d'exhaustif et que nous recensons pour essayer de synthétiser les paragraphes précédents :

- l'ontologie du mouvement adoptée, suivant les critères du chapitre 2, c'est à dire relatif/absolu, utilisant des points ou des régions, etc...
- la présence de raisonnement explicite sur les informations manipulées.
- la présence d'une représentation lexicale, ad hoc ou non.
- la prise en compte explicite des intentions d'observations pour la description.

Nous nous situons bien sûr dans une approche où nous nous préoccupons d'informations spatio-temporelles de nature qualitative et de leurs liens avec la sémantique du langage naturel. Nous allons donc essayer de voir comment articuler ces éléments dans un contexte d'observations de mobiles donnés par des capteurs numériques, ce qui va donc d'une certaine façon s'abstraire de la représentation numérique utilisée.

9.3 Intentions descriptives et données spatiales

Nous avons mentionné le problème du choix de descriptions d'une scène dans l'absolu ; on peut voir figure 9.2 un exemple réduit d'une scène avec trois mobiles (les disques) deux repères fixes (les polygones grisés) et quelques éléments de voies de communication. Dans l'absolu, on peut appliquer de nombreuses descriptions des localisations et des mouvements des objets présents. En contexte toutes n'ont pas la même portée. Si le système doit surveiller la scène pour protéger la banque par exemple, les relations des mobiles par rapport à la banque sont privilégiées. Cela correspond à l'intention de surveiller un repère particulier ou un type de repère particulier et de surveiller la présence de mobiles aux abords de la banque : ici "A et B passent devant la banque". On peut avoir des informations en plus sur certaines entités, par exemple que D est un cambrioleur (appelons le Dortmunder). Alors par rapport à la surveillance de la banque, cette entité est pertinente et il faut donc décrire que Dortmunder est en face de la banque. Le mouvement apporte des données en plus, si on surveille des mobiles, dans la mesure où les carrefours vers lesquels ils circulent vont donner une indication des endroits où ils peuvent passer. On peut vouloir aussi surveiller des séquences d'actions particulières, comme le fait pour un véhicule de passer plusieurs fois devant la banque. Nous ne prétendons pas donner une théorie de l'intention en général [2], ni même de construire exhaustivement toutes les situations possibles même dans un cadre restreint. Cependant nous pouvons définir quelques intentions de description de base, en rapport avec des propriétés spatio-temporelles, pour fournir quelques filtres qui vont utiliser les informations sur la nature des entités et leur comportement spatio-temporel.

2. On peut mesurer l'ampleur du problème en consultant le recueil d'articles sur le sujet dans (Cohen *et al.*, 1990).

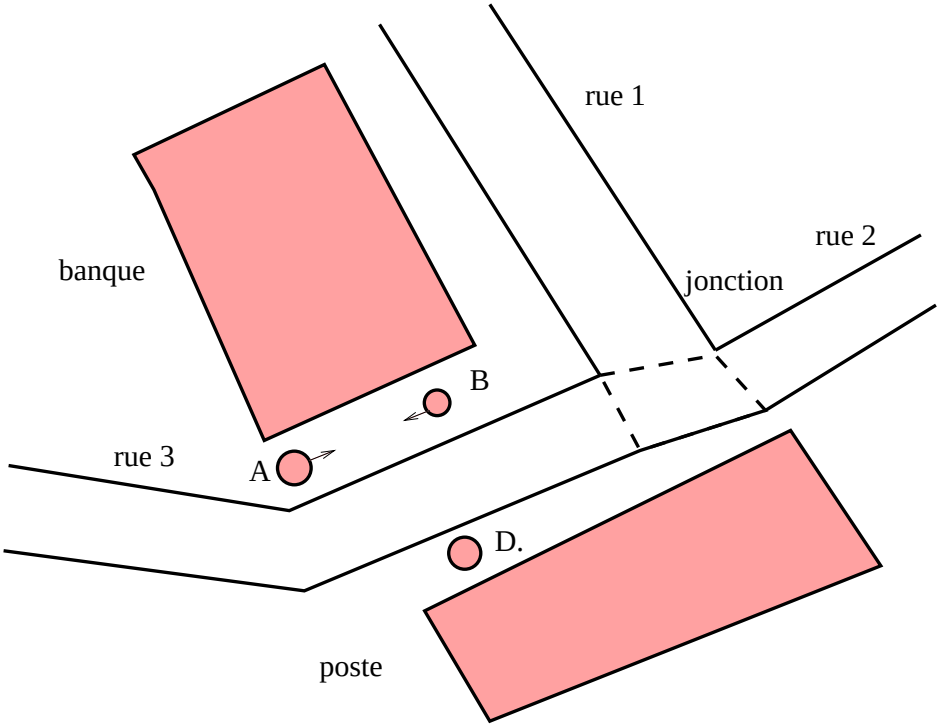

FIG. 9.2 - *Scénario exemple.*

9.3.1 Quelques exemples de descriptions simples

Nous entendons par description simple une éventualité homogène, comme un état ou un processus. Elle correspond à l'observation d'une propriété qui dure nécessairement pendant un intervalle de temps, comme *le véhicule est sur la route RN12*. Si une telle propriété est observée au temps d'observation t, on peut dire qu'il existe une éventualité e telle que le moment t fasse partie du temps de e.
Nous avons vu que les descriptions peuvent être guidées par des restrictions sur les entités mises en jeu ; on peut vouloir désigner des entités particulières ou bien des types d'entités. On peut les représenter sous forme de structure de traits pour avoir une forme proche de celle des entrées lexicales. La structure suivante correspond alors à la surveillance de véhicules par rapport à la constante banque_de_france :

$$\begin{bmatrix} \text{ARG1} & \text{x:vehicule} \\ \text{ARG2} & \text{banque\_de\_france} \end{bmatrix}$$

Une éventualité simple est donc formé sur le schéma suivant, que l'on pourrait appeler "localisation simple" :

$$\begin{bmatrix} \text{Event\_Str} & \begin{bmatrix} \text{éventualité} & \text{e:état} \vee \text{processus} \\ \text{ARG1} & \boxed{1} \\ \text{ARG2} & \boxed{2} \end{bmatrix} \end{bmatrix}$$

Et on peut instancier les traits [1] et [2] de diverses manières. Dans l'exemple de la banque, l'intention ci-dessus peut s'unifier avec les médians externes atéliques comme *longer*.
En représentant les prépositions sur le même schéma, par exemple *sur* :

$$\begin{bmatrix} \text{Event\_Str} & \begin{bmatrix} \text{éventualité} & \text{e:état} \\ \text{ARG1} & \boxed{1} \\ \text{ARG2} & \boxed{2} \\ \text{prédicat: sur}(\boxed{1},\boxed{2},\text{e}) & \end{bmatrix} \end{bmatrix}$$

Et en définissant la surveillance de la rue particulière dans la scène de la banque comme :

$$\begin{bmatrix} \text{Event\_Str} & \begin{bmatrix} \text{éventualité} & \text{e:état} \vee \text{processus} \\ \text{ARG1} & \text{x:vehicule} \\ \text{ARG2} & \text{rue\_3} \end{bmatrix} \end{bmatrix}$$

On peut unifier l'intention de surveillance avec la préposition *sur* :

$$\begin{bmatrix} \text{Event\_Str} & \begin{bmatrix} \text{éventualité} & \text{e:état} \\ \text{ARG1} & \text{x:vehicule} \\ \text{ARG2} & \text{rue\_3} \\ \text{prédicat: sur(x,rue\_3,e)} & \end{bmatrix} \end{bmatrix}$$

Et il revient alors au module de traduction numérique/symbolique de vérifier si un véhicule est bien sur cette portion de route à un moment donné, ce qui donne une valeur à x et un ou des temps contenus dans e. Avec ces objectifs simples de description on a bien sûr encore plusieurs possibilités, mais ce genre de filtre opère déjà une sélection non négligeable, et fonctionne de façon à intégrer les composantes lexicales et visuelles. Dans un contexte de véhicules circulant sur un ensemble de voies de communications déterminées on a vu au chapitre précédent que l'on a beaucoup plus de contraintes sur les descriptions pertinentes. Les localisations se faisant sur les segments, les types de relations se ramènent aux relations entre les véhicules et les tronçons lexicaux induits par les repères. Dans ce cas on peut définir des priorités de description, comme par exemple dire qu'une localisation sur un tronçon est plus précise qu'avant/après/vers le repère associé à ce tronçon, et plus précise que n'importe quelle relation par rapport à un repère situé sur un autre segment. Ces contraintes, d'ordre pragmatique, combinées aux schémas de sélections introduits plus haut permettent souvent de déterminer des localisations uniques. Par exemple, pour la surveillance des véhicules par rapport à la banque, on aurait uniquement des descriptions de véhicules situés sur un tronçon associé à la banque. Nous n'explorerons pas systématiquement ici les contraintes essentiellement pragmatiques qui peuvent permettre de guider l'interprétation en contexte puisque le but est surtout ici de voir comment articuler le différents composants que l'on pense indispensable au processus. Nous avons commencé dans (Maudet et Muller, 1998) à étudier les interactions entre divers principes qui contraignent la communication dans le cas de dialogues portant sur des informations spatiales, et il faudrait prolonger ces travaux pour pouvoir les impliquer dans les préoccupations de ce chapitre.
Par ailleurs, si à ce stade il n'y a pas de raisonnement à effectuer, on peut voir que s'il le fallait, la représentation lexicale introduit les prédicats de langage (*sur* dans l'exemple) qui font le lien avec la théorie du domaine spatio-temporel.

9.3.2 Événements

La description d'un mouvement correspondant à un changement spatial, peut s'exprimer aussi simplement, en faisant intervenir la structure de mouvement associée aux expressions verbales de déplacement.
Par exemple, le changement de localisation d'un véhicule par rapport à un repère peut s'exprimer comme suit :

$$\begin{bmatrix} \text{Event\_Str} & \begin{bmatrix} \text{éventualité} & \text{e:event} \\ \text{ARG1} & \text{x:vehicule} \\ \text{ARG2} & \text{repere24} \end{bmatrix} \\ \text{Mvt\_Str} & \begin{bmatrix} \textit{déplacement} & \\ \text{transitionnel} & + \end{bmatrix} \end{bmatrix}$$

On peut réduire les possibilités de description une nouvelle fois en ramenant le problème aux tronçons associés au repère, en considérant qu'une localisation par rapport à un repère correspond à être situé à un moment sur un tronçon associé. En fait cela revient à distinguer des types de localisation, celui sur lequel on met l'accent ici étant particulier aux voies de communications. On peut l'indiquer alors en typant les structures de traits correspondants. Cela revient à préciser la forme d'un déplacement en tenant compte de la contrainte évoquée au chapitre précédent :

$$\text{deplacement}(e) \rightarrow \exists t \exists y\ (\text{RS}(t) \wedge \text{associate}(t,\text{site}(e)) \wedge sur(\text{cible}(e), t, \text{PSP}(e)))$$

Le déplacement revient donc à combiner une localisation simple sur le tronçon associé au repère et identifier un changement. La relation temporelle entre le moment de la localisation et celui du changement observé déterminera une polarité, et le type du repère implique certaines contraintes sur le type de la relation de localisation (par exemple elle sera interne pour une région, de contact pour une frontière, spécifique pour un objet). On aura alors la classe de déplacement correspondant.

9.3.3 Intentions plus complexes.

On peut décrire des intentions plus complexes comme des chaînes d'intentions basiques identiques à celles déjà vues :

$$\begin{bmatrix} \textit{intention} & \\ \text{intention\_1} & \begin{bmatrix} \textit{intention} \\ \ldots\ \ldots \end{bmatrix} \\ \text{intention\_2} & \begin{bmatrix} \textit{intention} \\ \ldots\ \ldots \end{bmatrix} \end{bmatrix}$$

Les liens pouvant être indiqués par des restrictions supplémentaires (par exemples des contraintes temporelles ou spatio-temporelles entre les éventualités introduites). Par exemple, la surveillance de véhicules par rapport à une région combinée avec une localisation, comme quitter Toulouse puis aller vers l'aéroport de Blagnac pourrait être représenté sous la forme :

$$\left[\begin{array}{l} \textit{intention} \\ \left[\begin{array}{ll} \text{Event\_Str} & \left[\begin{array}{ll} \text{type} & e_1\text{:event}\boxed{1} \\ \text{ARG1} & \text{x:vehicule}\boxed{2} \\ \text{ARG2} & \text{Paris} \\ \text{prédicat} & \text{quitter(x,Toulouse,}e_1\text{)} \end{array}\right] \\ \text{Mvt\_Str} & \left[\begin{array}{l} \text{...} \end{array}\right] \end{array}\right] \\ \left[\begin{array}{ll} \text{Event\_Str} & \left[\begin{array}{ll} \text{type} & e_2\text{:state} \\ \text{ARG1} & \boxed{1} \\ \text{ARG2} & \text{aer\_bl} \\ \text{prédicat} & \text{vers(}\boxed{2}\text{,aer\_bl,}e_2\text{)} \\ \text{restriction} & \boxed{1} < e_2 \end{array}\right] \end{array}\right] \end{array}\right]$$

Avec ce niveau de complexité descriptive, on pourra avoir alors besoin de tester la validité d'une requête, ce qui ne peut se faire sans raisonnement spatio-temporel, et on peut donc utiliser ici notre modèle du chapitre 4 (et 7) ajouté aux contraintes propres aux itinéraires du chapitre 8. Par exemple le type de restriction temporelle introduite dans l'exemple peut être utilisé avec la table de composition d'information spatio-temporelle (donnée par les relations introduites par chaque intention) et d'information temporelle pour donner de nouvelles relations spatio-temporelles.

Par ailleurs pour reconnaître une séquence d'éventualités à partir d'observations ponctuelles, il faut pouvoir raisonner sur les extensions temporelles des relations que l'on cherche à reconnaître successivement et on a là aussi besoin d'évaluer les relations compatibles avec les observations, que cela soit fait avec un module dédié spécifiquement au raisonnement temporel comme ce qui est fait dans (Pensec, 1996), ou bien avec les types d'inférences déjà mentionnés.

9.4 Le cas des itinéraires.

On a introduit dans les section précédentes un vocabulaire destiné à expliciter les intentions d'observation que l'on pouvait avoir à modéliser. La forme que nous leur avons donnée est proche de celle des éléments lexicaux qui servent à décrire les situations, montrant le lien qui existe entre ces deux types d'informations, et pourquoi le langage naturel est un bon médium entre les données spatio-temporelles et certains objectifs de perception.
Pour voir comment on peut manipuler des intentions structurellement plus complexes, nous allons considérer le cas de descriptions d'itinéraires dans un contexte d'observation. On peut considérer deux exemples de surveillance d'itinéraires :

– dans un premier cas, l'objectif est de surveiller un parcours déterminé pour détecter les entités qui le suivent. Ce parcours est déterminé par un ensemble de repères ou de voies de communication

à surveiller. On peut en déduire une structure pour le parcours sur le modèle des itinéraires que l'on a présenté, en introduisant un ensemble de tronçons, déterminés par rapport aux repères. Le but est ensuite d'identifier les correspondances de trajectoires avec ce parcours.

– dans un deuxième cas, on peut vouloir surveiller des mobiles particuliers à certaines périodes (entre certaines prises de vues espacées arbitrairement dans le temps). Leur surveillance prend alors la forme de la description de leur itinéraire au cours du temps, ce qui revient à choisir les repères qui vont permettre ces descriptions (en tenant compte des aspects étudiés plus haut).

Sans avoir réellement une sortie correspondant à des phrases en langage naturel, on aurait alors des descriptions à base de prédicats du langage naturel qui serait facilement compréhensibles par un opérateur humain. C'est d'ailleurs l'objectif avoué du projet Wire, partiellement réalisé dans (Pensec, 1996).

On applique donc dans le premier cas l'algorithme suivant : Soit LO=$\langle O_1, ..., O_n \rangle$ la liste des objets déterminant l'itinéraire à surveiller.

1. on calcule LS= $\langle \langle O_1, L_1 \rangle, ..., \langle O_n, L_n \rangle \rangle$ avec $L_i = \{s_{i1}, ...s_{ip}\}$ l'ensemble des segment associés à l'objet O_i . On a alors $\forall i, j$:

   • si O_i est du type vc, alors $L_i = \{O_i\}$

   • sinon les s_{ij} sont les segments tels que $associate(s_{ij}, O_i)$

2. soit LD=Ø et LV=$\{x \mid vehicule(x)\}$

3. Tant que LV$\neq$ Ø faire

   • choisir x dans LV

   • chercher s'il existe e_1 tel qu'il y ait une localisation de x sur un élément de L_1. soit t_1 le temps de la prise de vue qui est pendant cet éventualité.

   • si oui on répète l'opération précédente en cherchant à chaque étape i un e_i et un t_i tel que $e_{i-1} < e_i$ et tel qu'il y ait une localisation de x sur un élément de L_i (une intention simple que nous avons déjà vue plus haut). Cette recherche s'arrête au pire à la dernière prise de vue [3]. Alors si on a $i = n$, c'est que le véhicule a suivi l'itinéraire à surveiller, et on l'ajoute à la liste des descriptions résultats LD=LD $\cup \langle x, t_1, t_n \rangle$.

   • sinon LV=LV/$\{x\}$

4. LD est alors la liste des véhicules ayant suivi l'itinéraire, avec les indices des scènes de début et de fin correspondantes.

3. On pourrait améliorer cette étape pour arrêter la recherche quand on trouve une impossibilité à un temps donné que l'itinéraire soit poursuivi ; nous n'avons pas exploré cette voie, mais elle nécessiterait de raisonner sur les localisations successives observées.

Pour le deuxième cas en rapport avec un itinéraire à surveiller, on applique l'algorithme suivant : on a en entrée un véhicule x entre les temps d'observation t_1 et t_n. On peut avoir aussi une intention particulière int portant sur le véhicule qui orientera la description finale (comme de privilégier certains types de repères).

1. soit $L_D = \emptyset$
2. soit pour chaque t_i entre t_1 et t_n :
   - on cherche une localisation de x sur un segment ou une jonction y_i (c'est une expression, notée D).
   - on calcule une description D' sur y_i compatible avec int. Si int=[], D'=D.
   - on ajoute D' à L_D.
3. on élimine les redondances de L_D.

On a donc un itinéraire décrit par la suite de segments ou de carrefours sur lesquels on a observé le véhicule si aucune intention ne vient contraindre la description, ou bien des descriptions guidées par de intentions de la forme de celle vues plus haut, que ce soit des localisation simples ou bien des événements.

9.5 Conclusion

Nous avons essayé de montrer ici comment il était possible d'articuler des composantes de représentation : lexicale, spatio-temporelle, et visuelle (en fait ici, des représentations numériques enrichies par certaines informations symboliques) pour enrichir un système de vision avec des informations symboliques de haut niveau, en prenant en compte explicitement la forme des intentions d'observation d'un opérateur humain. Nous avons insisté sur des représentations génériques du lexique, des intentions et du modèle symbolique spatio-temporel, afin de les garder indépendantes des traitements numériques qui permettent de passer des données issues de capteurs aux données symboliques. Cela permet d'une part de fournir des exemples de raisonnement spatio-temporels utilisables par un tel système, et d'autre part de dépasser quelques-unes des limitations observées par certaines des approches qui ont pour but d'intégrer des représentations du langage naturel avec des représentations visuelles. Le développement d'un système complet n'est pas notre but, et à ce titre ce que nous avons présenté est forcément embryonnaire ; il permet cependant de montrer la faisabilité d'une mise en œuvre pratique des concepts qui ont été étudiés ici. L'annexe C présente brièvement la plate-forme de test de ce qui a été présenté ici. Il manque à cette partie une étude algorithmique plus complète vis à vis des situations réelles rencontrées, ainsi qu'une estimation plus précise de la complexité de cette mise en œuvre. Nous n'avons fait qu'indiquer ce qui nous semble apporter des améliorations à ce problème si l'on considère les travaux par rapport auxquels nous nous plaçons.

Conclusion

Nous avions comme objectif de développer une théorie du mouvement de sens commun, prenant en compte la façon dont il est perçu par l'humain et la façon dont on communique en langage naturel à propos de situations spatio-temporelles. Nous avons vu dans un premier temps que beaucoup d'approches théoriques du mouvement étaient très influencées par le modèle de l'espace et du temps qui sous-tend la physique classique ; celui-ci implique une vision "permanente" des objets et en conséquence une séparation du temps et de l'espace en deux domaines distincts, qui sont de plus "absolus" : ils sont indépendants des objets et événements qui y prennent place. En s'inspirant d'une part de théories qui ont remis les objets et événements dans leur entièreté au centre de leurs investigations et non un espace et un temps abstraits constitués d'éléments ponctuels, et d'autre part de conceptions qui prônent une vision globale spatio-temporelle de ces éléments de notre perception, nous avons développé une théorie méréo-topologique d'entités spatio-temporelles que nous pensons plus adéquate pour décrire le mouvement, et indirectement l'espace et le temps. Nous nous appuyons notamment sur des études linguistiques ou d'ontologie formelle pour pouvoir caractériser la nature du mouvement de sens commun.

Afin de montrer l'expressivité de la théorie et son adéquation aux concepts cognitifs qui nous intéressent, nous avons montré comment elle permet de reformuler quelques problèmes liés aux objets et événements tels qu'ils sont décrits en langage naturel, et de donner ensuite une représentation sémantique d'expressions linguistiques décrivant le mouvement.

Les propriétés que la théorie prétend formaliser ont d'autre part été étudiées de plusieurs façons : les modèles de la théorie axiomatique présentée ont été étudiés, nous avons vu quelques types d'inférences permis par cette théorie qui semblent correspondre à l'intuition, et nous avons utilisé la théorie pour analyser les rapports entre le concept de continuité et les représentations qualitatives du temps et de l'espace qui sont le domaine d'étude de nombreux chercheurs en I.A.

Ensuite nous avons continué à explorer les utilisations possibles des concepts modélisés en les appliquant à la représentation de connaissances dans un domaine plus contraint qui est celui de la représentation d'itinéraires, domaine que nous estimons inséparables du raisonnement spatio-temporel et de la représentation du langage naturel. Cette intégration d'aspects spatio-temporels et linguistiques nous a permis une incursion dans un domaine qui est au confluent de ces préoccupations, à savoir l'interprétation de haut niveau de scènes visuelles, que nous avons pris comme exemple mettant en jeu tout ce que nous avions développé auparavant.

Nous avons tenté de respecter au cours de ce travail une méthodologie qui repose sur trois assises de types différents. La première est théorique et consiste à étudier la cohérence et la portée de notre étude en utilisant une approche axiomatique. La deuxième est de tester la validité cognitive de l'approche théorique en la confrontant à l'intuition au moyen de l'étude de l'adéquation au langage. Enfin l'étude pratique de raisonnements sur des informations spatio-temporelles dans des contextes particuliers est

un moyen de vérifier la possibilité d'utiliser le modèle de façon pratique.
Les objectifs propres à chaque élément de cette méthodologie ont été remplis à des degrés divers, étant données les contraintes qui président à la réalisation d'une thèse et nous avons indiqué les prolongements naturels de nos travaux dans chacun des aspects de notre démarche. Nous avons en effet volontairement restreint notre étude à ce qui nous paraissait le plus essentiel, la topologie et la méréologie de l'espace-temps, en ignorant de nombreux aspects complémentaires qui sont partie intégrante du mouvement naturel, orientation, distance, durée, etc. Cette étude doit maintenant se poursuivre en intégrant ces nouveaux aspects. De même certains aspects de ce que nous considérons comme faisant partie de la validation de notre approche ne pouvaient être qu'esquissés si nous voulions compléter la démarche globale, comme l'étude systématique des propriétés inférentielles du modèle spatio-temporel que nous avons proposé.