Niveaux de représentation pour la vision par ordinateur : indices d'image et indices de scène

Yves Demazeau

THESE

Présentée à

L'INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

pour obtenir le grade de
docteur de l'Institut National Polytechnique de Grenoble
(Spécialité : Informatique)

par

DEMAZEAU Yves

OOOOO

NIVEAUX DE REPRESENTATION POUR LA VISION PAR ORDINATEUR. INDICES D'IMAGE ET INDICES DE SCENE

OOOOO

Thèse soutenue le 22 Décembre 1986 devant la commission d'examen.

G. VEILLON	Président
M. BERTHOD	
J. CROWLEY	
J.C. LATOMBE	Examinateurs
R. MOHR	
G. TIBERGHIEN	

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

Président : Daniel BLOCH

Année 1987

Vice - Présidents : René CARRE
Jean-Marie PIERRARD

Professeurs des Universités

BARIBAUD Michel	ENSERG
BARRAUD Alain	ENSIEG
BAUDELET Bernard	ENSPG
BEAUFILS Jean-Pierre	ENSEEG
BESSON Jean	ENSEEG
BLIMAN Samuel	ENSERG
BLOCH Daniel	ENSPG
BOIS Philippe	ENSHMG
BONNETAIN Lucien	ENSEEG
BOUVARD Maurice	ENSHMG
BRISSONNEAU Pierre	ENSIEG
BRUNET Yves	IUFA
BUYLE-BODIN Maurice	ENSERG
CAILLERIE Denis	ENSHMG
CAVAIGNAC Jean-François	ENSPG
CHARTIER Germain	ENSPG
CHENEVIER Pierre	ENSERG
CHERADAME Hervé	UFR PGP
CHERUY Arlette	ENSIEG
CHIAVERINA Jean	UFR PGP
CHOVET Alain	ENSERG
COHEN Joseph	ENSERG
COUMES André	ENSERG
DARVE Félix	ENSHMG
DELLA-DORA Jean	ENSIMAG
DEPORTES Jacques	ENSPG
DOLMAZON Jean-Mar	ENSERG
DURAND Francis	ENSEEG
DURAND Jean-Louis	ENSIEG
FONLUPTJean	ENSIMAG
FOULARD Claude	ENSIEG
GANDINI Alessandro	UFR PGP
GAUBERT Claude	ENSPG
GENTIL Pierre	ENSERG
GREVEN Hélène	IUFA
GUERIN Bernard	ENSERG
GUYOT Pierre	ENSEEG
IVANES Marcel	ENSIEG
JAUSSAUD Pierre	ENSIEG
JOUBERT Pierre	ENSIEG
JOURDAIN Geneviève	ENSIEG
LACOUME Jean-Louis	ENSIEG
LESIEUR Marcel	ENSHMG
LESPINARD Georges	ENSHMG
LONGEQUEUE Jean-Pierre	ENSPG
LOUCHET François	ENSEEG
MASSE Philippe	ENSIEG
MASSELOT Christian	ENSIEG
MAZARE Guy	ENSIMAG
MOREAU René	ENSHMG
MORET Roger	ENSIEG
MOSSIERE Jacques	ENSIMAG
OBLED Charles	ENSHMG
OZIL Patrick	ENSEEG
PARIAUD Jean-Charles	ENSEEG
PAUTHENET René	ENSIEG
PERRET René	ENSIEG
PERRET Robert	ENSIEG
PIAU Jean-Michel	ENSHMG
POUPOT Christian	ENSERG
SAUCIER Gabrielle	ENSIMAG
SCHLENKER Claire	ENSPG
SCHLENKER Michel	ENSPG
SERMET PIERRE	ENSERG
SILVY Jacques	UFR PGP
SIRIEYS Pierre	ENSHMG
SOHM Jean-Claude	ENSEEG
SOLER Jean-Louis	ENSIMAG
SOUQUET Jean-Louis	ENSEEG
TROMPETTE Philippe	ENSHMG
VEILLON Gérard	ENSIMAG
ZADWORNY François	ENSERG

Professeur Université des Sciences Sociales (Grenoble II)

BOLLIET Louis

Personnes ayant obtenu le diplôme d'HABILITATION A DIRIGER DES RECHERCHES

BECKER Monique
BINDER Zdenek
CHASSERY Jean-Marc
COEY John
COLINET Catherine
COMMAULT Christian
CORNUEJOLS Gérard
DALARD Francis
DANES Florin
DEROO Daniel
DIARD Jean-Paul
DION Jean-Michel
DUGARD Luc
DURAND Robert
GALERIE Alain
GAUTHIER Jean-Paul
GENTIL Sylviane
PLA Fernand
GHIBAUDO Gérard
HAMAR Sylvaine
LADET Pierre
LATOMBE Claudine
LE GORREC Bernard
MADAR Roland
MULLER Jean
NGUYEN TRONG Bernadette
TCHUENTE Maurice
VINCENT Henri

Chercheurs du C.N.R.S

Directeurs de recherche 1ère Classe

CAILLET Marcel
CARRE René
FRUCHART Robert
JORRAND Philippe
LANDAU Ioan
MARTIN

Directeurs de recherche 2ème Classe

ALEMANY Antoine
ALLIBERT Colette
ALLIBERT Michel
ANSARA Ibrahim
ARMAND Michel
BINDER Gilbert
BONNET Roland
BORNARD Guy
CALMET Jacques
DAVID René
DRIOLE Jean
ESCUDIER Pierre
EUSTATHOPOULOS Nicolas
JOUD Jean-Charles
KAMARINOS Georges
KLEITZ Michel
KOFMAN Walter
LEJEUNE Gérard
MERMET Jean
MUNIER Jacques
SENATEUR Jean-Pierre
SUERY Michel
TEDOSIU
WACK Bernard

Personnalités agrées à titre permanent à diriger des travaux de recherche (décision du conseil scientifique)

E.N.S.E.E.G

BERNARD Claude
CHATILLON Catherine
CHATILLON Christian
COULON Michel
DIARD Jean-Paul
FOSTER Panayotis
HAMMOU Abdelkader
MALMEJAC Yves
MARTIN GARIN Régina
SAINTFORT Paul
SARRAZIN Pierre
SIMON Jean-Paul
TOUZAIN Philippe
URBAIN Georges

E.N.S.E.R.G

BOREL Joseph
CHOVET Alain
DOLMAZON Jean-Marc
HERAULT Jeanny

E.N.S.I.E.G

DESCHIZEAUX Pierre
GLANGEAUD François
PERARD Jacques
REINISCH Raymond

E.N.S.H.G

BOIS Daniel
DARVE Félix
MICHEL Jean-Marie
ROWE Alain
VAUCLIN Michel

E.N.S.I.M.A.G

BERT Didier
COURTIN Jacques
COURTOIS Bernard
DELLA DORA Jean
FONLUPT Jean
SIFAKIS Joseph

E.F.P.G

CHARUEL Robert

C.E.N.G

CADET Jean
COEURE Philippe
DELHAYE Jean-Marc
DUPUY Michel
JOUVE Hubert
NICOLAU Yvan
NIFENECKER Hervé
PERROUD Paul
PEUZIN Jean-Claude
TAIB Maurice
VINCENDON Marc

Laboratoires extérieurs

C.N.E.T

DEMOULIN Eric
DEVINE
GERBER Roland
MERCKEL Gérard
PAULEAU Yves

ECOLE NATIONALE SUPERIEURE DES MINES DE SAINT-ETIENNE

Directeur : Monsieur M.MERMET
Directeur des Etudes et de la formation: Monsieur J. LEVASSEUR
Directeur des recherches : Monsieur J. LEVY
Secrétaire Général : Mademoiselle M. CLERGUE

PROFESSEURS DE 1ère CATEGORIE

COINDE Alexandre	Gestion
GOUX Claude	Métallurgie
LEVY Jacques	Métallurgie
LOWYS Jean-Pierre	Physique
MATHON Albert	Gestion
RIEU Jean	Mécanique-Résistance des matériaux
SOUSTELLE Michel	Chimie
FORMERY Philippe	Mathématiques Appliquées

PROFESSEURS DE 2ème CATEGORIE

HABIB Michel	Informatique
PERRIN Michel	Géologie
VERCHERY Georges	Matériaux
TOUCHARD Bernard	Physique Industrielle

DIRECTEUR DE RECHERCHE

LESBATS Pierre	Métallurgie

MAITRE DE RECHERCHE

BISCONDI Michel	Métallurgie
DAVOINE Philippe	Géologie
FOURDEUX Angeline	Métallurgie
KOBYLANSKI André	Métallurgie
LALAUZE René	Chimie
LANCELOT Francis	Chimie
LE COZE Jean	Métallurgie
THEVENOT François	Chimie
TRAN MINH Canh	Chimie

Personalités habilitées à diriger des travaux de recherche

DRIVER Julian	Métallurgie
GUILHOT Bernard	Chimie
THOMAS Gérard	Chimie

Professeurs à l'UER de Sciences de Saint-Etienne

VERGNAUD Jean-Maurice	Chimie des Matériaux et Chimie Industrielle

Résumé

La thèse est relative à la Vision par Ordinateur. La première partie analyse les méthodes utilisées dans le domaine, justifie l'existence de différents niveaux de représentation et de traitement de l'information visuelle, puis explicite les cinq niveaux (IMAGE, INDICES D'IMAGE, INDICES DE SCENE, OBJET et SCENE) que nous distinguons. La seconde partie décrit, du niveau IMAGE au niveau INDICES DE SCENE, une expérimentation de l'inférence de formes à partir des contours dans le domaine restreint d'objets solides du monde des blocs. Elle met l'accent sur la distinction entre INDICES D'IMAGE et INDICES DE SCENE, et se prolonge par la résolution d'un problème industriel : la dépalétisation. S'appuyant sur les résultats obtenus, la troisième partie expose comment la stéréovision et la couleur s'intègrent au sein des niveaux préconisés, et comment ils permettent d'atteindre le niveau OBJET pour des objets flexibles filiformes. Ces apports sont illustrés par une autre application au domaine industriel : l'identification et la localisation de fils électriques dans un contexte d'automatisation de la production des ensembles câbles-connecteurs.

Abstract

This thesis deals with Computer Vision. The first part analyses the main methods that are used in this area. It justifies the existence of different levels for the representation and the computation of visual information. It finally specifies the five levels (IMAGE, IMAGE FEATURES, SCENE FEATURES, OBJECT and SCENE) we distinguish. The second part describes a shape-from contour experiment from the IMAGE level to the SCENE FEATURES level, in the restricted area of "block world" rigid objects. It emphasizes the distinction between IMAGE FEATURES and SCENE FEATURES and then describes the solution of an industrial problem: the unloading of ordered stacks of boxes. Relying on the obtained results, the third part explains how stereo vision and color vision become integrated into the advocated levels and how they allow to reach the OBJECT level for flexible wire-shaped objects. These contributions are illustrated by a second application to the industrial area: the identification and the localization of electrical wires in the context of Computer-Aided Manufacturing of wires and connectors sets.

Mots clés

Vision par ordinateur - Niveaux de représentation - Intelligence artificielle - Indices d'image - Indices de scène - Stéréovision - Couleur - Contours.

Remerciements

Je remercie :

- Gérard VEILLON (Professeur INPG), Directeur de l'ENSIMAG, pour m'avoir fait l'honneur de bien vouloir présider ce jury,
- Messieurs Marc BERTHOD (Directeur de Recherche INRIA à Sophia-Antipolis) et Roger MOHR (Professeur INPL), pour l'oeil critique qu'ils ont bien voulu porter sur ce manuscrit, pour les remarques qu'ils m'ont adressées et pour avoir accepté de juger ce travail,
- Messieurs James L. CROWLEY (Professeur INPG), Jean-Claude LATOMBE (Professeur INPG) et Guy TIBERGHIEN (Professeur UER Psychologie de Grenoble) pour avoir accepté de faire partie du jury de cette thèse,

- Monsieur Jean-Claude LATOMBE, Directeur Scientifique de cette thèse et point initial de ces travaux de recherche, qui a su me révéler mon goût pour la recherche. Je lui suis tout aussi reconnaissant des conseils qu'il a su me donner, et de la confiance qui m'a accordée,
- Messieurs Jose-Luis GORDILLO, Radu HORAUD, Gérard MEZIN et tout spécialement Augustin LUX pour les nombreuses discussions riches d'intérêt que nous avons eues ensemble durant ces années de recherche,
- Le groupe des chercheurs en Intelligence Artificielle, Robotique et Vision par Ordinateur, et plus généralement le laboratoire LIFIA, pour l'environnement de travail qui m'a été offert,

- Le CNRS et l'ENSIMAG, Messieurs Jean-Claude LATOMBE, Christian LAUGIER et Gérard VEILLON, qui ont rendu possible le financement de ces travaux,

- Messieurs Philippe CALOUD, Jose-Luis GORDILLO, Augustin LUX, Pierre PUGET, Olivier RAOULT, Fano RAMPARANY, Christian de SAINTE MARIE, Guy TIBERGHIEN et Madame Jocelyne TROCCAZ, pour avoir contribué à améliorer cette rédaction

- Messieurs Didier AUBERT et Eric DEMAZEAU, ainsi que Madame Jocelyne TROCCAZ pour leurs apports respectifs à la présentation de ce rapport,

- Toute ma famille, et particulièrement Sophie, ma femme, à laquelle je dédie ce manuscrit.

Sommaire

Liste des Figures

Introduction

Il n'existe pas de règles bien établies pour la décomposition du problème global de la vision en étapes successives entre l'image et l'interprétation finale. La définition des structures de données formant ces niveaux intermédiaires est pourtant un choix stratégique qui conditionnera le potentiel du système.

- Il conditionne le potentiel d'application. Plus la description de l'image est riche (types variés), et plus ses éléments correspondent à des éléments physiquement significatifs, plus il sera facile de résoudre un problème de vision à l'aide du système. Dans une bonne décomposition doivent apparaître des structures de données intermédiaires "utiles", c.à.d. correspondant à une réalité physique en vue, éventuellement, de la réalisation d'un réflexe visuel, ou significatives dans le dialogue avec un utilisateur.
- Il conditionne les possibilités de réalisation. La difficulté des problèmes algorithmiques posés par la construction des niveaux successifs conditionne les possibilités de réalisation et l'efficacité de l'ensemble. Une bonne décomposition est telle que le problème de construction d'un niveau de données à partir du niveau précédent admet des solutions algorithmiques efficaces.

Augustin Lux (1985)

Introduction

LE CONTEXTE :
un Système Intelligent opérant en Univers Réel

Un système intelligent opérant en univers réel doit être doté d'autonomie pour l'acquisition d'informations, pour la prise de décision et pour l'action. Il doit être capable d'exercer ces facultés dans un environnement complexe qui évolue avec le temps, et dont il ne possède qu'un modèle incomplet. La vision constitue l'une des plus puissantes sources d'information dont un système puisse disposer pour acquérir des informations sur son environnement. Dans ce dessein, le système de vision doit non seulement capter l'information brute (processus sensoriels), mais encore analyser, interpréter et comprendre les images constituées par cette information (processus perceptifs).

C'est dans ce cadre d'étude et de conception d'un système intelligent opérant en univers réel (objectif d'ensemble de l'équipe "Intelligence Artificielle, Robotique et Vision par Ordinateur" du "Laboratoire d'Informatique Fondamentale et d'Intelligence Artificielle") que s'inscrivent les travaux en **Vision par Ordinateur** ici présentés.

LE DOMAINE :
le Traitement Informatique des Images

Traitement d'Images, Reconnaissance des Formes, Vision par Ordinateur : trois disciplines qui sont souvent confondues au sein d'un même domaine : le traitement informatique d'images numérisées.

La principale raison à la continuelle ambiguité entre ces domaines est bien sûr issue de la communauté d'intérêt pour le traitement informatique (au sens propre

du "Traitement de l'Information") des images, mais aussi à la similarité d'ensemble des fins recherchées : Concourir à EXHIBER jusqu'à COMPRENDRE l'INFORMATION contenue dans des IMAGES.

Cependant, il ne devrait pas y avoir d'équivoque possible quant aux différences entre les moyens utilisés et les objectifs précis de chacun d'eux. Approchant l'image par des outils de Traitement du Signal, le **Traitement d'Images** transforme une image initiale en d'autres images mettant mieux en évidence certaines informations recherchées. La **Reconnaissance des Formes** restreinte au domaine de la Vision, cherche à caractériser des figures à l'aide d'outils mathématiques.

En admettant la validité de ces deux "définitions", quelle est alors la définition de la **Vision par Ordinateur?**

LA DISCIPLINE : la Vision par Ordinateur

> **VISION PAR ORDINATEUR : un ensemble de méthodes, spécifiques mais capables d'interagir de manière cohérente pour matérialiser un objectif commun : la construction d'un système capable de remplir artificiellement les fonctions du système visuel humain,** des fonctions les plus sensorielles de l'oeil aux fonctions les plus raisonnées auxquelles participent les aires visuelles du cortex, **pour aboutir à une description symbolique de la scène contenue dans le flot initial des images.**

L'objectif final visé est ambitieux. La Vision par Ordinateur en tant que discipline à part entière n'en étant qu'à ses "balbutiements", une définition quelle qu'elle soit ne peut être qu'entourée d'un certain flou. Tout comme en ontogénétique où il est courant d'étudier l'oeil comme faisant partie du cerveau, oeil et cerveau sont, dans notre définition, intimement liés au point d'être indissociables. Cependant, nous pensons que les processus mis en oeuvre doivent finalement permettre une "Compréhension de Scènes", réalisée en deux phases, l'une plutôt sensorielle, l'autre plutôt perceptive :

- Atteindre une description géométrique en termes de volumes des objets observés dans la scène (phase d'"**Analyse d'Images**"). A ce propos, il apparaît que le système visuel humain permet de décrire une scène observée en formes tridimensionnelles élémentaires, d'établir des relations spatiales entre ces formes, et même d'approcher le modèle de la scène sans nécessairement posséder de connaissances particulières sur celle-ci.

- Permettre une éventuelle désignation par nom des objets observés dans la scène, et aboutir à une description relationnelle et temporelle de la scène (phase de "**Compréhension de Scènes**" proprement dite). A ce stade d'étude, la prise en compte de connaissances autres que visuelles est nécessaire.

Des systèmes de vision combinant ces deux composantes existent déjà, car les recherches et les réalisations sont très nombreuses, mais... tous ces systèmes fonctionnent sur des domaines d'application restreints. A chaque nouvelle conception, on repart souvent à zéro. De même il existe des techniques connues et efficaces qui s'affinent et se spécifient de jour en jour mais... qui ne s'efforcent chacune qu'à se concentrer sur l'étude d'une caractéristique donnée du processus de vision : les contours, les ombres, la texture, la binocularité, etc...

Est-il alors pensable actuellement de satisfaire simultanément généralité, compétence et efficacité au sein d'un même système de Vision par Ordinateur? Le système idéal, fondé sur une base universelle, n'existe pas, du moins pas encore. Et il reste à chacun d'entre nous l'espoir, sinon de trouver tous les fondements de la Vision par Ordinateur, du moins, dans un premier temps, de découvrir quelques lois et théorèmes qui les régissent.

LA THESE :
analyse des Niveaux de Représentation de la Connaissance contenue dans une Image

> L'objet de ce document est l'analyse des niveaux cognitifs de représentation et de traitement de l'information dans un système de Vision par Ordinateur. **La thèse défendue est la suivante : même dans les cas de compréhension de scène paraissant les plus simples, et indépendamment des techniques utilisées, il est nécessaire de distinguer explicitement plusieurs "Niveaux de Représentation " entre celui qui décrit l'image initialement captée et celui, tant convoité, qui décrit symboliquement la scène observée.** Cette distinction n'est pas arbitraire mais repose sur de nombreux critères. En particulier, elle traduit un besoin de représenter une évolution discrète des notions d'"abstraction" et de "décentration", notions progressives qui caractérisent la succession des différentes structures élaborées au cours du processus de compréhension de scène.

L'objectif poursuivi conduit à prendre en compte un ensemble initial peu ou mal structuré de très nombreuses données souvent entachées d'erreur, et dont la plupart sont non significatives si elles sont prises isolément. Le problème de base consiste à interpréter ces données, afin d'en extraire des informations pertinentes au niveau de la description de scène.

Sur ce canevas, nous avons étudié les différents niveaux cognitifs nécessaires à l'interprétation d'images initialement reçues par des capteurs. Quatre principaux niveaux de représentation ont été distingués pour traduire les degrés d'**abstraction** (passages progressifs du numérique discret de l'image initiale au symbolique de la description finale) et de **décentration** (passage d'un repère initial lié à l'image à un repère absolu de la scène réelle observée) atteints dans le déroulement du processus d'interprétation de l'"IMAGE" en l'"OBJET". Et ceci plus précisément

dans le sens où un "OBJET" est une description en volumes et non le résultat d'une désignation. Un cinquième et dernier niveau, le niveau "SCENE", finalise le processus de compréhension de scènes.

La thèse défendue par ce document est concrétisée par des implantations effectives : des embryons de systèmes artificiels de vision, mais aussi des applications concrètes sur des problèmes industriels. Il s'agit ici de montrer qu'avec ce type de représentation, de tels systèmes "voient bien". Mais au-delà de ces résultats pratiques, nous avons aussi essayé d'évaluer le caractère de généralité des concepts introduits, en nous interrogeant de façon informelle sur le pourquoi de la nécessité à distinguer différents niveaux de représentation. En relation avec cette question, nous utilisons fréquemment, avec des sens qui nous sont propres, les termes d'"abstraction " et de "décentration". La simple évocation de ces deux termes montre combien, pour étudier notre sujet, nous avons eu besoin de nous évader du cadre de la Vision par Ordinateur pour nous tourner vers deux domaines où nous avons trouvé quelques inspirations : la psychologie et de la neurophysiologie.

LA DEMARCHE : Conceptualisation - Expérimentation - Application

La méthodologie que nous avons adoptée s'appuie sur une indispensable démarche expérimentale. Elle consiste à concevoir, à réaliser et à expérimenter des logiciels destinés à résoudre des problèmes réels, propres à révéler et à illustrer des aspects importants de la Représentation par Niveaux en Vision par Ordinateur. C'est ainsi que toute tentative pour conceptualiser et formaliser le processus de vision nous conduit à un besoin impératif d'expérimentation.

Réciproquement, c'est en expérimentant deux techniques bien connues, l'"inférence de forme à partir de contours" et l'"inférence de forme à partir de la stéréovision simple", qu'il nous est apparu essentiel de distinguer (et aussi d'essayer de formaliser) deux niveaux de représentation fondamentaux : les **"INDICES D'IMAGE"** (agrégats de données sensorielles, résultats d'une sensation) et les **"INDICES DE SCENE"** (entités locales de l'univers réel, résultats d'une perception).

Si le point de vue théorique de la Vision par Ordinateur est important, l'enjeu pratique l'est sans doute tout autant. En effet, satisfaction mise à part à remplir des fonctions incluant celles du système visuel humain (quand on y sera arrivé!), et indépendamment même des aspects économiques sous-jacents à la réussite d'une telle entreprise, il faut que le système de vision puisse s'intégrer dans un ensemble plus général, tout comme le système visuel humain fait partie de l'être humain et est utile à l'homme. C'est pourquoi, approfondissant les expérimentations et restreignant les domaines d'application des systèmes développés, nous avons aussi montré que les concepts défendus peuvent être exploités en milieu industriel.

LE RAPPORT :
sa constitution et sa lecture

Composé de trois parties incluant plusieurs études bibliographiques, le rapport présente la thèse développée de façon académique, puis décrit les deux expérimentations effectuées. La figure Orga montre les contributions respectives des différentes parties vis-à-vis de l'axe conceptualisation - expérimentation - application. Succintement, l'enchaînement des parties se présente comme suit :

En tout premier lieu, un synopsis dégage linéairement le cheminement de notre pensée au fur et à mesure de l'avancée dans les différents chapitres, en tentant d'en extraire les points forts. Quoique les trois parties soient suffisamment autonomes pour être consultées séparément, l'impression générale sera d'autant mieux rendue que le manuscrit sera lu séquentiellement.

La partie A du rapport expose **la nécessité d'expliciter différents niveaux de représentation intermédiaires au sein d'un système de Vision par Ordinateur.** Elle présente la Vision par Ordinateur comme un ensemble structuré de nombreuses techniques différentes, puis justifie l'existence des différents niveaux de représentation vis-à-vis du problème même de la Vision par Ordinateur, mais aussi par rapport au système visuel humain. Finalement, elle expose notre approche du problème en présentant les niveaux IMAGE, INDICES D'IMAGE, INDICES DE SCENE, OBJET et SCENE.

La partie B décrit **une première expérimentation ayant trait à l'inférence de formes à partir de contours, dans le domaine restreint d'objets solides du "monde des blocs".** Traitant d'images noir-et-blanc, elle met tout spécialement l'accent sur la distinction entre deux niveaux particuliers : le niveau INDICES D'IMAGE et le niveau INDICES DE SCENE. Suivant notre méthodologie, elle montre en outre ce qu'apporte la distinction entre ces niveaux dans le cas concret d'une application industrielle concernant la localisation de paquets sur une palette.

La partie C concerne **une deuxième expérimentation pour l'inférence de formes à partir de la stéréovision simple couleur, dans le domaine restreint des objets flexibles filiformes.** Elle montre pour cette seconde technique comment s'intègre la notion des niveaux, mais aussi quels sont les apports engendrés par l'utilisation d'images couleur. Du niveau IMAGE au niveau OBJET, elle détaille les différentes méthodes développées. Là aussi, la présentation s'achève sur la description d'une application industrielle : la localisation et l'identification de fils électriques, dans un contexte d'automatisation de la production d'ensembles câbles-connecteurs.

Finalement, une conclusion présente, sous la forme d'une discussion, plusieurs des idées que nous avons au sujet de la structure globale d'un système de Vision par

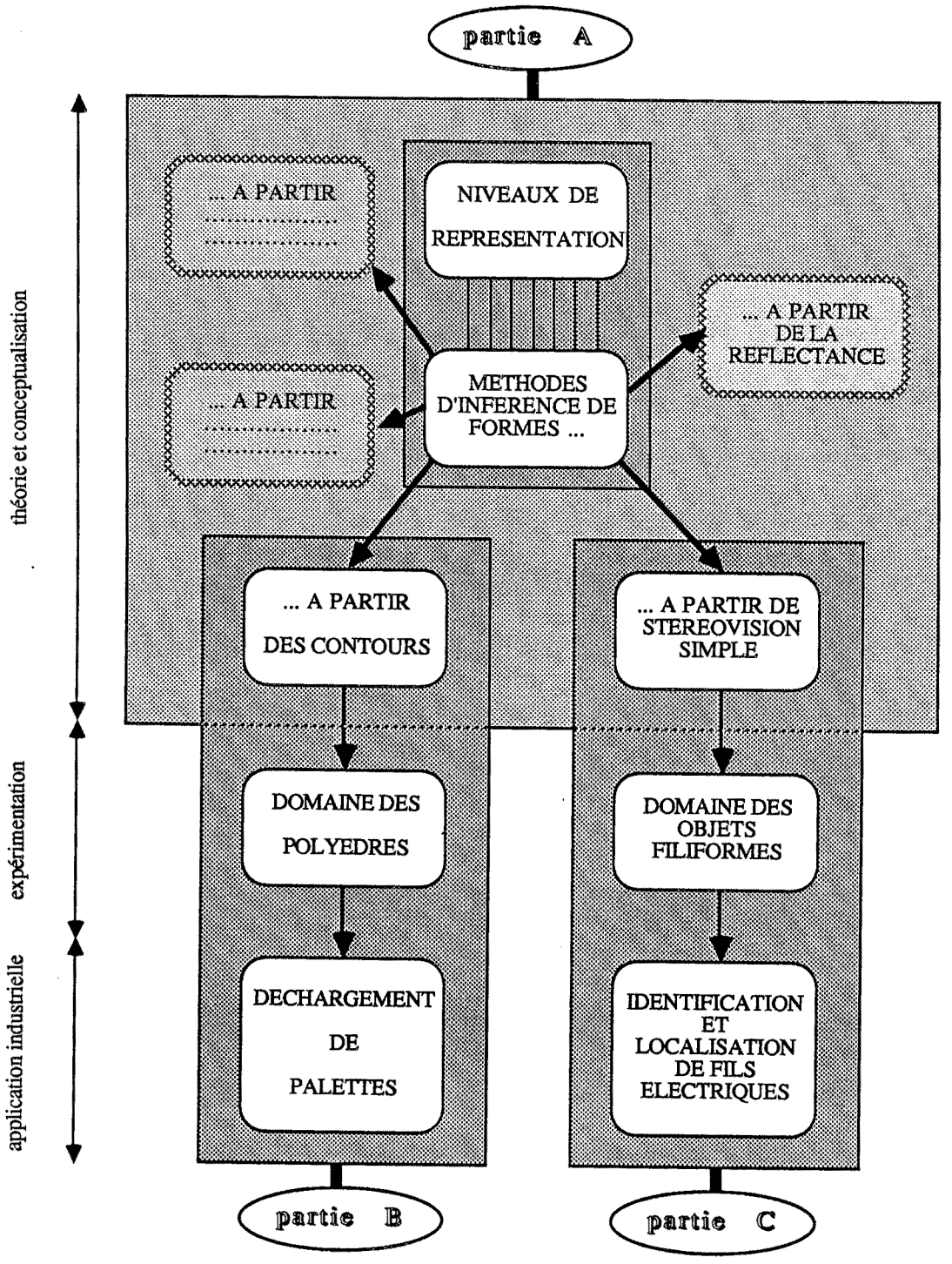


Figure Orga: Organisation du manuscrit

Ordinateur, en particulier par rapport aux problèmes du contexte, de la désignation et de la reconnaissance, puis développe nos vues prospectives.

POST-SCRIPTUM : la vision pour un système intelligent est naturellement 3-D

Nous n'avons pas fait allusion au terme "tridimensionnel". Les systèmes de Vision par Ordinateur ne seraient-ils pas "tridimensionnels"? Ou bien au contraire le seraient-ils tous? Etant donnés les objectifs poursuivis par la Vision par Ordinateur, ainsi que le type de représentation de la scène observée auquel aspirent tous ces systèmes, il est nous paraît clair qu'ils ne peuvent exister qu'à la condition nécessaire d'être *au moins* tridimensionnels! C'est la raison pour laquelle il nous arrive fréquemment (et souvent volontairement) d'omettre l'attribut "tridimensionnel ".

Synopsis

Synopsis

PARTIE A

Comme toute discipline encore trop jeune la Vision par Ordinateur a déjà été l'objet d'évolutions sensibles, démontrant par là même un certain manque de maturité : l'émerveillement devant les premiers résultats expérimentaux, qui laissaient espérer une résolution rapide au problème de "l'oeil du robot" a, après réflexion et déceptions, laissé place à des interrogations plus précises sur le comment du processus de vision. En effet, comme il l'est de toute discipline expérimentale, le succès d'une expérience visuelle en domaine restreint ne permet pas de démontrer une "loi de vision", mais ne peut que la suggérer a priori ou la confirmer a posteriori.

A.I Vingt années après ses débuts, la Vision par Ordinateur a pris un virage. Il ne s'agit plus de construire des systèmes de vision par extension (généralisation des expérimentations) d'un système donnant de bons résultats pour un domaine limité, jugeant ainsi empiriquement la puissance d'un système par rapport à la complexité réelle des scènes qu'il sait "comprendre". Il est désormais clair que la nature de l'information visuelle est différente selon que l'on s'attache à telle ou telle particularité intrinsèque à une image, par exemple l'illumination ou encore la texture. C'est ainsi que la tendance des recherches actuelles consiste à se concentrer sur un aspect particulier de la vision et de pousser l'étude au maximum des possibilités implicitement contenues par la caractéristique étudiée. L'étude de chaque caractéristique conduisant à des méthodes qui souvent lui sont propres, les embryons de systèmes ainsi développés sont plutôt jugés en fonction de leur capacité visuelle qu'en termes de complexité du domaine étudié. La systémique du problème de vision se trouve alors modifiée car ce n'est que de la conjonction et de la coopération de ces modules spécialisés que peut naître un système capable de "comprendre" les scènes qu'il observe.

L'ensemble des méthodes étudiées pour l'extraction d'informations 3-D d'une ou plusieurs images se partitionne en quatre sous-ensembles : méthodes d'inférence de formes monoculaires, méthodes d'inférence de formes multioculaires, méthodes par raisonnement direct dans l'espace tridimensionnel et méthodes actives de formation d'images tridimensionnelles. Les méthodes ne traitent pas des informations de même nature, ni ne fournissent des résultats de natures identiques (informations de surface et/ou de profondeur, en n'offrant que des modèles partiels de la scène). Finalement, aucune d'entre elles ne travaillant directement sur l'image (par exemple la vision stéréoscopique n'est en général pas réalisée sur la base de correspondance entre pixels), on est tout naturellement amené à se poser des questions sur la nature des représentations sur lesquelles elles opèrent et sur celle des descriptions qu'elles permettent d'engendrer. C'est ainsi que nous avons étudié la représentation de l'information visuelle à différents niveaux, sujet de la thèse défendue ici.

A.II La représentation de l'information visuelle et sa structuration en niveaux est un des problèmes centraux de la Vision par Ordinateur et plus encore de la compréhension de scène. Il est classique, encore que souvent artificiel et toujours difficile, de distinguer plusieurs niveaux d'opérations dans les phénomènes visuels : les approches courantes du problème utilisent, de l'image proprement dite à la description de scènes finale, quatre OU cinq niveaux de représentation.

Optimistes, nous pensons par ailleurs que les domaines de la psychologie et de la neurophysiologie peuvent contribuer non seulement à comprendre le fonctionnement du système visuel humain, mais aussi à guider l'élaboration d'un modèle informatique de la vision capable d'effectuer une compréhension de scène.

Réciproquement, nous espérons que la conception et la réalisation de modèles informatiques du processus de vision puissent contribuer à une meilleure compréhension du système visuel humain, en simulant des modèles dérivés de ces deux autres disciplines pour éprouver leur fiabilité et leur efficacité. Analysé par les neurophysiologues et les psychologues, chacun suivant un axe spécifique, le processus de compréhension de scènes visuelles chez l'homme cautionne l'existence de niveaux intermédiaires entre celui de l'image perçue par la rétine et celui de la description consciente de la scène tridimensionnelle.

A.III Bien que les méthodes étudiées ne soient pas toutes héritées de caractéristiques du système visuel humain, les premières justifications à l'existence de différents niveaux de représentation se trouvent pourtant sur le terrain de l'analogie avec la vision humaine.

- Les résultats obtenus dans les domaines de la psychologie et de la neurophysiologie constituent-ils une justification fondamentale à l'existence de différents niveaux pour un système de vision informatique? Une analogie source d'inspiration sûrement puisque certaines caractéristiques utilisées en Vision par Ordinateur sont identiques, mais une justification, pas vraiment. Car si le point commun majeur entre les deux types de vision considérés réside dans une égalité des "fins" recherchées, rien ne prouve que les méthodes informatiques doivent être identiques aux "moyens" utilisés par l'être humain, et

inversement.

- La complexité du processus de vision provient du fait que de nombreux facteurs interviennent simultanément dans l'image, parfois se masquant l'un l'autre, parfois se confondant. Comme tout phénomène complexe, le processus de vision ne peut être compris par une simple extrapolation des propriétés de ses composantes élémentaires qui sont ici les pixels. C'est pourquoi la seconde justification (justification informatique de complexité) à l'existence de niveaux intermédiaires traduit le besoin de décomposer le problème en sous-problèmes "élémentaires", pour répondre à des soucis de combinatoire et d'efficacité.

Les autres justifications sont d'ordre plus pratique, voire de bon sens, mais sont sans doute les plus importantes :

- Nécessité de représenter des passages très progressifs autant dans l'échelle des abstractions de représentation que dans celle de la décentration des repères de description (en passant d'un repère lié à la caméra à un repère lié à l'objet). Ne pouvant matériellement disposer d'une infinité de niveaux pour représenter ces processus continus, il s'agit de choisir ceux des niveaux qui sont les plus significatifs.

- Niveaux d'étude et communication avec d'autres domaines. Prenons le cas d'un système de vision intégré dans un système plus général, un robot de maintenance par exemple. Même s'il peut, comme tout système, se suffire à lui-même, le système de vision doit pouvoir dialoguer avec le "cerveau", mais aussi avec les autres "sens" du robot : c'est ainsi que les modèles perceptifs issus du processus de vision doivent pouvoir être utilisables pour la planification et le contrôle des actions du robot.

- Contrôle de la multiplicité des sources d'information et des natures différentes de ces informations. Chaque niveau développé doit être un potentiel lieu de partage, d'échange et de fusion d'informations entre méthodes étudiant des caractéristiques visuelles différentes, puisque chacune de ces méthodes constitue un élément de base du système de vision.

En regard des justifications à l'existence de niveaux intermédiaires, et à l'observation des résultats obtenus dans nos expérimentations, nous avons été conduits à effectuer la distinction fondamentale entre INDICES D'IMAGE (agrégats de données sensorielles) et INDICES DE SCENE (entités locales de l'univers réel). Sur les mêmes critères, nous définissons une suite de cinq niveaux de représentation adaptés à un objectif de compréhension de scène :

- le niveau IMAGE, qui traduit une valeur de signal en chaque point d'une image,

- le niveau INDICES D'IMAGE, qui rend explicite l'information importante contenue dans l'image, et particulièrement les changements d'intensité lumineuse, leurs distribution et organisation géométrique,

- le niveau INDICES DE SCENE, qui rend non seulement explicites l'orientation et l'apparence des surfaces visibles des objets en scène, mais aussi des relations spatiales symboliques (en adjacence et/ou en profondeur relatives) entre ces surfaces,

- le niveau OBJET, qui décrit les formes et leur organisation spatiale dans un repère lié à l'espace réel

- le niveau SCENE, qui constitue une description relationnelle et temporelle de la scène.

La distinction entre ces niveaux, conditionnée par les concepts d'abstraction et de décentration progressives, nous amène à choisir des modèles de représentation adéquats à chaque niveau, mais aussi à nous préoccupper des points très importants qui caractérisent la dynamique du système de Vision par Ordinateur. C'est ainsi par exemple que nous sommes conduits à construire des procédures intra-niveaux pour enrichir la structure d'un niveau par rapport à son exploitation probable. Par ailleurs, l'existence de différents paliers pose le problème des moyens de communication qui les relient. Les liaisons inter-niveaux se partitionnent schématiquement en deux catégories : les algorithmes d'"inférence" permettant à partir d'informations d'un niveau donné d'obtenir des informations d'un niveau d'abstraction supérieur (processus ascendants) et les procédures dites de "contrôle", susceptibles d'être utilisées pour diriger le processus de compréhension de scène dans sa globalité, mais aussi destinées à organiser localement l'extraction et l'interprétation des indices d'un niveau inférieur (processus descendants).

PARTIE B

Dans ce cadre conceptuel, nous avons d'abord étudié les niveaux indices d'image linéaires et indices de scène en considérant les méthodes d'inférence de forme à partir des contours. Premières méthodes (d'un point de vue historique) à avoir fourni des résultats tangibles en Vision par Ordinateur, ces méthodes trouvent un terrain d'expérimentation privilégié dans le domaine du "monde des blocs".

B.I On peut dire que le monde des blocs, de façon générale, a constitué pendant toute une époque LE terrain privilégié de toute première expérimentation en

Vision par Ordinateur. La première des raisons en est que les nombreuses restrictions géométriques imposées aux objets de ce domaine (polyèdres) conduisent à des réalisations logicielles rapides et facilitent la reconnaissance d'objets. La seconde reflète simplement une vérité selon laquelle, bien qu'éloigné du monde réel et des scènes d'extérieur, le monde des blocs permet une modélisation approchée de nombreux objets quotidiens.

Les différentes approches d'étude de ce monde (géométrique numérique, géométrique symbolique, et plus récemment algébrique), si elles n'ignorent pas (parfois même implicitement) la différence entre indices d'image et indices de scène, ne distinguent toutefois pas explicitement les structures créées à ces deux niveaux.

B.II Il est vrai que la distinction entre niveaux, si elle apparaît évidente pour des scènes réelles, le semble moins dans le cas d'études en domaine restreint : il est en effet couramment admis que la connaissance de restrictions sur les objets étudiés implique des contraintes fortes sur les indices d'image, ce qui suffit alors à interpréter la scène observée. Il nous semble cependant indispensable d'offrir à un système de vision (et même s'il fonctionne dans le monde des blocs) les moyens de contrôler le bien-fondé (i.e. la "réalité scénique") des interprétations qu'il effectue.

L'un de ces moyens est précisément la distinction entre les structures créées aux niveaux indices d'image et indices de scène. En effet, les particularités (en abstraction et en décentration) spécifiques de chacun de ces types d'indices mises à part, une telle différenciation prévient, par exemple, de l'incomplétude des interprétations d'indices d'image en indices de scène. Autorisant une multiplicité explicite de l'interprétation d'une structure d'indices image initiale, le système est prémuni contre toute éventualité de non-reconnaissance et il lui est aussi possible de comprendre des images "à sens multiples" ou encore certaines illusions.

B.III L'approche suivie pour cette première expérimentation de compréhension de scènes composées d'objets solides, est fondée sur l'analyse d'une image digitalisée unique noir et blanc avec extraction, puis interprétation d'indices d'image rectilignes et simples (donc des segments de droite). Les méthodes d'interprétation utilisées sont constituées d'une analyse contextuelle locale accompagnée d'une technique de propagation de contraintes, le plus indépendamment possible de tout contexte global d'environnement. Approchant ainsi les niveaux OBJET et SCENE, il est possible (mais d'un intérêt secondaire) de reconnaître des objets polyédriques présents dans une image à l'aide d'une banque de modèles prédéfinis, sur des simples critères de forme et grâce à l'utilisation de théorèmes de base en géométrie (comme la conservation de la convexité et du parallélisme).

B.IV Sur le plan pratique, cette méthode a été utilisée pour l'étude de faisabilité d'une application industrielle en Robotique : le déchargement de palettes constituées de paquets parallélépipédiques. A première vue, le problème posé semble pouvoir être résolu par des méthodes bidimensionnelles conventionnelles. Mais ce n'est vrai que dans une certaine mesure : en effet, la présence d'étiquettes sur les paquets, et surtout la forte texture des surfaces observées, due à l'utilisation d'em-

ballages déformables (typiquement en carton) justifie l'exploitation de l'approche consistant à distinguer les indices d'image et les indices de scène. Assurant un meilleur taux de reconnaissance que les systèmes bidimensionnels conventionnels, cette réalisation fournit en outre un exemple tangible d'utilisation industrielle d'un logiciel d'inférence de forme à partir de contours, opérant dans le monde des blocs.

PARTIE C

Après l'étude dans le monde des blocs, nous avons abordé les méthodes d'inférence de forme à partir de la stéréovision couleur simple, pour des scènes constituées d'objets filiformes flexibles. Dans un premier temps, il nous faut délimiter le champ d'action de ces nouvelles données par rapport aux différents niveaux de représentation préconisés.

C.I Se distinguant d'une image noir et blanc par la présence en chacun de ses pixels d'informations plus riches, l'image couleur peut soit être traitée de la même manière qu'une image noir et blanc, soit être le sujet d'analyses spécifiques. Quelles que soient les méthodes choisies, le type de représentation de l'information de couleur qui peut être utilisé est fortement conditionné par le niveau d'abstraction auquel le problème de vision est étudié : un modèle hérité de la transformation de Karhunen-Loeve est, par exemple, bien adapté à l'extraction des indices d'image alors qu'au niveau OBJET, c'est plutôt un modèle de Munsell qui représente le mieux l'information de couleur.

Tout comme la couleur, et au contraire de l'inférence de forme à partir des contours qui est étudiée depuis une vingtaine d'années, la stéréovision simple n'a que peu donné l'occasion de traiter les niveaux de représentation de l'information binoculaire. La raison principale en est que la stéréovision simple (pour laquelle deux capteurs rentrent alors en jeu et non plus un seul) est couramment abordée en séparant le problème en quatre sous-problèmes bien distincts : 1/ la détermination des types d'indices à rechercher dans les deux images de la paire stéréoscopique, 2/ la mise en correspondance entre indices des deux images, 3/ la détermination des positions relatives des deux caméras, et 4/ le calcul de la distance des objets observés aux caméras. C'est ainsi qu'en général, les niveaux présents dans les systèmes qui suivent ce schéma général ne font que refléter l'ordre temporel dans lequel les différents problèmes sont activés et résolus.

Il est fréquemment admis, par exemple, que dans le cas d'une stéréovision simple, un calcul de "disparité" (notion géométrique qui correspond effectivement chez l'homme à la notion de disparité rétinienne) suffit à localiser les objets de la scène observée. Cette méthode correspond à une vision de loin chez l'être humain pour une perception imprécise du relief de la scène. A l'opposé, nous croyons que le but essentiel de la stéréovision est d'acquérir une information de profondeur précise dans un espace restreint et proche du capteur. C'est pourquoi nous avons développé un capteur dont la simplicité de conception et la précision du calibrage autorisent à envisager une coopération bras-oeil efficace.

Le problème de recouvrement des informations tridimensionnelles étant résolu et notre modèle défini, il reste qu'un processus de mise en correspondance fondé

sur un calcul de disparité, appliquée à des caractéristiques des niveaux IMAGE et INDICES D'IMAGE, devient d'autant plus inefficace que l'angle formé par les deux capteurs est plus obtus ("stéréovision large"). Quels sont alors, pour de telles configurations, les indices à extraire dans les images de la paire stéréoscopique? Comment les mettre en correspondance? Ces deux interrogations reflètent les deux sous-problèmes (1/ et 2/) les plus délicats et les moins étudiés de la stéréovision. La forte disparité inhérente à la stéréovision large nous a tout naturellement conduit à raisonner avec des indices de niveaux élevés, et tout particulièrement avec ceux du niveau INDICES DE SCENE. En effet, et là encore mises à part leurs spécificités en abstraction et décentration (qui contribuent fortement à empêcher une explosion de la combinatoire de correspondance), ces indices peuvent par exemple guider le processus de correspondance, bien au-delà des effets de la traditionnelle contrainte dite "des lignes épipolaires".

L'approche que nous avons suivie se justifie d'autant plus lorsqu'il s'agit de traiter des objets "gauches" (notion définie par extension de la définition mathématique des surfaces gauches) et flexibles, qui requièrent des traitements bien plus raffinés que les objets du monde des blocs. Gauches, les contraintes sur les indices de scène n'ont pas toujours de traduction "évidente" au niveau des indices d'image. Flexibles, il n'est pas possible d'utiliser un modèle géométrique simple pour guider la recherche. C'est ce type d'objets que nous avons voulu traiter pour notre seconde expérimentation. Cependant, pour des raisons pratiques, nous nous sommes limités à l'étude des objets gauches et flexibles mais qui soient de plus filiformes.

C.II Fréquemment présente dans les images du monde réel, la classe des objets filiformes est étudiée dans des domaines aussi différents que la médecine et la robotique. Depuis près de cinq années, le sujet a donné lieu à plusieurs réalisations, bidimensionnelles ou tridimensionnelles, qui font chacune preuve d'ingéniosité. Notons à ce propos qu'autant les méthodes adoptées peuvent être différentes pour un problème de Vision par Ordinateur précis (comme par exemple l'identification et la localisation de fils électriques), autant les résultats obtenus sont souvent comparables.

C.III Abordant parallèlement les trois sujets de couleur, de stéréovision simple et de filiformité, la méthode que nous avons utilisée est constituée des éléments suivants : extraction d'indices d'image couleur (rectilignes ou courbes, simples ou doubles), interprétation de ces indices en INDICES DE SCENE, puis reconstitution des OBJETS pour finalement décrire la SCENE observée. Les types des indices d'image et des indices de scène recherchés constituent une extension presque immédiate de ceux de la première expérimentation (monde des blocs). Les méthodes d'interprétation des indices d'image en indices de scène relèvent encore d'une analyse contextuelle locale, accompagnée de l'utilisation de la propagation de contraintes. L'interprétation des indices de scène en objets fait usage d'une mise en correspondance qui bénéficie des informations de couleur et du guidage lié au modèle de stéréovision que nous avons adopté.

C.IV Les méthodes développées ont conduit au - mais ont aussi bénéficié du - développement d'une application sur un problème industriel. Le logiciel spécifique développé pour cette application analyse des scènes composées de fils électriques. L'identification et la localisation des fils utilisent une version simplifiée des méthodes élaborées pour l'inférence de formes à partir de la stéréovision couleur simple. Elles bénéficient en outre de connaissances spécifiques sur l'environnement des fils, qui accélèrent et optimisent les techniques employées.

Cette réalisation fait partie d'une étude de faisabilité d'une cellule d'automatisation de la production des ensembles câbles-connecteurs (étude effectuée sous contrat avec la société des Machines BULL - Belfort). Les fils électriques doivent être localisés pour être saisis par un robot, puis sertis et finalement insérés à leur emplacement adéquat dans un connecteur.

Cette première partie expose la thèse défendue, c'est-à-dire la nécessité d'expliciter différents niveaux de représentation au sein d'un système de Vision par Ordinateur. Elle présente la Vision par Ordinateur comme un ensemble structuré de nombreuses techniques différentes. Elle justifie ensuite l'existence de différents niveaux de représentation vis-à-vis du problème même de la Vision par Ordinateur, ainsi que par rapport au système visuel humain. Finalement, elle expose notre approche du problème fondée sur la distinction entre cinq niveaux de représentation : les niveaux IMAGE, INDICES D'IMAGE, INDICES DE SCENE, OBJET et SCENE.

Partie A

Vision Tridimensionnelle et Niveaux de Représentation

On peut dire: *"J'ai soif, je vais prendre un verre d'eau"* ou bien: *"Le degré insuffisant de mon hypothalamus a donné à mon cortex des instructions grâce auxquelles je me dirige vers un point d'eau"*. Ce sont deux manières de représenter une même réalité. Il y a donc une certaine autonomie des niveaux d'explication d'un phénomène qui rend souvent bien illusoire les tentatives de réduction d'un niveau au niveau inférieur.

Michel Imbert (1985)

Chapitre A.I

Méthodes en Vision par Ordinateur (V.O.)

Comme nous l'avons déjà dit dans l'introduction, la vision par ordinateur est par nature et relativement à ses ambitions essentiellement tridimensionnelle. L'intérêt porté à la perception de l'espace tridimensionnel et à la compréhension des scènes observées, est né dès l'apparition des premiers systèmes de vision artificielle. L'objectif était déjà d'obtenir des informations riches à partir d'un traitement informatique des images [ROBER-65].

Les principales méthodes connues pour l'étude de la vision tridimensionnelle peuvent être regroupées en quatre grandes catégories :

- les méthodes d'inférence de forme monoculaires,
- les méthodes par raisonnement direct dans l'espace 3-D,
- les méthodes d'inférence de forme multioculaires,
- les méthodes actives de formation d'images tridimensionnelles.

La nature des informations initiales que traitent ces méthodes (resp. 2-D, 2-D, Nx2-D, 3-D), et celle des informations qu'elles peuvent engendrer (resp. "2,5-D", 3-D, 3-D, 3-D) permettent d'établir une première distinction. Les deux premiers types

de méthodes analysent des images bidimensionnelles, obtenues de façon classique par une caméra de type quelconque. Elles s'efforcent essentiellement d'extraire de ces images des informations tridimensionnelles précises quant à *l'orientation* des surfaces observées dans la scène, et si cela s'avère possible (cas du raisonnement direct dans l'espace 3-D), quant à la localisation en profondeur des objets observés. A l'inverse, les méthodes des deux derniers types s'attachent surtout à *localiser en profondeur* les objets de la scène (et ce avec une très bonne précision) et peuvent aussi être utilisées dans le but d'obtenir l'orientation des surfaces constituant ces objets. Finalement, s'il n'existe pas, à l'heure actuelle, de système de Vision par Ordinateur anthropomorphe, la plupart des systèmes des trois premiers types sont inspirés de la vision humaine, ce qui n'est pas le cas des systèmes du dernier type.

Dans ce chapitre, nous effectuons un rapide, mais relativement exhaustif tour d'horizon de ces méthodes. Pour chacune d'entre elles, nous donnons une description du principe, des problèmes soulevés et des techniques développées. Nous reviendrons dans la suite de ce rapport sur deux de ces méthodes qui se rapportent plus spécifiquement à nos travaux : l'inférence de formes à partir des contours (partie B) et l'inférence de formes à partir de la stéréovision simple (partie C).

Pour plus de détail sur chacune de ces méthodes, le lecteur pourra se reporter aux références citées. Il trouvera aussi des synthèses suivant d'autres points de vue dans [BRADY-81], [COHEN-82], [CROWL-84] et [LUX-85a].

1 La naissance des Méthodes d'Inférence de Formes

Avant d'analyser les méthodes d'inférence de formes, il est bon de faire un retour en arrière dans les années 65-75, l'époque des premiers sytèmes de vision tridimensionnels.

C'est Roberts qui construit le tout premier système voué à la reconnaissance d'objets tridimensionnels, dans le domaine restreint du monde des blocs. La démarche qu'il suit [ROBER-65], purement numérique, tant du point de vue de l'extraction de droites (segmentation) que de la mise en correspondance entre droites extraites et modèles, montre rapidement ses limites.

Guzman montre alors que les droites détectées peuvent être interprétées et rendues explicites par un raisonnement symbolique, suite à l'observation suivante [GUZMA-68] : les "lignes" et les "jonctions" (ou "noeuds") de l'image sont les projections respectives des arêtes et des sommets des objets en scène. Il élabore un ensemble d'heuristiques sur les noeuds qui permettent à son système de connecter entre elles les régions susceptibles d'appartenir au même objet. L'approche purement séquentielle ascendante des systèmes de Roberts et de Guzman, comme tous les systèmes suivant cette même démarche, reste conditionnée par la qualité du processus de segmentation. **Mais, même dans le monde des blocs, il est impossible d'extraire parfaitement une droite dans l'image.** En effet, un tracé de droite ne peut être jugé parfait qu'après interprétation correcte (vis-à-vis d'un modèle par exemple), mais le succès d'une interprétation dépend de la qualité

parfaite de l'extraction de cette droite!

L'une des solutions partielles à ce problème est d'apporter aux programmes des éléments de connaissance sur leurs domaines d'étude. C'est ainsi par exemple, que le programme de Shirai [SHIRA-75] parvient à poursuivre dans l'image des droites incomplètement extraites. Chez Falk [FALK-72], l'aide à l'interprétation d'un "dessin de lignes" extrait de l'image se traduit par la connaissance du modèle d'objet.

A ce point des études, Huffman et Clowes font une remarque majeure : les "droites de contour" (observées lorsqu'un objet en cache un autre ou s'inscrit sur le fond de la scène) doivent être distinguées des "droites de connexion" (observées au lieu où deux surfaces d'un objet s'intersectent). Tous les deux exploitent ces observations en classifiant toutes les interprétations de lignes et de noeuds. Cette classification est facilitée par deux hypothèses simplificatrices : toutes les faces des objets sont planes et les sommets sont triédriques [HUFFM-71] [CLOWE-71]. Chaque interprétation de droite (étiquetage de ligne) est associée à des contraintes locales sur l'occupation spatiale des deux sommets correspondants. Les contraintes sur l'interprétation des noeuds de l'image sont propagées le long des lignes de l'image, sachant que la nature physique d'une arête plane est constante le long de cette arête. C'est sur ce schéma que Mackworth développe l'espace gradient[1] représentation qui lui permet de mieux exploiter la restriction de planarité des surfaces introduite par ses prédécesseurs : il devient alors possible de raisonner sur des propriétés de scène comme la convexité, la concavité, mais aussi sur la perpendicularité et l'uniformité [MACKW-73].

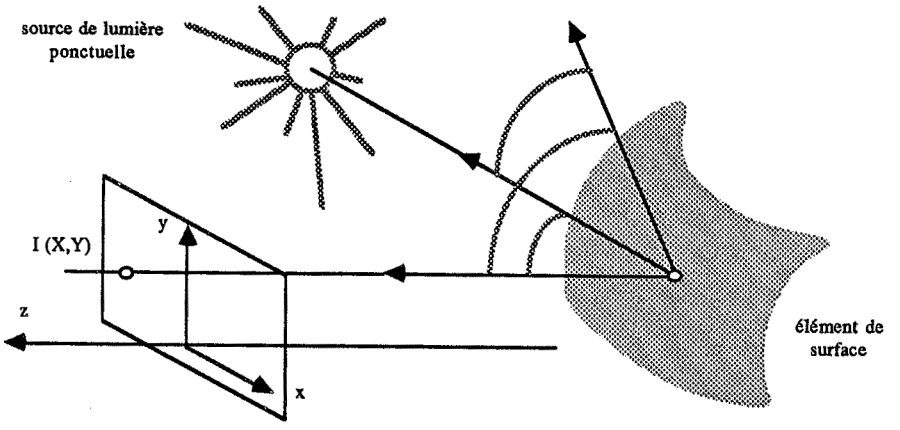
Jusqu'aux travaux de Huffman et Clowes, on peut parler de la technique utilisée comme de l'"inférence de formes à partir d'un dessin de lignes" ("shape from line drawings"). Waltz généralise l'éventail de classification en tenant compte des ombres et des "cassures" (droites d'intersection entre deux objets lorsque l'un cache l'autre) [WALTZ-75]. Ces travaux annoncent une séparation entre "méthodes d'inférence de forme à partir des contours" et "méthodes d'inférence de forme à partir des ombres ". Jusqu'à ces travaux y compris, les méthodes utilisées regroupaient dans une même étiquette les informations issues par exemple de l'illumination, des cassures, des occultations, ou encore du type d'arête. Développer de nouveaux systèmes consistait alors, tout en gardant la même démarche, à étendre (par généralisation des expérimentations) le domaine dans lequel les systèmes précédents fournissaient de bons résultats.

Et là sans aucun doute réside la différence essentielle entre la Vision par Or-

[1] **espace gradient** : représentation plane de l'orientation des surfaces visibles. L'orientation d'un élément de surface (x, y, z) est décrite par les deux paramètres de son gradient de surface (vecteur normal), communément appelés (P, Q). Si $-z = f(x, y)$ représente l'équation de l'élément de surface (cf. figure A.01.a), le vecteur normal à la surface est de la forme $(P, Q, 1)$ et le gradient (P, Q) de cette surface est défini comme :

$$P = \frac{\partial f}{\partial x} = \frac{\partial - z}{\partial x} \qquad Q = \frac{\partial f}{\partial y} = \frac{\partial - z}{\partial y}$$

(d'après [COHEN-82-1] (pp. 261 - 263))



A.01.a

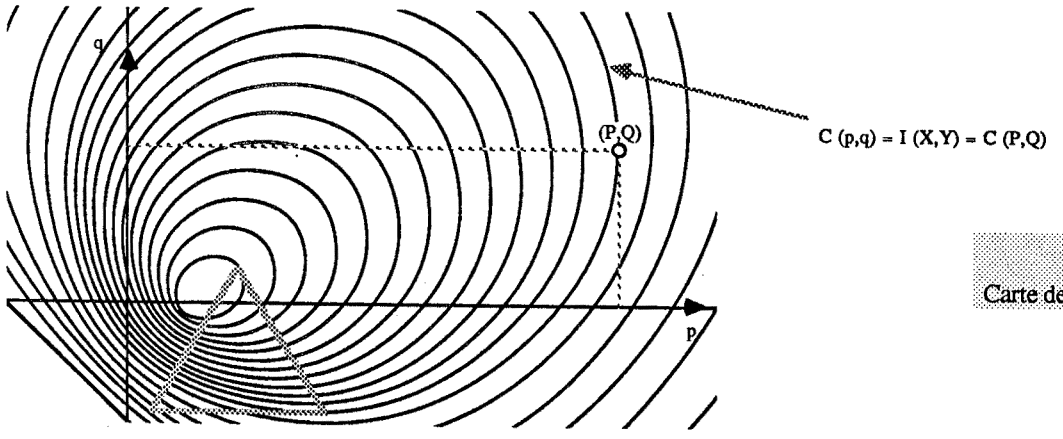
Modèle de formation d'une image

Si on suppose
- une illumination incidente parallèle et constante en chaque position de la surface
- une projection de la scène orthographique
- une réflection uniforme quelque soit l'angle de vue

alors, pour une configuration donnée, l'intensité I (X,Y) du point observé dans l'image correspond à un point (P,Q) de l'espace gradient qui appartient à une courbe continue C (p,q) appelée "courbe d'iso-intensité". Le lieu des points C (p,q) correspondant à une même intensité I (X,Y) observée est défini par l'équation:

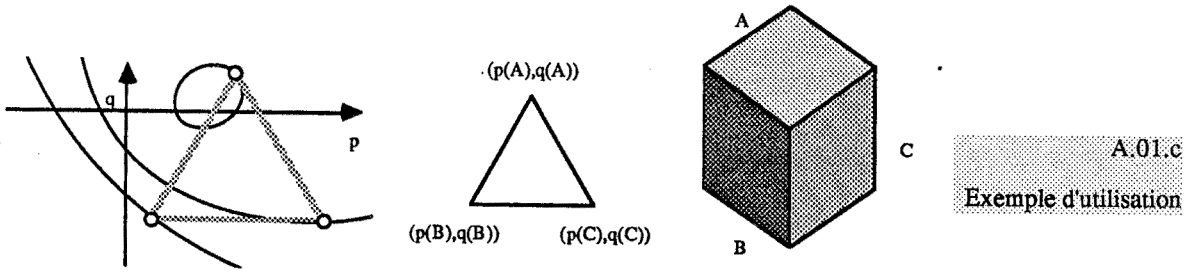
$$I(X,Y) = r \frac{Ps\,p + Qs\,q + 1}{\| (Ps, Qs, 1) \| \; \| (P, Q, 1) \|} = C(P,Q)$$

(Ps , Qs , 1) : vecteur qui pointe de la surface vers la lumière
(P , Q , 1) : orientation de l'élément de surface
r : constante de réflectance, liée à l'albedo de l'élément de surface



A.01.b

Carte de réflectance



A.01.c

Exemple d'utilisation

Figure A.01: Méthode d'Inférence de Formes à partir de la Réflectance

dinateur telle qu'elle était abordée à cette époque et telle qu'elle l'est maintenant. La tendance actuelle est la focalisation sur des thèmes qui peuvent être identifiés comme des modules du système visuel humain. Chacune des méthodes d'inférence de formes va donc s'attacher à l'étude d'un thème particulier. Dans chacun des cas, il s'agit d'*isoler* une source d'information particulière et de l'*exploiter* au maximum pour *montrer jusqu'où* il est possible d'aller dans l'interprétation de l'image initiale. **Cependant, si chacune de ces méthodes fait appel à un processus indépendant, il est important de noter que c'est la conjonction de plusieurs de ces méthodes qui permet d'aboutir à une description tridimensionnelle complète de la scène.** Les travaux actuels sont jugés plus en fonction de la capacité visuelle de la méthode utilisée que de la complexité du domaine que le système est capable d'étudier [BRADY-81] [MARR-82] [CROWL-84].

Cette démarche est particulièrement appropriée pour les méthodes d'inférence de forme monoculaires : l'analyse de l'information initiale ne peut permettre à elle seule d'inférer un modèle tridimensionnel complet de la scène sans le concours de connaissances extérieures. par exemple géométriques ou fonctionnelles, mais aussi plus "simplement" fournies par d'autres méthodes perceptives.

2 Les Méthodes d'Inférence de Formes Monoculaires (2-D → 2,5-D)

Les méthodes d'inférence de formes, qu'elles soient monoculaires ou multioculaires, fournissent des informations tridimensionnelles sur les surfaces présentes dans l'image, à partir d'informations bidimensionnelles. Cependant, aucune méthode d'inférence de formes monoculaires n'est capable de fournir la totalité des six coordonnées (trois en position, trois en orientation) qui déterminent parfaitement un élément de surface dans l'espace réel.

2.1 L'Inférence de Formes à partir des Contours

Disposant d'une image bidimensionnelle, il s'agit ici d'extraire les informations relatives au contour des objets (**lignes extérieures** définissant le **profil des objets**), mais aussi celles relatives à l'apparence des surfaces visibles des objets (**lignes intérieures au contour de l'objet**) ("shape from contours") [WITKI-82].

L'étude de ce type de méthode est la continuité historique de "l'inférence de formes à partir de dessins de lignes". La poursuite de cette tradition se manifeste dans les travaux où l'on se restreint à des objets composés de surfaces planes comme le font par exemple [DRAPE-81], [KANAD-81] ou encore [DEMAZ-84a] (première de nos expérimentations).

Kanade combine la notion d'espace gradient introduite par Mackworth et celle de l'étiquetage de lignes. Il propose deux notions supplémentaires de "parallélisme " et de "symétrie oblique" pour mieux contraindre les orientations de surface.

Draper, quant à lui, préfère effectuer un raisonnement symbolique de profondeur et/ou d'adjacence relatives entre les faces planes des objets.

A l'opposé, [SUGIH-79] et [BARRO-81], parmi d'autres, s'intéressent à l'interprétation de scènes composées d'objets "gauches" (i.e. possédant des surfaces courbes). Tous ces travaux, comme celui de [MALIK-85], font appel à un étiquetage destiné à appréhender un phénomène nouveau lié à la nature gauche des objets : certaines "lignes-image" observées ne correspondent à aucune arête réelle des objets.

C'est aussi de cette deuxième famille de méthodes que relèvent en partie les travaux de [DEMAZ-85a], traitant d'objets gauches et filiformes (deuxième de nos expérimentations).

Nous reviendrons sur ce type de méthodes (en montrant les techniques que nous avons utilisées) au cours des parties B et C, puisque les lignes de contours constituent *les* indices importants de nos deux expérimentations. Que ce soit pour traiter des objets gauches ou non, il apparaît que les méthodes d'inférence de forme à partir des contours sont non seulement assez immédiates à mettre en oeuvre et conduisent à des réalisations efficaces, mais encore qu'elles sont puissantes, au sens de leurs capacités visuelles [BINFO-81].

2.2 L'Inférence de Formes à partir des Ombres

On trouve aussi dans [BINFO-81] un intérêet pour l'étude de la caractéristique visuelle "ombres". Née à la scission de l'inférence de forme à partir des dessins de lignes, l'inférence de formes à partir des ombres ("shape from shadows") étudie l'apport fourni par la **détection des ombres** dans une image.

Waltz, en généralisant les travaux de Huffman et Clowes par l'introduction des étiquetages d'ombres [WALTZ-75] est le premier à considérer que les ombres d'une image ne sont pas forcément une gêne pour la compréhension d'une image. Se plaçant dans le cas d'une unique source d'illumination éloignée, Waltz montre que les ambiguités d'interprétation qui sont possibles sans source de lumière sont souvent facilement résolues dès que les ombres existantes sont prises en compte et interprétées.

C'est de ce type de méthodes que relèvent les travaux de [SHAFE-82]. Etant donné un dessin de lignes aux zones d'ombre identifiées, Shafer montre comment, sous l'hypothèse d'une projection orthographique, la forme des ombres permet d'engendrer des contraintes sur les orientations des surfaces mises en jeu.

La caractéristique "ombres" est particulièrement mise en valeur avec le système WIRESIGHT de [TSUJI-83] où elle constitue le fondement même du système. Ce système (décrit en détail dans la partie C) localise des fils électriques. Disposant d'une source de lumière ponctuelle et d'un plan de travail dont les géométries sont parfaitement connues, l'information de distance sur tous les points d'un fil est obtenue par mise en correspondance entre les projections sur l'image 1/ du fil lui-même et 2/ de son ombre projetée sur le plan de travail. Les éventuelles indéterminations

résiduelles sont résolues par l'utilisation séquentielle de plusieurs sources de lumière.

C'est aussi de cet esprit que relèvent les travaux de [KENDE-86], dans lesquels l'extraction des formes des objets est assurée par observation de l'évolution de l'ombre propre d'un objet pendant que la source de lumière ponctuelle se déplace. Cependant, malgré ces liens évidents avec l'étude de la caractéristique "ombre", l'esprit de la méthode utilisée est beaucoup plus proche de celui de la "stéréovision photométrique", une des méthodes d'inférence de formes à partir de la réflectance.

2.3 L'Inférence de Formes à partir de la Réflectance

Une image bidimensionnelle fournit souvent des informations présentant des ambiguités : pour pouvoir acquérir la forme tridimensionnelle des objets en scène, des hypothèses reliant les caractéristiques de l'image et celles de la scène sont nécessaires. Cette démarche est typique de l'inférence de formes à partir de la réflectance ("shape from shading"), qui est fondée sur l'observation de **variations d'intensité lumineuse** perçue sur les surfaces, en supposant connu un modèle de formation de l'image étudiée. Il est en effet possible, dans certains cas, d'estimer les modifications locales de l'orientation des surfaces par observation de tels changements d'intensité lumineuse sur ces surfaces.

C'est [HORN-77] qui, l'un des premiers, a travaillé sur un projet relatif à la capacité humaine d'inférer des formes à partir de la réflectance. Utilisant l'espace gradient (p,q) pour représenter les orientations des éléments de surface visibles, il montre comment la notion de "carte de réflectance" permet de relier la normale de l'élément de surface observé, les caractéristiques de réflectance, l'illumination ponctuelle ambiante, et la valeur d'intensité enregistrée au point correspondant dans l'image (cf. figure A.01.a). La figure A.01.b montre que sous les trois conditions suivantes,

1. illumination incidente *parallèle* et *constante* en chaque position de la surface,
2. projection de la scène *orthographique* (ces deux premières conditions correspondent au cas où la source de lumière et l'observateur sont tous les deux éloignés de l'objet observé), et
3. réflection *uniforme* quel que soit l'angle de vue (c'est en partiulier le cas des surfaces lambertiennes introduites par [WOODH-78]),

l'intensité observée en un pixel de l'image contraint son orientation de surface à se trouver sur le "contour d'iso-intensité" (courbe continue de l'espace gradient) correspondant. Mais cette contrainte n'est pas suffisamment forte pour déterminer sans ambiguité l'orientation de la surface observée : seul l'apport de contraintes supplémentaires permet de fournir les informations 3-D recherchées.

Le recouvrement de l'orientation de la surface observée peut s'effectuer de différentes manières, la plus "simple" étant la "stéréovision photométrique". En effet, si la même scène est observée par illumination avec une seconde source de lumière, la nouvelle intensité observée en (x,y,z) correspond à une deuxième courbe de l'espace gradient. En général, les deux courbes s'intersectent en deux points de l'espace (p,q), dont l'un représente effectivement l'orientation de la surface observée. Une troisième mesure obtenue à l'aide d'une troisième source d'intensité lumineuse permet de lever l'ambiguité résiduelle.

Une autre solution pour lever cette indétermination consiste à posséder des connaissances sur les objets observés. Prenons par exemple le cas d'une scène comportant un cube (cf. figure A.01.c) : les *gradients des trois faces* observées *forment un triangle équilatéral* dont la position dans l'espace gradient est connue à une translation et une homothétie (de centre celui du triangle) près. Si ce cube est vu sous les conditions requises, les trois intensités observées (l'intensité est constante sur une face plane) contraignent alors les orientations de surface à *se situer sur les contours d'iso-intensité* correspondants. La prise en compte simultanée de ces deux contraintes permet de déterminer exactement les orientations des surfaces.

L'inférence de forme à partir de la réflectance est un domaine en pleine évolution. C'est dans cette même ligne que les travaux de Pentland montrent comment faire une estimation de l'orientation des éléments de surface à partir du Laplacien de l'image [PENTL-82]. La majorité des travaux publiés sont basés sur la supposition que la direction de la lumière incidente est connue; et pour simplifier encore le problème, les surfaces étudiées sont supposées lambertiennes. A l'opposé, les surfaces usuelles du monde réel cumulent généralement propriété lambertienne *et* propriété miroir (cette dernière engendre en particulier des reflets sur les autres objets) : ces deux phénomènes concourent à un effet de surbrillance dans l'image et rendent évidemment le problème plus difficile à traiter. Des travaux comme ceux de [CASTA-84] montrent qu'il est néanmoins possible de calculer l'orientation des surfaces, en se passant même de toute connaissance a priori sur la lumière incidente et l'albedo de la surface.

2.4 L'Inférence de Formes à partir de la Texture

La texture d'une surface peut, elle aussi, être une clé importante dans la détermination de la géométrie d'une surface. Historiquement, le **changement de texture** était connu pour son utilité dans la discrimination des formes d'objet. Mais ce n'est que très récemment que la texture a été étudiée mathématiquement, afin que les systèmes de vision puissent mettre en relation la texture avec la forme d'un objet ("shape from texture"). Difficilement définissable de façon formelle, une texture peut être caractérisée statistiquement et structurellement comme une région de l'image qui contient plusieurs répétitions d'éléments de surfaces identiques.

Un parallèle assez étroit peut être fait entre l'inférence de forme à partir de la réflectance et l'inférence de forme à partir de la texture. Il suffit d'imaginer chaque

petit élément de texture ("texel") comme un pixel. L'apparence d'un texel d'une image change avec l'orientation de la surface, tout comme l'intensité d'un pixel. Tout comme dans le cas de la réflectance, l'analyse des éléments de texture conduit à poser un certain nombre d'hypothèses. L'une de ces hypothèses est l'homogénéité de la texture de surface, ce qui permet d'affirmer que toute variation observée dans l'image est due au changement soit de forme, soit de point de vue (notion analogue à la propriété lambertienne dans le cas de l'étude de la réflectance). Et là encore, ces hypothèses contraignent les orientations des éléments de surface, mais ne suffisent pas à les déterminer précisément.

Ces suppositions effectuées, il existe au moins trois manières différentes d'obtenir une information de surface à partir de l'observation des déformations des éléments de texture [KENDE-80].

- Si la forme exacte de l'élément de texture est parfaitement connue (un triangle ou un cercle par exemple), il est facile de calculer la transformation spatiale que cet élément a subi.

- La seconde consiste à supposer que tous les éléments de texture sont approximativement de la même taille. Des calculs de dérivée de l'importance des déformations permettent de déduire l'orientation locale de surface.

  Ikeuchi, travaillant sur la sphère gaussienne[2], propose une extension des deux premières méthodes pour la détermination des orientations de surface à texture régulières [IKEUC-80].

- La troisième méthode repose sur une analyse des contours d'éléments de texture supposés colinéaires [STEVE-81]. Les lignes de contour peuvent par exemple être le résultat de l'intersection d'une surface gauche avec plusieurs plans parallèles. La technique utilisée relève alors autant de l'inférence de forme à partir de textures de surface que de l'inférence de forme à partir de contours et s'appuie tout particulièrement sur la vérification de la continuité de la dérivée.

Finalement, pour les méthodes d'inférence de forme à partir de la texture (comme d'ailleurs à partir de la réflectance), une dernière alternative, dans le cas où aucune connaissance supplémentaire ne vient contraindre le problème, consiste à utiliser les deux contraintes universelles d'*unicité* (chaque point de l'image est associé au plus à une orientation de surface) et de *continuité* (sur une même surface, l'orientation varie peu, sauf au voisinage des contours de l'objet). La connaissance de l'orientation d'un élément de surface permet ainsi d'acquérir des connaissances sur celle des éléments de surface voisins. C'est sur cette idée d'une interprétation continue que repose la méthode statistique de [WITKI-81], qui permet de déterminer

[2]**sphère gaussienne** : alternative à l'espace gradient pour représenter les orientations de surfaces. La sphère gaussienne est une sphère unité dont l'axe Z est défini par le centre de la sphère et une direction représentant celle de l'observateur (traversant donc la sphère en un point appelé "pôle Nord"). Placé au centre de la sphère unité, un élément de surface voit son orientation représentée par le point d'intersection de sa direction normale avec la sphère.

jusqu'à des orientations de surface de scènes d'extérieur, en sélectionnant l'interprétation la plus uniforme.

La méthode générale est particulièrement efficace dans le cas de scènes composées d'objets gauches, pour lesquels il faut déterminer les orientations de surface pour chaque point, plutôt que pour toute une surface. De plus, certains points de l'image donnent des informations explicites et peuvent servir de points d'ancrage aux algorithmes de "relaxation"; ainsi, sur le contour d'un objet gauche, la ligne de vue est tangente à la surface de l'objet pour tous les points du contour et la normale à la surface est simplement définie comme la perpendiculaire commune à la ligne de contour et à la ligne de vue. Dans le cas de la texture, la propagation de contraintes permet de faciliter la recherche d'une solution globalement correcte, tout en résolvant localement les erreurs sur les texels, fortement sujets au bruit de discrétisation. C'est en cela que réside une des forces des méthodes d'inférence de forme à partir de la texture : des travaux comme ceux de [WITKI-81], utilisant très peu d'hypothèses mais des méthodes similaires à l'approche par raisonnement direct dans l'espace 3-D, montrent bien toute la puissance de la caractéristique "texture".

3 Le Raisonnement Direct dans l'Espace 3-D (2-D → 3-D)

La "reconstruction en profondeur" d'une scène observée à travers une simple image joue un rôle majeur pour les méthodes d'inférence de formes et constitue de ce fait un thème de recherche fondamental. Pourtant, certaines tâches de reconnaissance en Vision par Ordinateur (et parfois même dans le cas de la vision humaine) sont bien mieux accomplies **sans passer par une reconstruction en profondeur** : c'est ce que pensent plusieurs chercheurs (dont [BARNA-83] [HORAU-85] et [LOWE-85]) qui, par les méthodes qu'ils engendrent, définissent une "approche par raisonnement direct dans l'espace 3-D". Tous utilisent en entrée une unique image bidimensionnelle et l'approche suivie est définie par le respect simultané de quatre qualités :

- Extraction d'indices visuels susceptibles de posséder des propriétés conservées par le processus projectif de formation d'une image.
- Détermination de certains "groupements perceptuels" possédant des propriétés descriptives intéressantes, voire "intrinsèques" (i.e. invariantes en translation et en rotation).
- Rétroprojection dans l'espace 3-D" (i.e. projection inverse de l'espace image dans l'espace réel) des groupements perceptuels sélectionnés. A ce stade, les résultats sont d'autant plus discriminants que l'image est fortement perspective.
- Dans un objectif de *reconnaissance* : mise en correspondance des indices 3-D construits avec des modèles géométriques.

- Dans un objectif de *description* : raisonnement de cohérence spatiale à la recherche d'une solution globale la plus "consistante", voire la plus "évidente ", par minimisation d'un ou plusieurs critères de stabilité, qui correspondent à des propriétés globales de l'environnement.

C'est ainsi que Lowe reconnaît des objets par mise en correspondance spatiale de groupements perceptuels sélectionnés (parallélogrammes) avec une base de modèles [LOWE-85].

Barnard, quant à lui, détermine l'orientation de plans par rétroprojection de contours polygonaux fermés (la courbure d'un contour plan est une propriété intrinsèque, tout comme les angles d'un polygone plan qui eux, de plus, sont invariants aux changements d'échelle) [BARNA-83].

Dans le même esprit, Horaud [HORAU-85] effectue la reconnaissance d'objets composés de surfaces planes par mise en correspondance spatiale de sommets et d'angles.

Il faut remarquer que les indices visuels auxquels ces méthodes s'intéressent sont pour l'instant les mêmes que ceux qui sont utilisés dans les méthodes d'inférence de forme à partir de contours. Inversement, les travaux en inférence de forme à partir de contours de [BRADY-83] (avec minimisation d'un critère de compacité), ou ceux de [WITKI-81] traitant de la texture, sont à mettre en parallèle avec l'approche par raisonnement direct dans l'espace 3-D.

Notons finalement que l'originalité de ce type de méthode réside dans l'esprit même de leur démarche : elles offrent une alternative aux méthodes classiques qui reconstruisent d'abord des modèles de profondeur avant d'inférer la forme des objets. Ce séquencement est particulièrement caractéristique des méthodes d'inférence de formes multioculaires que nous présentons maintenant.

4 Les Méthodes d'Inférence de Formes Multioculaires (N x 2-D → 3-D)

Comme leur nom l'indique, ces méthodes ne se distinguent des méthodes d'inférence de forme monoculaires que par le seul fait qu'elles analysent **plusieurs** images, obtenues simultanément ou en séquence. La prise en compte de la multiplicité des images en entrée permet d'inférer des informations tridimensionnelles, tant en position qu'en orientation. Et tout comme dans le cas des méthodes d'inférence de forme monoculaires, les axes d'étude choisis sont souvent inspirés par des caractéristiques (parfois supposées) du système visuel humain.

4.1 L'Inférence de Formes à partir de la Stéréovision Simple

L'une des méthodes les mieux connues, car une des plus étudiées pour reconstituer la forme des surfaces des objets, est la "vision stéréoscopique", aussi appelée "vision binoculaire" ("shape from simple stereo"). Il s'agit ici d'étudier **la caractéristique**

de binocularité de la vision humaine, l'une des principales sources primaires d'informations que possède l'être humain pour percevoir des objets tridimensionnels dans leur environnement.

La vision stéréoscopique *simple* fournit une information tridimensionnelle à partir de la localisation d'une même forme dans deux images prises simultanément de deux points de vue différents par deux caméras dont les caractéristiques (internes et externes) sont *parfaitement connues* du système.

La figure A.02 détaille le principe bien connu de la stéréovision simple : ayant déterminé dans les deux images d'une même scène quels points des images correspondent à un même objet de l'espace réel, si la connaissance des positions relatives des deux caméras est correcte et si les caméras sont linéaires, alors l'objet correspondant aux points observés doit être localisé à l'endroit de l'espace réel où les deux lignes de vue de l'objet s'intersectent.

L'extraction de l'information de profondeur d'une paire d'images stéréoscopiques est classiquement décomposé en quatre composantes principales :

- la **détection** des indices visuels qui sont facilement reconnaissables dans les deux images,
- la **mise en correspondance** de ces indices dans les deux images,
- la **détermination des positions** relatives des deux caméras,
- le **calcul des distances** aux caméras des objets dont les indices caractéristiques ont été reconnus dans les deux images.

La vision stéréoscopique simple se distingue de la "vision stéréoscopique généralisée" par le fait que **la géométrie des caméras, exactement connue, permet de contraindre le processus de mise en correspondance** entre les images : la figure A.02 montre que l'objet responsable de l'indice P_g dans l'image gauche doit se situer le long du rayon C_gP_g. L'image de l'objet dans l'image de droite se trouve donc sur la projection du rayon C_gP_g sur le plan de l'image droite, soit sur la ligne D. La recherche pour la mise en correspondance d'un indice donné d'une image est donc restreinte à l'étude le long d'une ligne dans la seconde image (contrainte dite "de la ligne épipolaire"), plutôt que dans toute l'image.

Cette méthode constituant l'une des bases de notre seconde expérimentation, nous reviendrons dans la partie C sur la résolution des quatre phases de recouvrement des informations tridimensionnelles, ainsi que sur les suppositions et les implications liées à l'utilisation de la méthode [DEMAZ-85a].

La stéréoscopie simple a donné lieu à des réalisations révélatrices de la nature des indices visuels à détecter : les indices peuvent être ceux qu'on peut extraire d'une simple image, mais des travaux comme ceux de [GRIMS-81] et de [NINIO-81] montrent qu'il existe aussi des indices propres à la seule stéréovision, dynamiquement révélés par un processus de mise en correspondance fondé sur un calcul de

L'image d'un objet dans chacune des vues est engendrée par un rayon de lumière ayant pour origine l'objet et passant par le centre de la lentille. Inversement, selon le principe de "rétroprojection dans l'espace 3-D", le centre Cg de la lentille et un point de l'image Pg déterminent une ligne unique le long de laquelle l'objet se trouve. Le lieu M de la ligne CgPg où l'objet se situe peut facilement être détecté à l'aide d'une image prise d'une seconde caméra: s'il est possible de découvrir dans la seconde image ce qui peut correspondre au point de la première image, alors ce nouveau point Pd et le centre de la seconde lentille (soit Cd), déterminent un second de lumière CdPd. Si la connaissance des positions relatives des deux caméras est correcte et si les caméras sont linéaires, alors l'objet doit être localisé au point où les deux droites s'intersectent.

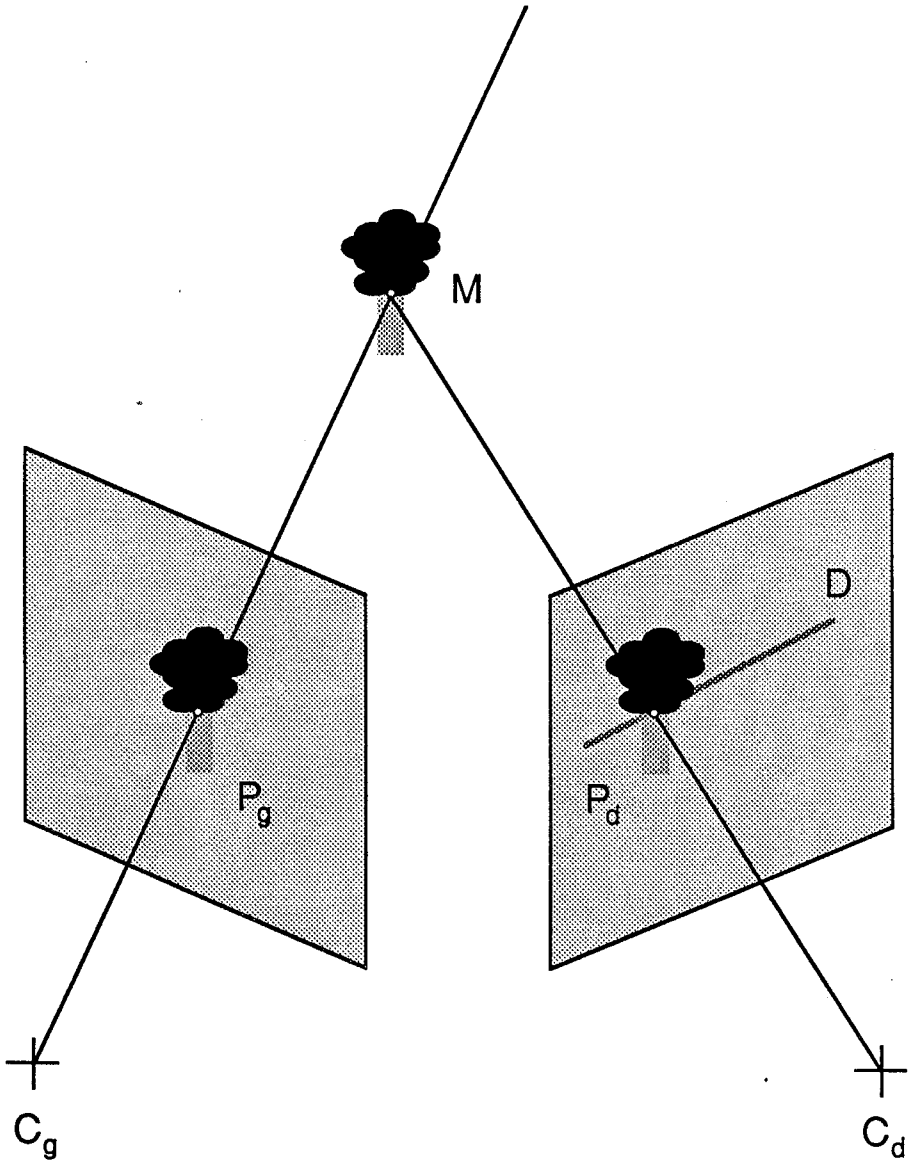


Figure A.02: Méthode d'Inférence de Formes à partir de la Stéréovision Simple

disparité. Dans ces systèmes, tout comme dans les travaux de [KONIS-84], les deux images étudiées sont prises simultanément par deux caméras. Mais l'utilisation de la stéréoscopie simple se rencontre également dans le domaine des robots mobiles. Souvent, les deux images sont alors prises par une caméra unique, en début et en fin d'un déplacement élémentaire du robot [BOLLE-85] [AYACH-85] [THORP-83b]. Dans tous les cas et malgré ses caractéristiques simplificatrices, la stéréoscopie simple reste un sujet très étudié car fournissant d'excellents résultats, même lorsqu'elle est appliquée isolément.

4.2 L'Inférence de Formes à partir de la Stéréovision Généralisée

L'homme, tout comme de nombreux animaux (les mammifères, mais aussi les oiseaux, dont la direction de l'oeil est fixe par rapport à celle de la tête) acquiert énormément d'informations tridimensionnelles en déplaçant sa tête dans une direction quelconque. La détermination de la forme des objets à l'occasion de ces mouvements élémentaires constitue le sujet spécifique des méthodes dites "de stéréoscopie généralisée" ("shape from generalized stereo" ou "shape from occlusion").

Ne disposant d'aucune information sur la géométrie du capteur utilisé, ces méthodes sont fondées sur la détection et l'exploitation de deux indices particuliers, fortement liés lors d'un déplacement du capteur de vision : les **lignes d'occultation** et les **changements de taille**.

Les lignes d'occultation correspondent généralement aux lignes de contours. La détection et le suivi de ces indices entrent de façon prépondérante dans la détermination de la *forme* des objets, tout particulièrement lors d'un déplacement *latéral* du capteur.

Pour la *localisation* des objets, c'est au contraire les changements de taille qui constituent l'indice prépondérant. De fait, révélée préférentiellement lors d'un déplacement *longitudinal* du capteur (la caméra se rapproche ou s'éloigne de la scène), l'amplitude des changements de taille des objets reflète leur distance à la caméra.

Bien que les résultats obtenus par cette méthode soient encore en nombre très restreint, et que le problème de mise en correspondance soit aussi difficile que dans le cas de la vision stéréoscopique simple, le potentiel inhérent à cette méthode est largement reconnu [MARR-82].

Dans l'instant, l'utilisation de cette méthode est bien souvent liée à un déplacement du capteur et se retrouve tout naturellement dans le domaine des robots mobiles. D'ailleurs, outre la méconnaissance de la géométrie du capteur (différence avec la "stéréovision simple"), la spécificité de cette méthode réside aussi dans le fait que le mouvement du capteur est contrôlé, et donc soumis à stratégies de contrôle (différence avec l'inférence de forme à partir du mouvement). C'est ainsi que les travaux de [TSUKI-85] et de [THORP-83a] s'attachent à montrer comment évolue l'apparence de l'environnement (en particulier l'étiquetage associé au "dessin

de lignes" observé) lors d'un déplacement de leurs ensembles robot-vision, et ceci indépendamment de toute connaissance précise sur la géométrie de leurs capteurs.

4.3 L'Inférence de Formes à partir du Mouvement

Tout comme il est possible d'acquérir des informations tridimensionnelles lors d'un déplacement du capteur par rapport à la scène observée, il est enrichissant d'observer le mouvement des objets lorsqu'ils se déplacent par rapport au capteur. Mais le but de ces méthodes n'est pas tant de détecter les **changements induits par le mouvement** des objets, que de **les mesurer** et de **les utiliser** pour, par exemple, séparer la scène en objets distincts et reconstruire les structures tridimensionnelles des objets en mouvement ("shape from motion").

Un objet rigide peut être décrit par un ensemble de points de référence sur sa surface. Deux points de référence définissent un vecteur (donc une valeur de norme et une direction) caractéristique de l'objet et de la rigidité de l'objet. Lorsqu'un objet rigide se déplace à l'intérieur de la scène observée, la norme et la direction tridimensionnelle d'un tel vecteur restent constantes, alors que leurs valeurs projetées et observées dans l'image peuvent évoluer. C'est ainsi que la localisation dans une séquence d'images de points de référence d'un objet en déplacement peut permettre de déterminer les valeurs (en norme et en direction) des vecteurs joignant ces points.

Les problèmes liés à une telle approche sont à mettre en parallèle avec ceux qui sont soulevés lorsqu'il s'agit d'utiliser la vision stéréoscopique :

- détermination des points de références
- mise en correspondance entre points de références
- inférence de la forme tridimensionnelle de l'objet à partir des informations de correspondance.

Généralement limités à l'observation d'objets solides et aisément résolvables pour des équations de mouvement simples (uniquement composées de translations), les différents sous-problèmes soulevés pour des mouvements plus complexes (où interviennent des translations mais aussi des rotations), conduisent à des analyses beaucoup plus poussées, surtout quant à la détermination et la mise en correspondance de points de référence.

Une première approche du problème consiste à remarquer que, du fait de l'évolution tridimensionnelle des objets de la scène observée, ce ne sont pas tellement des "points" de référence qui sont recherchés, mais des indices tridimensionnels dont les extrémités puissent fournir de tels points. Il apparaît pourtant, à la lumière des expérimentations menées par [ULLMA-79], que les indices tridimensionnels ne sont pas les plus adéquats au processus de mise en correspondance : au contraire, on peut se contenter de configurations bidimensionnelles en tant qu'indices de référence. Dans le schéma défini par Ullman (projection orthographique), l'observation de quatre points dans trois images successives permet de reconstruire un objet solide.

L'approche de [LONGU-80], plus connue sous le nom de "flot optique", est fondée sur l'analyse de deux images perspectives. Elle détecte des vecteurs en assez grand nombre pour pouvoir calculer les dérivées spatiales première et seconde du champ de vélocité : bien que ce champ soit ici induit par un déplacement absolu du capteur et non des objets, la méthode utilisée fait partie des méthodes d'inférences de forme à partir du mouvement.

Les méthodes d'inférence de formes à partir du mouvement se distinguent de la stéréovision généralisée non seulement dans le sens où les premières (resp. dernières) sont souvent étudiées lors d'un déplacement des objets par rapport au capteur (resp. du capteur par rapport à son environnement) mais surtout car elles s'attachent à la caractéristique "champ de vélocité" ([HILDR-82]) et non pas aux caractéristiques "lignes d'occultation" et "changements de taille" de la vision humaine.

4.4 L'Inference de Forme a partir des Reflets

Nous avons déjà vu comment l'existence de sources de lumière ponctuelles permettait d'extraire des informations sur l'orientation des surfaces (ombres, réflectance). Nous avons vu de même à quel point cette source de renseignements était difficile à appréhender du fait que l'intensité des pixels ne dépend pas seulement de l'orientation des surfaces étudiées. Malgré tout, face à une image du monde réel, l'être humain voit son attention attirée par certaines caractéristiques comme les arêtes, mais aussi par certains points, particuliers du fait de leur très forte brillance ("reflets "). Habituellement, et tout particulièrement en milieu industriel, on cherche plutôt à éviter l'existence des reflets, en agençant astucieusement les lumières ponctuelles ambiantes et/ou en choisissant des matériaux peu réfléchissants, si cela est possible. C'est pourtant la **détection de ces reflets et leur suivi** lors d'un déplacement relatif de du capteur par rapport aux objets, qui constitue la base de la méthode d'inférence de formes du même nom ("shape from highlights"). Elle correspond donc à une étude d'un cas particulier de réflectance avec séréovision généralisée.

Ne disposant là encore d'aucune information sur la géométrie du capteur utilisé, la démarche proposée est à rapprocher de celle de l'inférence de formes à partir de la stéréovision généralisée. Etant données la connaissance de la direction de la source de lumière et de la position approximative d'une surface dans le monde réel, l'*orientation* de certains éléments de surface peut être déterminée précisément en utilisant l'existence de reflets (cf. figure A.03).

Une méthode consiste dans un premier temps à détecter et à caractériser ces reflets avec des attributs de formes adéquats, du genre de ceux qui sont utilisés dans le domaine de la Reconnaissance des Formes. Dans un second temps, il s'agit de suivre les bords des reflets détectés lors d'un déplacement relatif de l'objet et de la caméra, pour finalement en déduire une description locale de l'orientation des éléments de surface ainsi particularisés.

Cette méthode d'inférence de formes est une des toutes dernières nées, et n'a

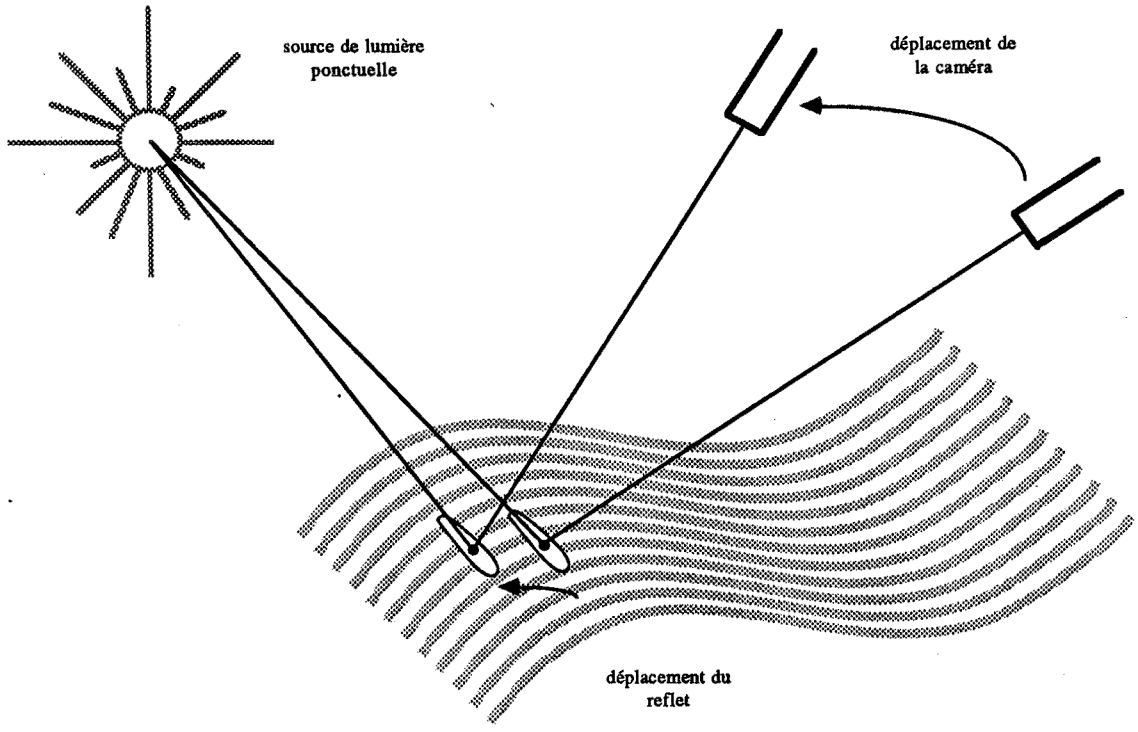


Figure A.03: Méthode d'Inférence de Formes à partir des Reflets

pas encore été vraiment étudiée. Néanmoins, ce sont bien les reflets qui sont utilisés comme indices primaires dans les travaux de [HEALE-85], qui exhibent des relations entre les propriétés des reflets dans une image et les caractéristiques des surfaces réelles correspondantes.

Une autre utilisation de la caractéristique "reflets" apparaît dans le cadre de la levée d'indétermination inhérente à la méthode d'inférence de formes à partir de la réflectance [BLAKE-85] : la détection des reflets sur les éléments de surface, en autorisant l'extraction d'informations sur la concavité de ces surfaces, fournit en effet des contraintes sur l'orientation de ces surfaces.

5 Les Méthodes Actives de Formation d'Images 3-D (3-D → 3-D)

L'une des caractéristiques principales des méthodes actives de formation d'images 3-D réside dans leur non-anthropomorphisme. Appelées aussi "méthodes directes ", elles disposent en entrée d'informations tridimensionnelles obtenues grâce à l'utilisation de capteurs de distance et bénéficient en outre de modèles classiques en Mathématiques, comme la triangulation.

5.1 Les Méthodes Optiques

La plus populaire des méthodes actives optiques concerne l'utilisation de lumière comme outil de mesure de distance. Le principe général en est parfaitement illustré par le système développé par Borianne [BORIA-84] : un plan de lumière cohérente (qui peut être émis soit à l'aide d'une lentille cylindrique, soit par déplacement rotatif plan d'un rayon de lumière) intersecte les objets de la scène selon une courbe. Cette courbe de l'espace 3-D, contenue dans le plan de lumière, est perçue par une caméra qui, si elle est munie d'un filtre monochromatique de même longueur d'onde que la lumière utilisée, permet au système de disposer d'une image révélant un profil des objets présents dans le plan de lumière (cf. figure A.04).

Par déplacement relatif des objets et/ou du plan de lumière, il est possible de construire une collection de profils tridimensionnels. Pour assurer le déplacement relatif des objets et du plan de lumière alors que la caméra est fixe, il est possible :

- soit d'utiliser un plateau translatable [BORIA-84] ou rotatif [LELAN-84] sur lequel sont posés les objets,
- soit de déplacer le plan de lumière en s'assurant que le profil reste observable dans l'image [FAUGE-84].

C'est le principe même d'acquisition de ces images, permettant de calculer *directement* une image tridimensionnelle sans difficulté majeure, qui explique pourquoi ces méthodes sont appelées "méthodes actives de formation d'images tridimensionnelles".

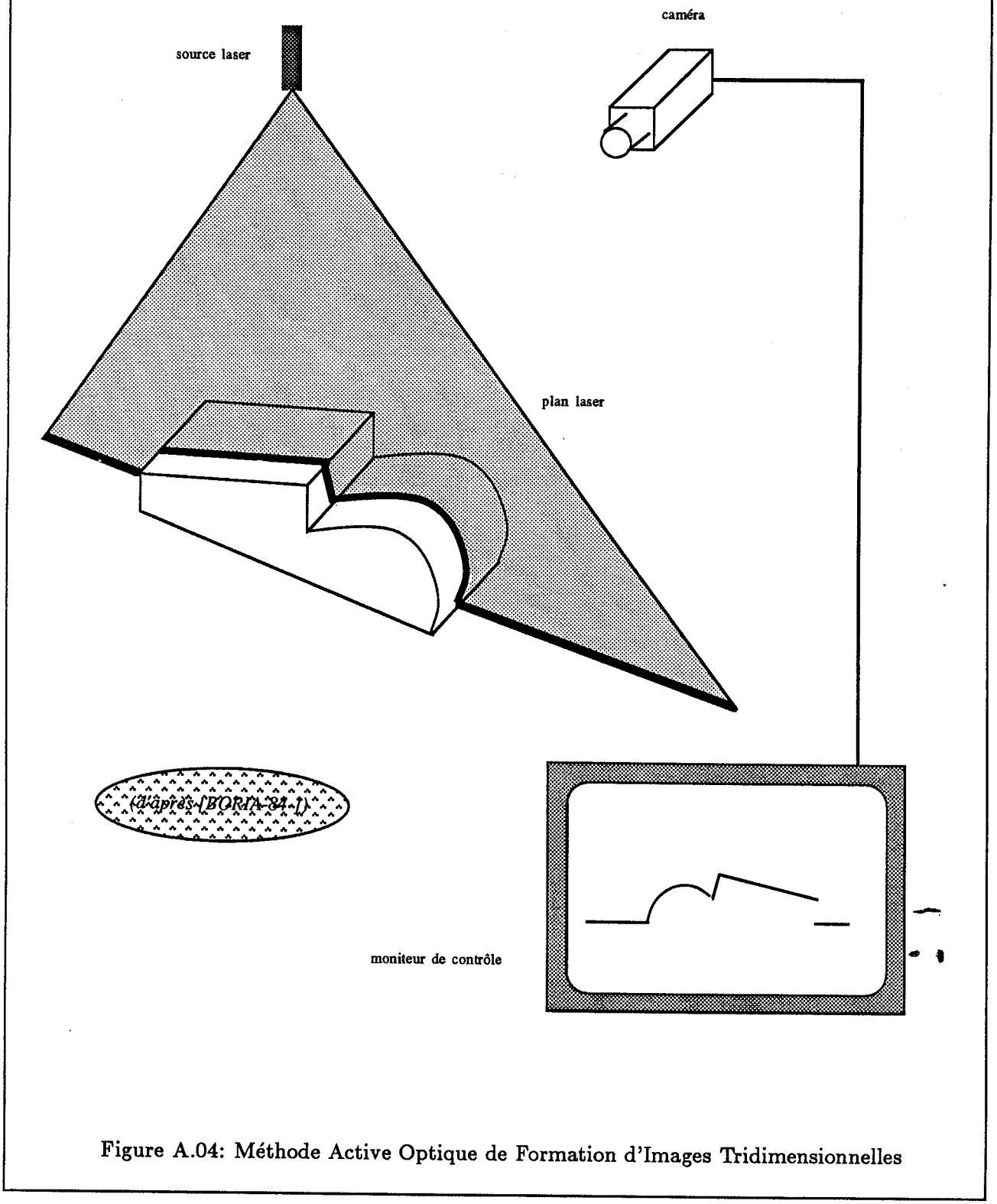


Figure A.04: Méthode Active Optique de Formation d'Images Tridimensionnelles

Dans un second temps, il reste à analyser les différents profils tridimensionnels (recherche de plans et de surfaces gauches, détection d'arêtes) pour pouvoir reconstituer un modèle tridimensionnel de la scène observée. Bénéficiant de données relativement complètes, à l'inverse des méthodes du type "inférence de forme", les techniques applicables sont plus aisées. Parmi les nombreuses réalisations de reconstruction de modèles tridimensionnels de scène suivant ce principe figurent particulièrement celles de [OSHIM-81] (reconnaissance d'objets quelconques décrits par régions), de [BOLLE-83] (orientation de surfaces planes) et de [FAUGE-84] (identification d'objets décrits par éléments de surface).

Notons finalement que nombre de ces systèmes sont d'ores et déjà utilisés dans le monde industriel, comme c'est le cas des systèmes de [BORIA-84] et de [LELAN-84]. Réciproquement, il faut noter que les systèmes tridimensionnels installés dans l'industrie sont presque tous basés sur l'utilisation de ces méthodes optiques.

5.2 Les Autres Méthodes

Non optiques, les autres méthodes sont cependant fondées sur le même principe d'acquisition initiale d'informations tridimensionnelles suivie de leur analyse. Seul le type des informations initiales diffère : les méthodes actives acoustiques sont par exemple fondées sur le principe d'un balayage de la scène par un signal ondulatoire et mesure du "temps de vol" que met ce signal pour revenir après avoir été émis. Ces méthodes sont fréquemment utilisées en Robotique Mobile [GIRAL-84] [CROWL-85]. Leur gros avantage réside dans leur faible coût temporel, ce qui permet par exemple d'envisager l'étude de mouvements réflexes d'évitement d'objets mobiles évoluant dans l'environnement du robot. L'information traitée étant souvent moins riche qu'une information visuelle et généralement de faible précision, les méthodes actives non optiques permettent plus une "sensation" qu'une réelle "perception" de la scène Nous ne nous attarderons d'ailleurs pas plus sur les méthodes actives non optiques car, hors leur analogie aux méthodes actives optiques, elles ne présentent aucun caractère "visuel" et ne font pas vraiment partie de la Vision par Ordinateur.

Chapitre A.II

Niveaux de Représentation en Vision

Si on adhère à la méthodologie actuelle des études menées en Vision par Ordinateur (conjonction d'études isolées de caractéristiques inspirées du système visuel humain), il faut cependant remarquer que tous les modules ne travaillent pas *directement* sur l'image. Par exemple, il est clair que la vision stéréoscopique n'est pas réalisée sur la base de correspondances au niveau des pixels. Pour chaque étude d'un thème particulier, on est donc conduit à se poser les questions sur la nature des représentations sur lesquelles ces méthodes opèrent et sur celle des descriptions qu'elles permettent de produire. Un autre problème à résoudre est celui de la collaboration entre ces différentes méthodes, tant sur le plan des représentations qui doivent être partagées que sur celui des moyens effectifs de communication à mettre en oeuvre.

Telles sont deux des principales réflexions qui nous ont conduit à étudier la "représentation" de l'information visuelle à différents "niveaux".

"Représentation" - "Niveau" : deux mots que nous manipulons couramment, mais qu'il paraît bien difficile de définir. Dans un premier temps, nous nous contenterons des définitions suivantes, extraites du "Petit Robert" (Dictionnaire Alphabétique et Analogique de la Langue Française des Editions Robert) :

niveau élévation comparative, degré comparatif, échelon d'une organisation.

représentation le fait de rendre sensible (un objet absent ou un concept) au moyen d'une image, d'une figure, d'un signe, etc ...

L'objet des paragraphes qui constituent ce chapitre n'est effectivement pas tant la contribution qu'ils peuvent offrir à la **définition** de ces termes que la présentation des **constats** et les **justifications** qu'ils apportent en faveur de l'existence informatique de structures intermédiaires indispensables à tout processus de compréhension de scènes visuelles.

1 Les approches des Niveaux de Représentation en V.O.

La représentation de l'information visuelle et sa structuration en niveaux est un des problèmes centraux de la Vision par Ordinateur. D'un point de vue informatique, il est classique, encore que souvent artificiel et toujours difficile, de distinguer plusieurs niveaux d'opérations dans les phénomènes visuels (cf. figure A.05).

1.1 Le Paradigme Segmentation - Interprétation

Les premiers systèmes de reconnaissance constituent l'exemple type des systèmes où les niveaux de représentations semblent les plus naturels, sans qu'il n'aït jamais été montré qu'ils étaient suffisants. De [ROBER-65] à [WALTZ-75], la conception globale des niveaux de représentation d'un système de vision suit une même philosophie : mis à part le niveau IMAGE, le système est composé de deux autres niveaux (cf. figure A.05.a) :

- le niveau INDICES VISUELS, incluant des lignes (lignes de contraste) et des régions (région d'homogénéité). Il est atteint grâce à des processus de *segmentation* (vision de "bas niveau") qui permettent d'extraire des informations pertinentes en termes d'indices visuels.

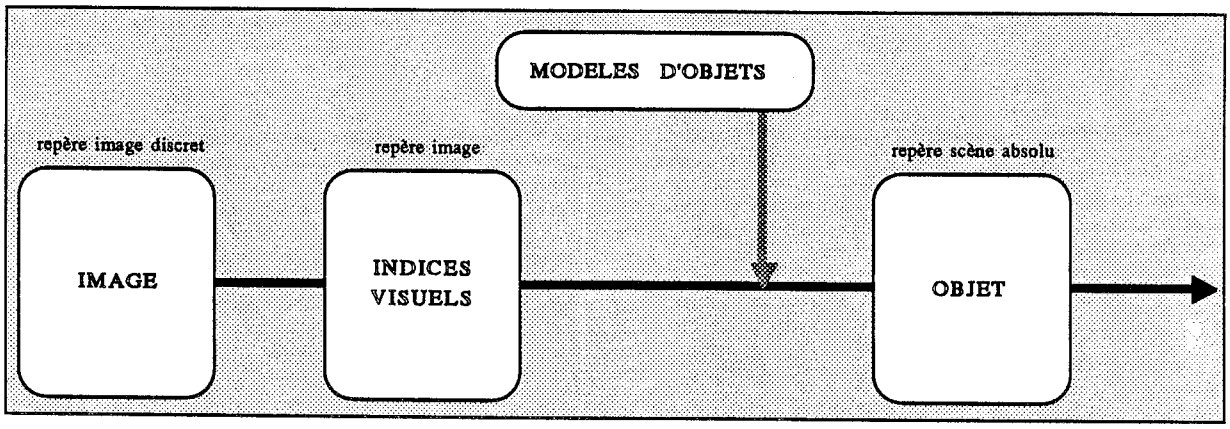
  Le repère utilisé par cette structure est le repère de l'image.

- le niveau OBJET, qui est le résultat d'une *interprétation* fondée sur une mise en correspondance directe entre les indices visuels et des modèles d'objet.

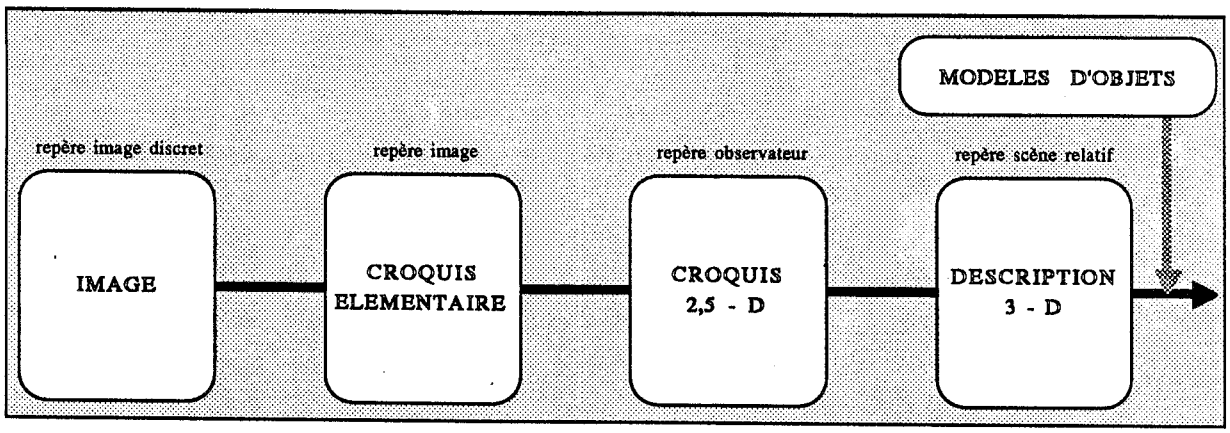
  Le repère utilisé par cette structure est le repère lié à la scène observée (repère scène absolu).

Une telle démarche suppose que chaque indice visuel de ligne corresponde à un contour ou à une arête interne d'un objet, et que chaque région correspond à une face réelle d'un objet. Le système est dépendant de la qualité du processus de segmentation et l'interprétation reste fortement combinatoire. L'impossibilité constatée d'une bonne segmentation conduit soit à rendre le système hétérarchique, soit à revoir entièrement la structure du système.

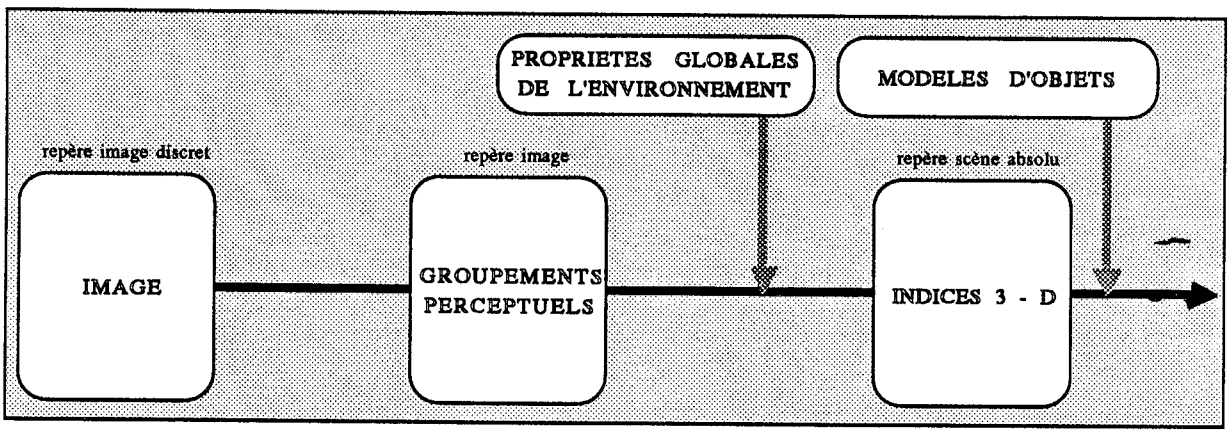
Les processus sont représentés de sorte que toute connaissance externe au système intervient le plus tard possible dans la hiérarchie des niveaux considérés. Dans le même esprit, les possibilités de contrôle prédictif ne sont pas représentées.



A.05.a Le paradigme segmentation - interprétation (1965 --> 1975)



A.05.b L'intermédiaire reconstruction en profondeur (1975 -->)



A.05.c Le raisonnement direct dans l'espace 3-D (1983 -->)

Figure A.05: Les trois approches informatiques d'un système de V.O.

1.2 La reconstruction en profondeur et/ou en orientation

Une nouvelle structuration, c'est ce que Marr propose dès 1975. Exposée au plus complet dans [MARR-82], la structure préconisée introduit un niveau supplémentaire, le niveau "croquis 2,5-D", qui correspond à une reconstruction intermédiaire de l'information de profondeur d'un indice et/ou de son orientation. De nombreux systèmes ont depuis choisi ce modèle, car il reflète une nécessité incontournable dès que sont utilisées les méthodes d'inférence de formes monoculaires ou multioculaires. Les niveaux qui y sont distingués sont les suivants (cf. figure A.05.b) :

- le niveau CROQUIS ELEMENTAIRE ("primal sketch") : il s'agit d'obtenir des représentations convenables à partir des structures et des changements présents dans l'image. Cela sous-entend un certain nombre de choses telles la détection de changements d'intensité lumineuse, la représentation et l'analyse de structures géométriques locales, et la détection des effets lumineux (lumière ambiante, source de lumière ponctuelle, transparence). Ce niveau rend donc explicite l'information contenue dans l'image bidimensionnelle, et tout particulièrement les changements d'intensité lumineuse et leurs distribution et organisation géométriques dans l'image.

  Le système de coordonnées dans lequel la représentation est construite est celui de l'image.

- le niveau CROQUIS 2,5-D ("2,5-D sketch") : il s'agit d'appliquer un certain nombre de primitives opérant sur la représentation du "primal sketch" pour dériver une représentation de la géométrie des surfaces visibles dans l'image. Les processus utilisés sont des évaluations d'orientation de surfaces locales, de discontinuité en surface ou en profondeur, de distance de l'observateur à la scène.

  La représentation construite à ce niveau est réalisée dans un système de coordonnées lié à l'observateur.

- le niveau DESCRIPTION 3-D ("3-D model") : il s'agit d'obtenir une description des formes et de leur organisation spatiale dans un repère lié à l'objet, en utilisant une représentation hiérarchique modulaire où figurent des primitives de volume (représentant les volumes d'espace occupés par l'objet) aussi bien que des propriétés de surface.

  La représentation est donc exprimée dans un repère scène relatif.

L'intérêt de la démarche réside dans son caractère d'interprétation ascendante : une description de la scène 3-D est construite sans utilisation de connaissances spécifiques sur la scène observée.

1.3 Le raisonnement direct dans l'espace 3-D

L'intermédiaire reconstruction en profondeur est-elle absolument nécessaire? Non, selon un dernier courant de pensée qui correspond à l'"approche par raisonnement direct dans l'espace 3-D", dont nous avons exposé les principes au chapitre

précédent. Les niveaux de représentation qui sont alors distingués [BARNA-84] (parfois implicitement) sont les suivants (cf. figure A.05.c) :

- le niveau GROUPEMENTS PERCEPTUELS : chaque élément de cette structure possède des propriétés descriptives intrinsèques qui sont conservées par le processus projectif de formation d'une image.

  Le repère dans lequel la structure est exprimée est le repère de l'image.

- le niveau INDICES 3-D : les indices de scène sont obtenus par simple rétroprojection et compte tenu d'un raisonnement de cohérence spatiale faisant intervenir les propriétés globales du contexte de la scène.

  La structure ainsi engendrée est exprimée dans un repère scène absolu.

L'intérêt de la démarche réside dans la possibilité d'acquérir *directement* une représentation 3-D partielle de la scène en ne tenant compte que des propriétés globales de l'environnement, et donc, là encore, indépendamment de toute connaissance spécifique.

2 L'existence de Niveaux pour la Vision Humaine

La méthodologie actuelle des études menées en Vision par Ordinateur conduit tout naturellement à tenir compte des résultats obtenus dans les domaines de la psychologie et de la neurophysiologie. Or si d'un point de vue informatique, trois approches du problème des niveaux de représentation partitionnent les différents systèmes existants, qu'en est-il de l'étude du système visuel humain?

Voir est une faculté qui a confondu philosophes et hommes de science pendant des siècles, et continue de le faire (cf. figure A.06). En effet, même si certains processus perceptifs peuvent être conscients, les autres processus visuels humains sont indépendants de la conscience. En particulier tous les processus sensoriels sont trop complexes pour que l'homme puisse de lui-même les prendre en compte de façon consciente. Réciproquement, le libre arbitre ne semble rien avoir à gagner à considérer des problèmes comme par exemple celui de la correspondance stéréoscopique.

A l'heure actuelle, notre compréhension de la machinerie visuelle du cerveau est loin d'être parfaite. Nous savons qu'elle possède de nombreuses unités, chacune spécialisée dans un rôle particulier. Cependant, le monde visuel dont nous sommes conscients est à l'évidence un monde totalement intégré. Et ce sont quelques-uns des résultats obtenus en psychologie et en neurophysiologie concernant l'existence de différents niveaux de représentation dans le processus visuels que nous exposons dans les paragraphes qui suivent.

2.1 Le triptyque Neurophysiologie - Psychologie - Informatique

Trois domaines, la neurophysiologie, la psychologie et l'informatique, en qualité de sciences cognitives, se penchent sur le problème de la vision : le problème posé nous

EPISODE 12

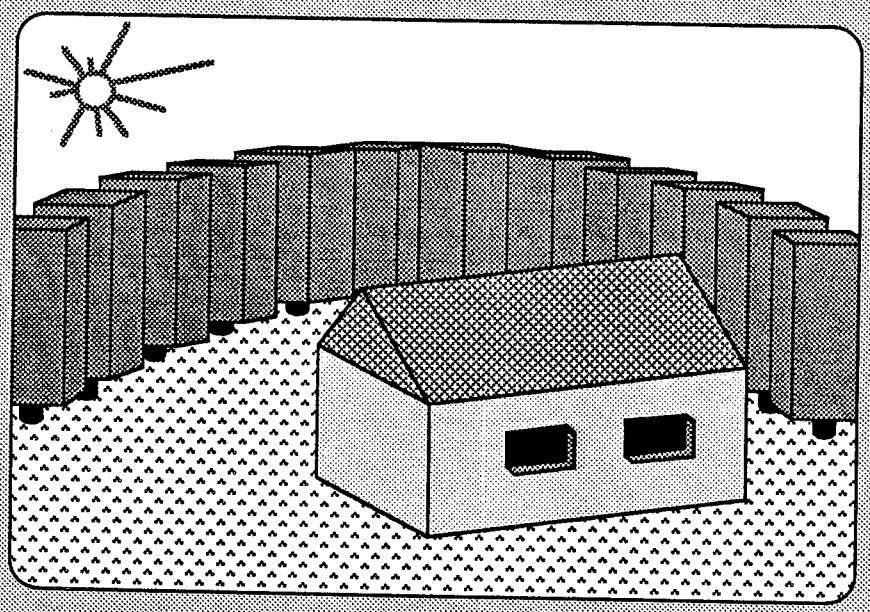
Ils avaient prévu de faire le casse ce soir-là, et étaient partis dans l'après-midi reconnaitre une dernière fois la maison dans laquelle ils allaient opérer de nuit.

Bobby, l'homme de main qui devait assurer le bon fonctionnement de l'opération, était assis dans la voiture à la position du mort, et cela ne lui plaisait qu'à moitié. Sortant de l'enfer de ses précédentes missions, on l'avait surnommé "Cerbère"; il avait accepté cette mission sans aucun problème. La récompense promise était alléchante et il devait travailler avec le renommé compère qui lui avait été attribué et qui, cet après-midi là, devait lui donner les premières indications. Car en effet, pour l'instant, il ne savait même pas où tout cela devait avoir lieu.

L'homme qui était au volant s'appelait Charly et était bien connu du milieu pour les onze précédentes missions périlleuses qu'il avait menées à bien. Plutôt connu sous le nom d'"Hercule" et bien que d'apparence massive, ce roc cachait une grande érudition doublée d'une expression synthétique qu'il aimait retrouver chez les autres.

Son attention captivée par la circulation l'empêchait de détailler les maisons qu'ils longeaient actuellement; pourtant, il était sûr qu'elle était dans cette rue, sur la droite. Aussi, il demanda à son acolyte de lui décrire les maisons qu'il voyait à travers la vitre.

Charly: "Dis Bobby, et celle-là, elle est comment?, Qu'est-ce que tu vois?"
Bobby: "C'est une villa avec des fenêtres sur l'un de ses cotés, entourée d'un jardin limité par une haie de thuyas, arrangés en cercle."
Charly: "C'est bon c'est celle-là. Je vais repasser plusieurs fois devant afin que tu t'imprègnes bien de la situation".



Et c'est ainsi, que possédant une vision claire de la scène, Charly et Bobby réussirent le coup qu'ils avaient projeté. Mais, suivant les ordres qui lui avaient été donnés, Charly supprima Bobby car il en savait beaucoup trop.

Et c'est ainsi qu'Hercule envoya Cerbère au royaume des morts, rendant par là même un grand service à l'humanité.

Mais cela, l'histoire ne le dit pas ...

Figure A.06: Compréhension humaine d'une scène

paraît avoir d'autant plus de chances d'être résolu si chaque domaine tente de tenir compte des résultats obtenus dans les deux autres. Autrement dit, il nous apparaît important, dans la mesure du possible, de trouver des méthodes susceptibles de ne conduire à des contradictions, ni avec des données physiologiques, ni avec les considérations relatives aux structures opératoires des processus intelligents.

Si ces trois domaines concourent complémentairement à trouver une solution, ils sont néanmoins de natures bien différentes, par les méthodes qu'ils emploient et par le niveau des informations qu'ils s'attachent à étudier. *"Par leurs racines, en effet, les mécanismes perceptifs relèvent du domaine de la physiologie du système nerveux, tandis que leurs formes supérieures, pour autant que l'on est conduit à distinguer plusieurs paliers au sein des activités perceptives, rejoignent les adaptations élémentaires de l'intelligence"* [PIAGE-61]. Chaque chercheur concerné est de ce fait assis sur une seule chaise. C'est ainsi que si le physiologiste étudie le "hardware" des systèmes visuels biologiques (à l'aide de micro-électrodes), et alors que le psychologue fournit des méthodologies pour étudier les performances d'"entrée-sortie" du système visuel (souvent à l'aide d'illusions), il s'agit pour l'informaticien d'essayer de construire effectivement un système de vision capable de remplir artificiellement les fonctions du système visuel humain (à l'aide de matériel et de logiciel).

Cependant, nous sommes conduits à penser que les domaines de la psychologie et de la neurophysiologie peuvent contribuer non seulement à comprendre le fonctionnement du système visuel humain, mais aussi à guider l'élaboration d'un modèle informatique de la vision capable d'effectuer une compréhension de scène. Réciproquement, nous espérons que la conception et la réalisation de modèles informatiques du processus de vision peuvent contribuer à une meilleure compréhension du système visuel humain. Car selon nous, **le rôle de l'informatique de la vision est aussi de simuler les modèles dérivés (tant connexionnistes [BALLA-84] que macroscopiques) des autres disciplines pour éprouver leur fiabilité et leur efficacité**. C'est ce que pensent certains informaticiens [POGGI-84], et c'est aussi ce qu'attendent les psychologues [BARLO-85] et les neurophysiologistes. *"Décrire et expliquer sont deux opérations différentes. Même si un neurobiologiste était capable de décrire le comportement de chaque neurone, il ne pourrait pas prédire l'état du cerveau. Le monde obéit à des lois déterministes dont le résultat est imprévisible. Même si nous pouvions parvenir à une connaissance exacte de tous les mécanismes physico-chimiques élémentaires du système nerveux, même si nous pouvions déduire ce fonctionnement d'une description exhaustive du cerveau, nous aurions encore des difficultés à expliquer les opérations d'un système aussi complexe que le cerveau"* [IMBER-85]. C'est ainsi que les travaux de Marr apportent une certaine lumière quant à la conceptualisation du processus de vision [MARR-78] [MARR-82] : ils définissent des niveaux de représentation (croquis élémentaire, croquis 2,5-D, modèle 3-D) qui s'appuient sur des études effectuées avec des neurophysiologistes et des psychologues, et fondées sur le respect de propriétés physiques régissant le monde réel.

Pourtant, malgré les efforts conjugués des chercheurs de ces trois domaines, nous

sommes encore loin de pouvoir construire une machine qui rivalise avec la capacité de l'homme à décrire des scènes naturelles complexes, tout comme Bobby peut le faire si naturellement pour Charly (cf. figure A.06).

2.2 La théorie de l'"Ecran Intérieur"

La première idée qui vient à l'esprit pour expliquer la vision humaine est la théorie dite de l'"écran intérieur" [FRISB-79]. Chaque oeil travaille comme une caméra. Il est muni d'une lentille (le cristallin) et possède une rétine, constituée de petites unités réceptrices. Le but de la lentille est de projeter sur la rétine une image du monde extérieur. Ainsi stimulée, la rétine transmet des messages au sukjet de cette image le long des fibres du nerf optique jusqu'au cerveau. Celui-ci est composé de millions de petits composants appelés "cellules". Certaines de ces cellules sont spécialisées dans la vision et disposées en forme de "feuille de papier", sorte de "tableau noir" bidimensionnel : c'est l'"écran intérieur". Chaque cellule de l'écran peut à tout moment être active ou inactive, en traduisant la brillance du point correspondant de l'image. Les cellules de l'écran intérieur, dans leur ensemble, présentent une structure d'activité dont la forme réfléchit directement celle de l'image rétinienne reçue par les yeux.

Cette théorie est aisée à comprendre et attrayante pour l'esprit, car nos expériences visuelles paraissent, en quelque sorte, "cadrer" avec le monde extérieur. Il est donc naturel de supposer qu'existent dans le cerveau des mécanismes de la vision permettant un cadrage des plus simples, un cadrage matériellement identique, c'est-à-dire "photographique". L'image photographique cérébrale de la scène regardée, élaborée par le système visuel humain constitue selon la théorie de l'écran intérieur la base de notre expérience visuelle. Mais cette théorie bute sur son incapacité à expliquer comment nous sommes en mesure, par exemple, de reconnaître les différents objets représentés par les images cérébrales. Présentée ainsi, cette théorie n'est d'ailleurs a priori relative qu'à la description de scène par représentation dans une image plane, bidimensionnelle. Pourtant, en général, le système visuel a affaire à des scènes en trois dimensions, et accomplit "parfaitement" son travail en fournissant une **description explicite** sur la localisation des objets fixes de la scène, mais aussi sur les objets qui évoluent dans l'espace. Dans ce sens, le système visuel humain satisfaisait à la faculté de "compréhension de scène", telle que nous l'avons définie. La théorie de l'écran intérieur, quant à elle, peut difficilement être étendue de manière logique en émettant la proposition que l'écran intérieur est bien une structure tridimensionnelle, masse imposante de "cellules nerveuses", qui représentent les points individuels de la scène, quelle que soit par exemple la distance de ces points à la rétine.

D'un point de vue de l'informaticien, le satisfecit initial de la théorie de l'écran intérieur s'oppose au défaut de ne rien expliquer en termes ... informatisables! De plus, et comme de nombreux travaux effectués en neurophysiologie et en psychologie expérimentale le montrent, elle laisse de nombreux phénomènes visuels inexpliqués.

Et c'est dans une direction qui s'est avérée riche en résultats intéressants qu'ont été entreprises des études interdisciplinaires, faisant appel à des connaissances tant sur le plan neurophysiologique que sur les plans psychologiques et informatiques.

2.3 Les apports de la Neurophysiologie

Tant en neurophysiologie qu'en psychologie, on est souvent amené à distinguer plusieurs niveaux d'opération dans les phénomènes visuels. La capacité humaine à comprendre une scène et à raisonner sur cette scène concerne les niveaux où interviennent des phénomènes psychologiques complexes comme la catégorisation, la mémoire ou l'attention. Les opérations de détection et d'analyse des indices visuels représentent en revanche une série d'étapes indispensables aux opérations perceptives ultérieures. C'est l'élucidation de ces étapes qui constitue le programme de la neurologie du système visuel humain. L'étendue des travaux effectués dans le domaine n'est qu'effleurée dans les paragraphes suivants et le lecteur intéressé pourra se référer aux nombreuses publications du domaine spécialisé, et en particulier à [IMBER-83] où à [FRISB-79].

Suivons rapidement le trajet de l'image optique de la scène observée par l'oeil de Freddy. L'image initialement perçue par la **rétine** emprunte la voie du nerf optique pour transiter dans les **corps genouillés latéraux** avant d'être transmise au **cortex cérébral** (cf. figure A.07). Quels sont les traitements que cette information subit?

2.3.1 Une compression de l'information visuelle dès la rétine

Dès 1950, les expériences neurophysiologistes ont permis de dégager une notion fondamentale : chaque cellule du système visuel, qu'il s'agisse d'un neurone ganglionnaire de la rétine, d'un neurone du corps genouillé latéral ou d'un neurone des aires visuelles des hémisphères cérébraux, possède un "champ récepteur". Il s'agit d'une aire d'étendue et de forme donnée sur la rétine qui lorsqu'elle reçoit un stimulus lumineux, évoque une réponse au niveau du neurone. Les neurones ganglionnaires répondent par une augmentation (cellules à centre "ON") ou une diminution (cellules à centre "OFF") de leur émission spontanée d'impulsions lorsqu'une petite tache lumineuse tombe dans leur champ récepteur. Et plutôt que de dresser un tableau impressionniste voire pointilliste de la scène visuelle, ils détectent en fait les **différences d'illumination** entre deux zones contigues à l'intérieur de leur champ récepteur.

Cette organisation antagoniste des champs récepteurs est l'un des moyens utilisés par le système visuel pour extraire de façon économique les aspects pertinents des objets ou des évènements extérieurs nécessaires à la perception du monde visuel. De fait, ce sont des procédés de ce type qui permettent au système nerveux de l'oeil une **compression de l'information** à un point tel que l'image oculaire, découpée par 150 millions de bâtonnets et 6 ou 7 millions de cônes, est intermédiairement réduite à un million de messages qui transitent le long des fibres optiques.

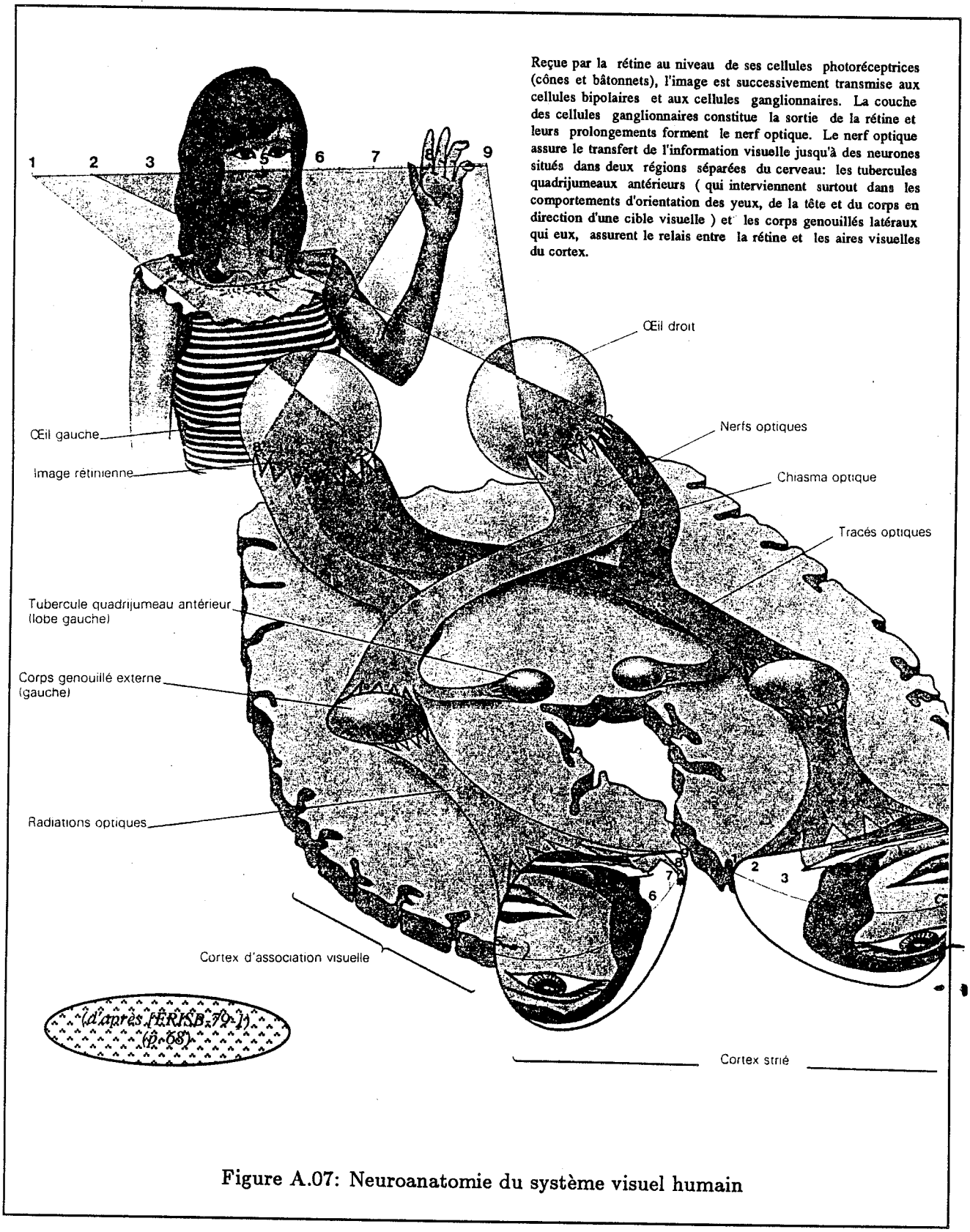


Figure A.07: Neuroanatomie du système visuel humain

2.3.2 Un traitement séquentiel de l'information visuelle ...

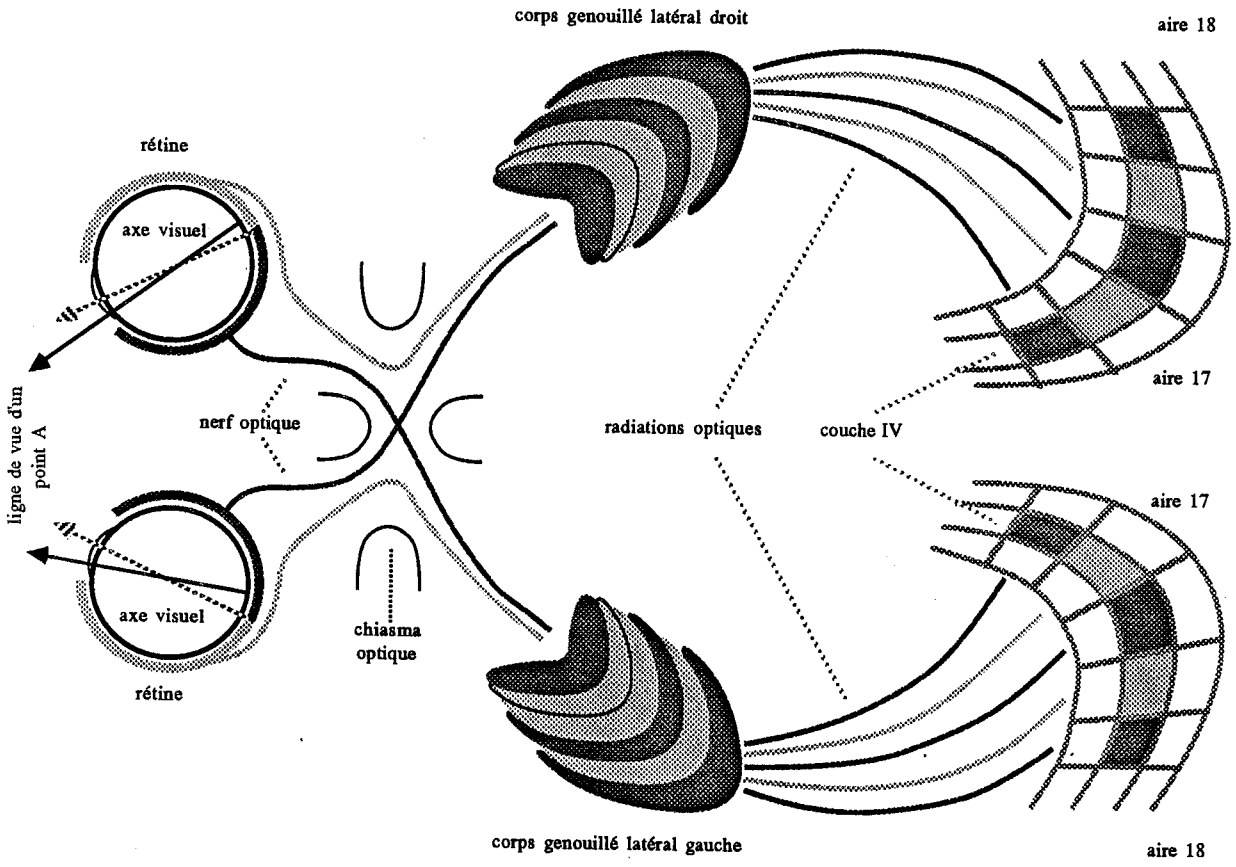
Transmis aux cellules des corps genouillés latéraux, le message en provenance de la rétine y est l'objet d'une réorganisation plus que d'un véritable traitement, du fait que les neurones des corps genouillés latéraux ont les mêmes caractéristiques fonctionnelles que ceux de la rétine (à centre "ON" ou "OFF"). D'un point de vue fonctionnel, la différence majeure avec le niveau rétinien est d'ordre quantitatif dans le sens où les neurones des corps genouillés latéraux sont encore plus sensibles aux différences d'illumination entre le centre et la périphérie de leur champ récepteur.

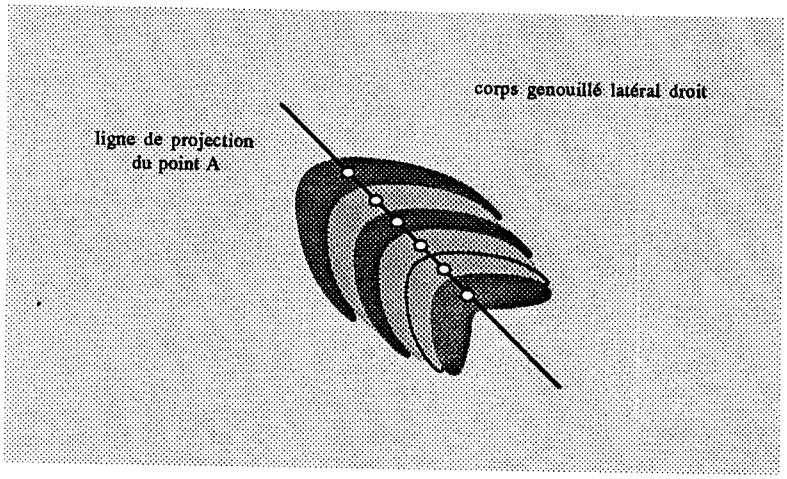
L'organisation du message visuel, elle, est par contre totalement différente (cf. figure A.08). L'asymétrie visuelle observée (unique chez l'homme), est provoquée au niveau du chiasma optique où se croisent les nerfs optiques : la partie droite (resp. gauche) de la scène observée par chaque oeil est transmise au corps genouillé latéral gauche (resp. droit) avant d'être "vue" par l'hémisphère gauche (resp. droit) du cerveau. Chaque corps genouillé latéral est composé de six couches, représentations "cartographiques" du demi-champ visuel, de sorte que chacune de ces couches, issue d'un oeil donné (les couches 2,3,5 du corps genouillé latéral droit sont issues de l'oeil droit alors que les couches 1,4,6 sont issues de l'oeil gauche) "voit" la totalité du demi-champ visuel. De plus, ces couches sont parfaitement superposées : si l'on descend une électrode perpendiculairement aux couches, les neurones rencontrés dans les différentes couches sont relatifs au même champ récepteur (**"colonnes de dominance oculaire"**). En définitif, on peut dire que le corps genouillé est un relais des voies visuelles au niveau duquel le message visuel n'est pas radicalement transformé.

Les fibres issues des neurones des corps genouillés latéraux passent à travers la substance blanche des hémisphères cérébraux, forment ce que l'on décrit couramment du terme de radiations optiques, et se terminent dans une zone délimitée de la substance grise cérébrale appelée **aire visuelle primaire** (aussi appelée "aire striée " ou "aire 17"). Au voisinage de l'aire 17 se situent les aires visuelles secondaires (aires 18 et 19) qui ne reçoivent des informations visuelles que de manière indirecte, à partir de l'aire 17 mais aussi à partir de trajets sous-corticaux nettement plus complexes.

2.3.3 ... mais une organisation parallèle en modules fonctionnels de base

Au niveau de l'organisation du cortex visuel primaire, on retrouve un ordre topographique précis, une véritable "carte" de la rétine (cf. figure A.07) : la moitié supérieure du champ visuel se projette sur la zone inférieure de l'aire 17, la moitié inférieure sur la zone supérieure. La région centrale de la rétine y est cartographiée à l'échelle la plus grande et les régions périphériques correspondent à des représentations corticales de plus en plus petites au fur et à mesure que l'on s'éloigne de la région centrale. Les neurones de l'aire visuelle 17 (très nombreux et de formes très diverses), sont distribués dans six couches superposées, traditionnellement numérotées de I à VI par les neuroanatomistes (la couche I étant la plus





(d'après [IMBER-85a] pp. 601-608)

Figure A.08: Réorganisation de l'image rétinienne

externe). Les fibres issues des corps genouillés latéraux se terminent principalement au niveau des neurones de la couche IV. Les nombreuses connexions entre neurones de couches différentes permettent alors à l'information visuelle de suivre un trajet particulièrement compliqué (cf. figure A.09.a).

Le fonctionnement du cortex n'est pas simple à expliquer en quelques lignes. Notons simplement qu'il est composé de "cellules simples" et de "cellules complexes" qui ont en commun de réagir préférentiellement à une orientation donnée du stimulus lumineux. Lorsqu'on descend une électrode perpendiculairement aux couches de l'aire 17, on ne rencontre que des neurones préférant tous la même orientation de stimulus. Il y a donc dans le cortex des **"colonnes d'orientation"** analogues aux colonnes de dominance oculaire que nous avons explicitées antérieurement.

Le schéma anatomique d'organisation fonctionnelle du cortex visuel primaire proposé par Hubel et Wiesel [HUBEL-77] consigne ces différents résultats (cf. figure A.09.b). Considérons le volume de tissu cortical (aussi appelé "hypercolonne ") formé de la manière suivante : une colonne de dominance oculaire gauche et une colonne de dominance oculaire droite constituent une des dimensions de ce volume; un ensemble de colonnes dont l'orientation varie de façon continue, mais moyennant une discrétisation, de 0^o à 180^o forme une autre dimension du volume, la dernière dimension étant formée de l'épaisseur du cortex visuel et comprenant donc les cellules simples et complexes des six couches. A l'intérieur de ce petit volume cortical, une petite aire de champ visuel sera donc vue par l'oeil gauche et par l'oeil droit selon toutes les possibilités d'interaction binoculaire et selon toutes les orientations des contraste lumineux. Selon Hubel et Wiesel qui ont mené leurs expériences sur des macaques, une telle hypercolonne constitue le module fonctionnel de base du cortex strié. Autrement dit le cortex visuel de l'aire 17 tout entier peut se concevoir comme une espèce de chaussée pavée de ces petits modules élémentaires, et l'ensemble de la scène est analysée au moyen de ce dallage.

2.3.4 Un fonctionnement par spécialisations successives

Quel type d'analyse globale d'une scène est réalisé par le cortex visuel? Nous avons évoqué précédemment une circulation compliquée des signaux nerveux dans le cortex strié qui vont de la couche IV aux couches II et III, de là descendent vers les couches V puis VI pour finalement revenir dans la couche IV. On peut supposer que **cette circulation permet d'extraire progressivement des attributs de plus en plus abstraits** des messages visuels : Différentes descriptions de caractéristique sont élaborées à partir des mesures effectuées par les cellules.

Lorsqu'elles sont complètes, elles sont envoyées en d'autres localisations cérébrales, mais nous ne savons pas encore (et tout spécialement du point de vue de la neurophysiologie) quel est le code par lequel ces informations sur les caractéristiques sont rédigées. La seule présomption forte est qu'il existe sans doute des phases ultérieures de traitement qui extraient de cette masse de données une description complète de l'ensemble de l'image.

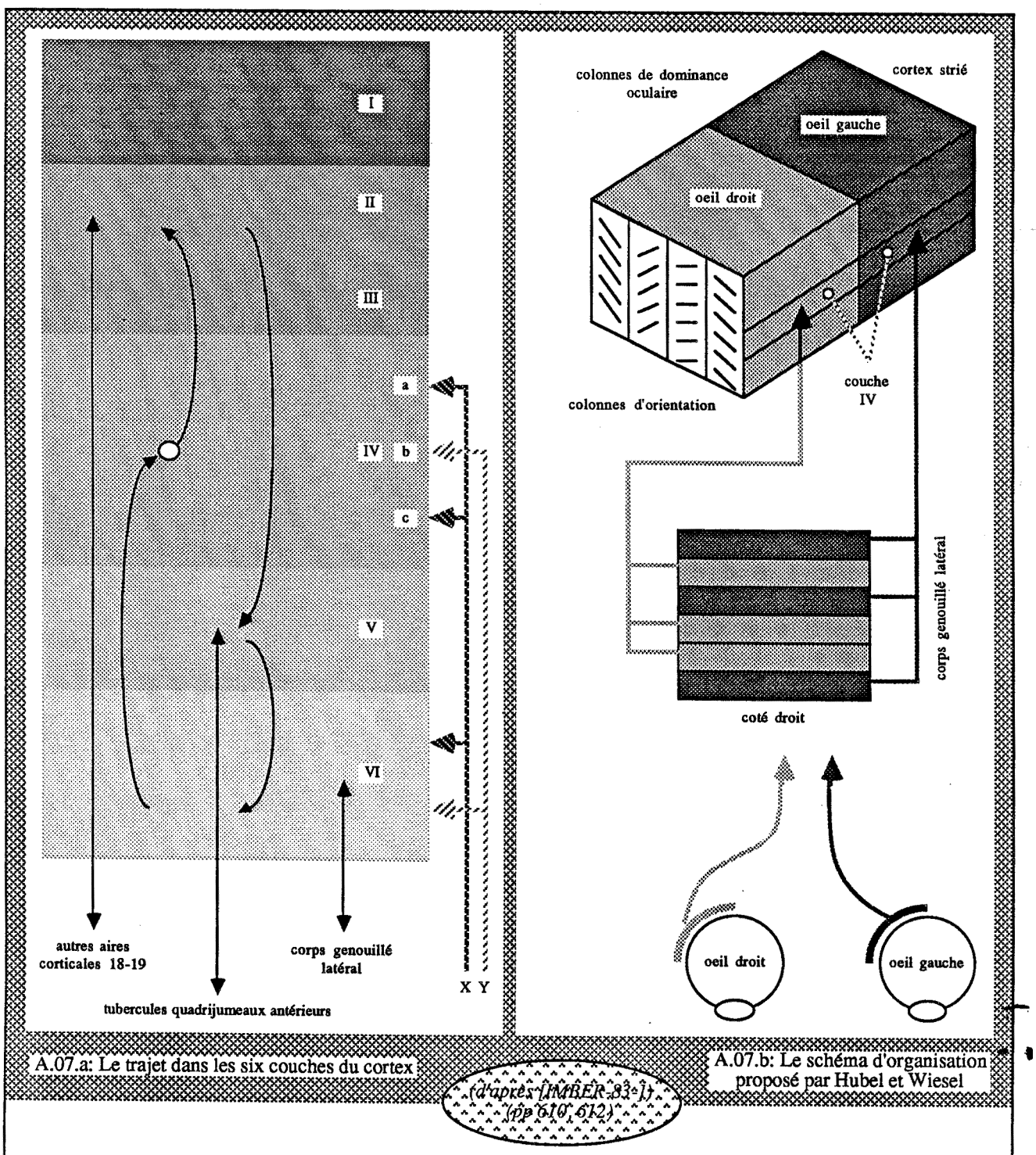


Figure A.09: Trajet et Organisation de l'information visuelle dans le cortex

La destination des fibres issues du cortex dépend de la position exacte dans les différentes couches du neurone d'origine. Par exemple, on pense généralement que les neurones de la couche VI jouent un rôle déterminant dans la mise au point en position du message visuel, alors que ceux de la couche V prennent part à l'élaboration et à l'exécution des réactions sensori-motrices. Quant aux neurones des couches II et III, leurs axones se distribuent de façon ordonnée dans les régions visuelles voisines. Il semble que les attributs différents codés dans l'aire primaire 17 soient redistribués de façon différentielle dans ces diverses aires secondaires. Ainsi retrouve-t-on abondamment dans l'aire 19 du macaque des neurones activés spécifiquement par un stimulus en mouvement dans une direction donnée. De même, l'aire visuelle secondaire 18 du macaque contient au moins quatre représentations visuelles différentes, dont l'une serait spécialisée dans le traitement de la couleur, alors que d'autres seraient pour leur part plus spécialement composées de neurones binoculaires "précisément accordés" à une distance déterminée (disparité sélective).

Les explications des résultats issus des expériences neurophysiologiques possèdent l'avantage d'être matérielles et matériellement causales : il s'agit de **constater** une relation causale. Tel n'est pas le cas des modèles issus des expérimentations psychologiques : elles aussi mettent en relation des évènements matériels et des réponses matérielles (comportement), mais la relation fonctionnelle causale doit évidemment être **interpolée théoriquement**. Mais il ne faut pas pour autant croire que les modèles issus de la psychologie ne sont que des simples vues de l'esprit : leur rôle pour l'étude de la vision est en effet analogue à celui de la déduction mathématique en physique.

Cette différence entre neurophysiologie et psychologie mise à part, il ne faut pas oublier que la neuro-physiologie n'est "concrète" que dans la mesure où elle en est à ses débuts : et dans la mesure où elle deviendra exacte, elle prendra inévitablement une forme plus physico-mathématique, et donc de plus en plus "abstraite".

2.4 Les apports de la Psychologie Expérimentale

La théorie de l'écran intérieur n'est pas satisfaisante par le simple fait que ce que l'on voit diffère souvent radicalement de ce qui se trouve devant nos yeux, ne serait-ce que parce que la projection rétinienne est à deux dimensions et qu'elle traite un espace tridimensionnel. Même en admettant par exemple que les scènes tridimensionnelles complexes puissent être recréées dans le tissu cérébral d'une façon matérielle directe, une telle théorie ne peut alors pas expliquer le fait que nous puissions construire dans notre cerveau un modèle tridimensionnel d'objets qui sont pourtant physiquement impossibles, comme l'escalier des moines du dessin d'Escher intitulé "Montant et Descendant" (cf. figure A.10). Les psychologues travaillant dans le domaine de la perception regardent les phénomènes d'illusion comme un moyen important pour essayer de comprendre les processus perceptifs, que ceux-ci fournissent une réponse correcte ou, au contraire, conduisent effectivement à des illusions. Expérimentalement, il apparaît en effet que nombre d'illusions visuelles offrent des indices sur l'existence de mécanismes perceptifs chargés de construire

ILLUSTRATION EN ATTENTE D'AUTORISATION
DE REPRODUCTION

une description de scène explicite. Ainsi, des illusions comme celles provoquées par le dessin d'Escher nous apprennent par rapport au problème de la représentation symbolique de la notion de profondeur dans le cerveau : 1/ que de petits détails servent à construire une description explicite pour des parties locales de la scène vue dans son entier, et que la cohérence d'ensemble de la représentation n'est pas le facteur le plus important 2/ que réciproquement, la cohérence globale inhibe les contradictions locales.

Certaines des suggestions fournies par les études psychologiques sont reprises dans les paragraphes suivants. Ces thèmes, particulièrement liés avec notre étude, sont rapidement évoqués, et pour une étude plus générale, le lecteur pourra par exemple se référer à [FRISB-79] et à [PIAGE-61].

2.4.1 Une description symbolique

Ni les objets, ni leurs propriétés (comme la luminosité), ni leurs fonctions ne résident en tant que tels dans notre cerveau, pourtant appelé à fournir une description symbolique de la scène observée. Le mot "chaise" représente un objet précis fait pour s'asseoir, mais ne constitue pas l'objet lui-même tel qu'il peut apparaître dans une image. Il est vrai qu'il est difficile, et même peu naturel, de séparer la "perception d'une scène" de la "scène elle-même", mais cette distinction doit être faite pour comprendre ce qu'est la vision : lorsque cela est fait, la conclusion que la vision implique une description symbolique existant dans notre tête devient plus facile à accepter.

La facilité avec laquelle notre système visuel délivre sa description symbolique d'une scène réelle est telle qu'on peut, a priori, douter que ce processus concerne la nature de la vision. Afin de s'en convaincre, il n'est pas inutile de présenter des figures truquées qui déroutent le système visuel, et révèlent quelque chose du processus de description de scènes en action. Prenons le cas de la bande dessinée de Gustave Verbeek (cf. figure A.11). Par manque de place dans la revue où ses dessins devaient être publiés, Gustave Verbeek prit le biais de remplir chacune des cases par deux dessins, de sorte que regardé à l'endroit ou à l'envers, le dessin soit perçu de l'une ou de l'autre des deux façons possibles. Une seule image est présente, mais elle conduit le système visuel à réaliser des descriptions de scènes différentes qui dépendent de la disposition spatiale de la représentation. La théorie de l'écran intérieur a bien du mal à rendre compte du phénomène produit par l'inversion d'une telle figure "bas-en-haut", car selon cette théorie, l'inversion ne devrait produire qu'une perception de la même image, mais de bas en haut. En réalité, l'inversion modifie subtilement la nature du témoignage dans l'image rétinienne sur ce qui compose la scène, et le système visuel suit en fonction. Les deux lectures du dessin paraissent différentes et pas seulement à cause du texte qui est associé aux deux interprétations possibles de ce dessin. En fait, outre le vocabulaire employé, c'est véritablement deux scènes différentes qui sont interprétées à partir de la même image.

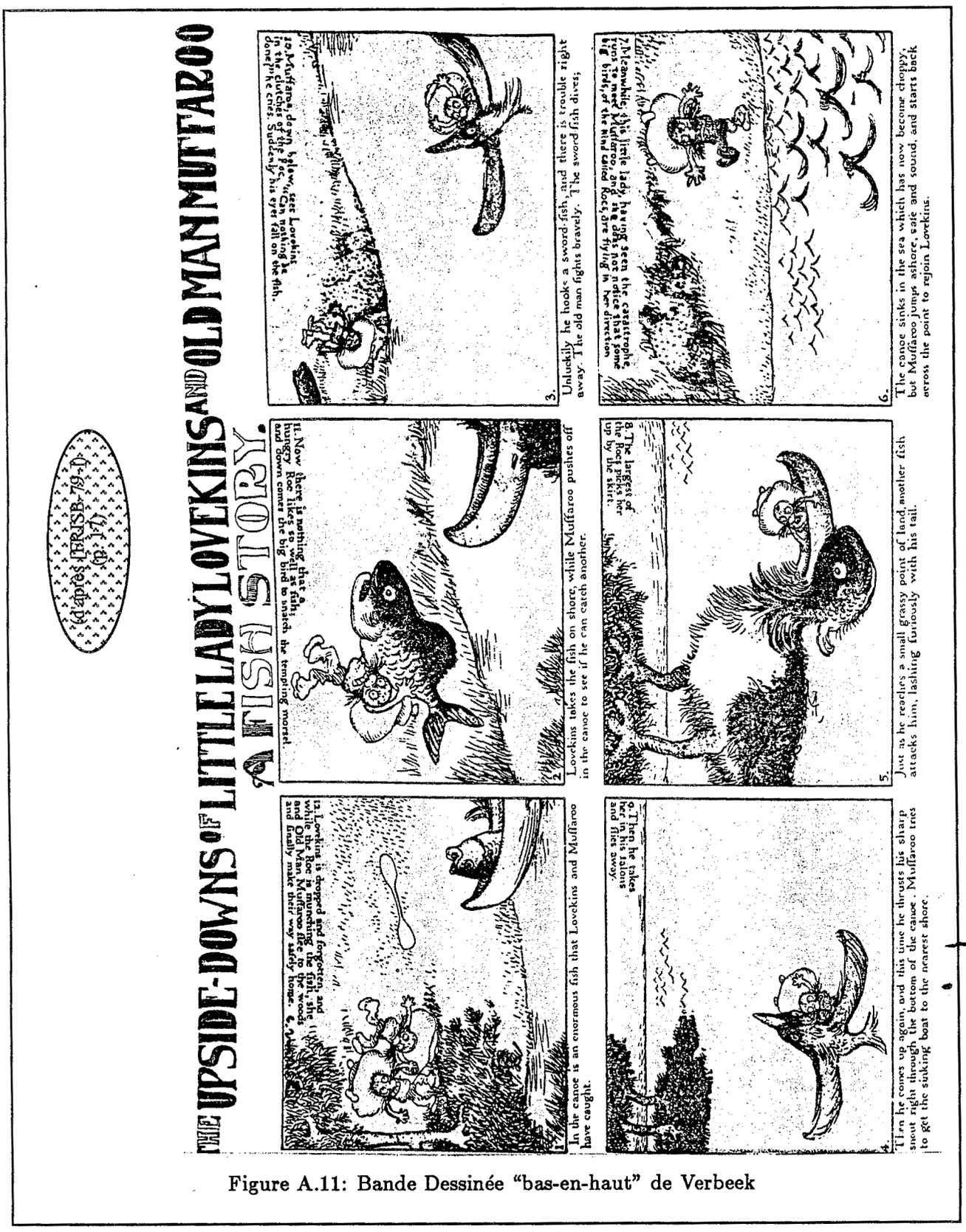


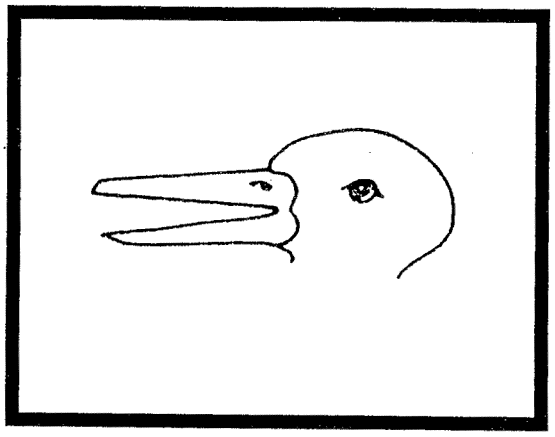
Figure A.11: Bande Dessinée "bas-en-haut" de Verbeek

2.4.2 Notion d'Invariant et importance du Contexte

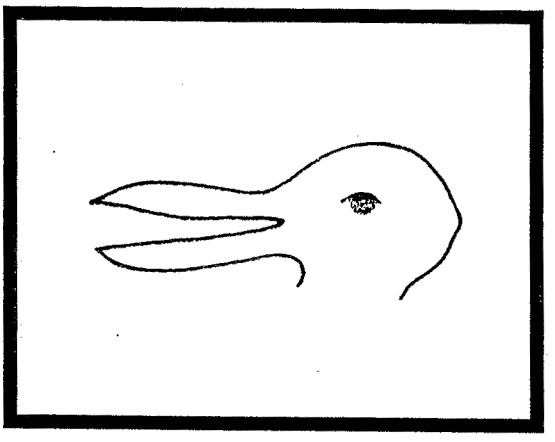
Une autre manière de déjouer les capacités de description de scène du système visuel humain est de lui fournir une donnée ambigue lui permettant d'aboutir alternativement à différentes descriptions. Prenons le cas de la célèbre série du "canard/lapin " (cf. figure A.12), ou encore la séquence animée qui constitue le générique de l'émission télévisée "THALASSA". D'une image à l'autre, de nombreux aspects de la description de scène (et en particulier de nombreux indices visuels qui composent la scène) demeurent invariants mais l'impression d'ensemble change subtilement au fur et à mesure que s'impose l'une ou l'autre des deux possibilités. Le contexte d'interprétation global le plus évident prend le pas sur l'invariance des interprétations locales. La description de scène adoptée détermine la façon dont nous voyons les images. Tout comme pour les figures "bas-en-haut", ce n'est pas simplement une question de terminologie : l'ensemble de la description de scène, incluant aussi bien les éléments décrits que l'interprétation globale traduit vraiment, à chaque fois, l'expérience visuelle.

2.4.3 Une structuration en Niveaux

La principale faiblesse de la théorie de l'écran intérieur se traduit par le fait qu'une description de scène en termes de brillance en chaque point est une forme extrêmement limitée de description. Les nombres constituent une description dans la mesure où ils rendent explicites les niveaux de gris de l'image d'entrée, c'est-à-dire opérationnels pour les phases ultérieures du traitement d'images. Mais la description d'une scène existant dans notre tête ne se limite pas simplement à la brillance individuelle des points de la scène observée, mais nous en révèle considérablement plus : nous pouvons savoir quels sont les objets que nous sommes en train de regarder, ou comment nous nous montrons capables d'en décrire les principales caractéristiques (forme, matière, mouvement, dimensions) ou encore leurs relations spatiales réciproques. Ceci est totalement négligé par la théorie de l'écran intérieur et semble pourtant bien constituer la base de la vision. Parvenir à une restitution aussi parfaite est l'aboutissement d'un processus incroyablement complexe qui requiert un grand nombre d'interprétations toujours plus fouillées à partir de la qualitativement "pauvre" mais potentiellement "riche" information contenue aux pixels d'une image à niveaux de gris. En réalité, une description par niveaux de gris n'est que le premier et le plus médiocre des buts, tant du point de vue de l'étude psychologique du système visuel que de son étude informatique.



(d'après [FRISB-79-1] (p. 19))



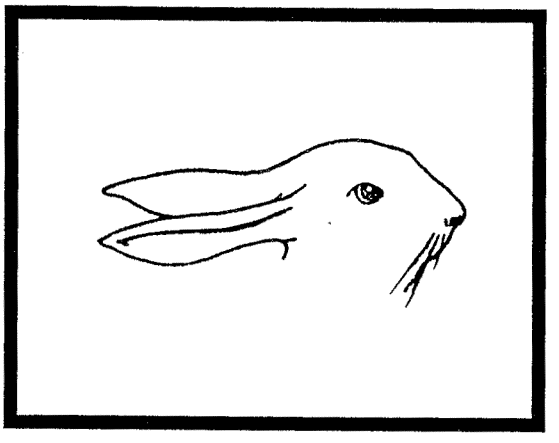


Figure A.12: La série dite du "canard-lapin"

Chapitre A.III

Notre Approche

1 La nécessité de Niveaux de Répresentation en V.O.

Pour de nombreuses réalisations informatiques, la distinction entre plusieurs niveaux est souvent arbitraire et il n'existe aucune sémantique qui puisse justifier de la présence des informations explicitées aux niveaux distingués. Mais d'ailleurs, pourquoi faut-il absolument, du point de vue de l'informaticien, qu'un système de compréhension de scènes soit structuré en niveaux? Et surtout quels niveaux doivent être distingués et pourquoi? C'est à ces questions que nous répondons dans les paragraphes suivants.

1.1 L'analogie avec le Système Visuel Humain

Sans constituer une réelle justification à l'existence de niveaux informatiques, l'analogie avec le système humain est "troublante", d'où notre position assez ambigue à propos de l'anthropomorphie des systèmes de compréhension de scènes. Etudions un moment la figure A.13, second des manuscrits retrouvés dans les archives de l'auteur : mise à part la dernière des cinq descriptions proposées par Freddy du même lieu de villégiature, les autres peuvent apparaître pour un être humain sinon arbitraires, du moins tout à fait inadaptées, à moins évidemment qu'elles ne constituent l'objectif explicitement visé. C'est-à-dire que dans le cas général, il ne viendrait jamais à l'idée de quelqu'un de décrire une scène observée par une succession de représentations de plus en plus abstraites, avant de décider finalement que

la scène est suffisamment bien décrite et exprimée dans un vocabulaire assez riche et intelligible pour être compris des autres. Et si il est vrai que ce phénomène se produit lors de l'observation d'un objet éloigné qui se rapproche progressivement, la succession des descriptions est alors liée à un problème de *résolution* et non à un quelconque problème d'*abstraction*.

A partir de la figure A.13 et de la brève étude tant psychologique que neurophysiologique effectuée au chapitre précédent, il ressort que plusieurs caractéristiques du système visuel humain peuvent inspirer la construction d'un système de vision intégrant l'existence de différents niveaux de représentation. Par exemple :

la formidable compression de l'information initialement perçue par la rétine lorsqu'elle parvient au niveau du cortex, mais aussi sensible dès les premières descriptions de Freddy.

la multiplicité et la spécialité des unités de la machinerie visuelle mises en jeu dans l'interprétation de l'image perçue par Freddy.

le fossé qui existe entre la nature des informations perçues par la rétine (et même du cortex), et celle des informations qui sont nécessaires pour, par exemple, permettre à Freddy d'affirmer que "le prisme rouge est posé sur un parallélépipède blanc cassé", ou encore qu'"il s'agit d'une maison".

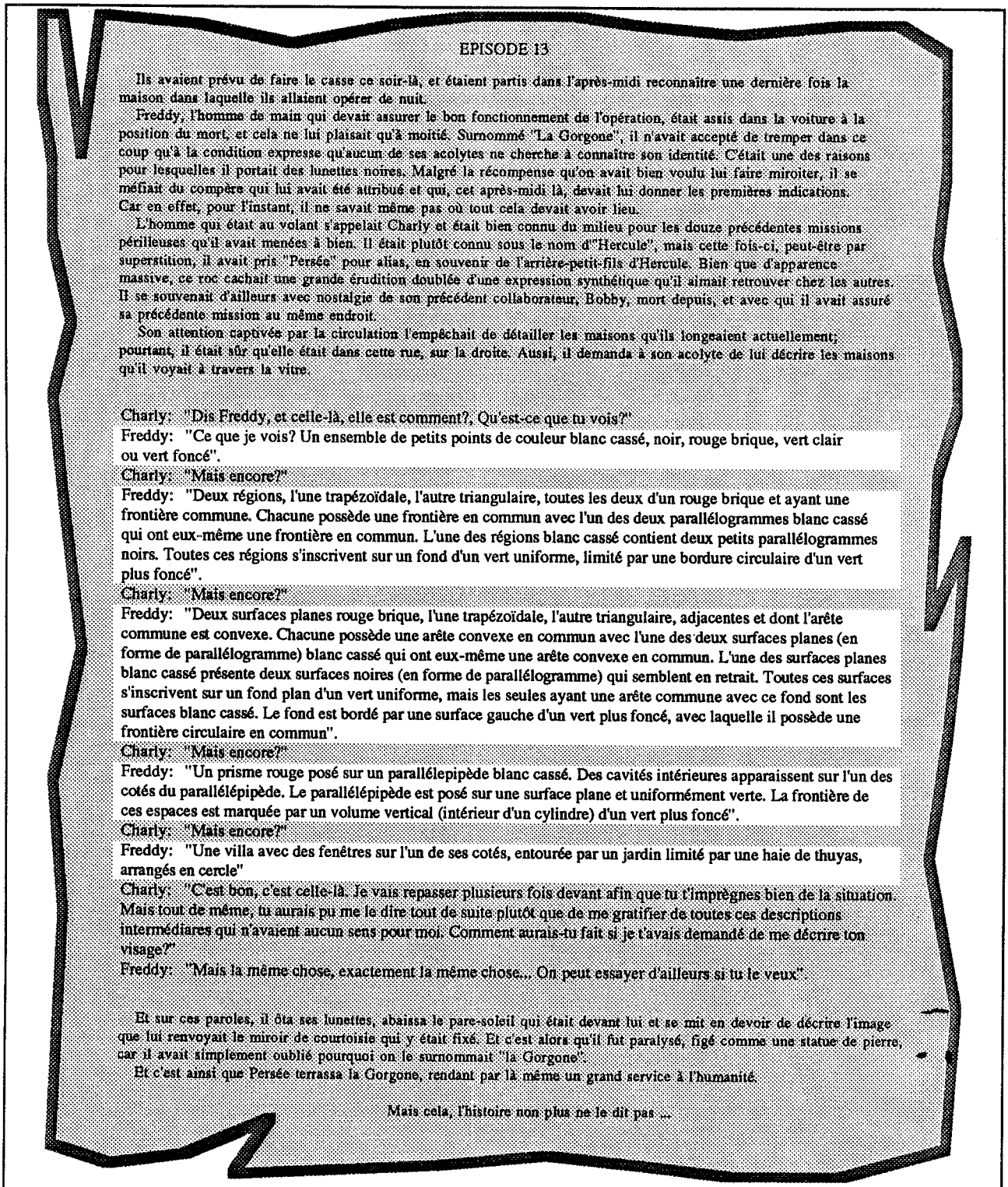
l'écart dimensionnel qui existe entre les deux dimensions des images initialement perçues par les yeux, et les trois dimensions de la description en volumes (prisme, parallélépipède).

la différence entre les référentiels par rapport auxquels sont exprimées les descriptions. Dans le cas de l'image perçue par les yeux de Freddy, la description est liée à l'observateur (viewer-centered), alors que dans le cas de la description en volumes élémentaires, la description est liée aux objets (object-centered).

la difficulté de compréhension d'une scène décrite à un niveau de formulation trop peu élevé, comme les premières descriptions de Freddy qui exaspèrent tant Charly. Cette incompréhension se traduit par l'impossibilité qu'a Charly à choisir entre les différentes interprétations que lui suggère chacune de ces descriptions de bas niveau.

la capacité humaine à décrire explicitement les qualités attachées à chaque caractéristique de chaque objet présent dans le monde observé, sa forme, sa couleur, sa distance, dire s'il est mobile ou non, etc ... Nous avons simultanément conscience du fait que chaque caractéristique est une seule entité, et qu'un ensemble de caractéristiques est constitué d'éléments d'un tout perceptif plus vaste.

D'une façon ou d'une autre, les résultats des analyses séparées doivent être réunis et chaque description de caractéristique d'une propriété doit être reliée à celle d'une autre. Nous l'avons déjà dit : "nous croyons que les domaines de la psychologie et de la neurophysiologie peuvent contribuer non seulement à comprendre



EPISODE 13

Ils avaient prévu de faire le casse ce soir-là, et étaient partis dans l'après-midi reconnaître une dernière fois la maison dans laquelle ils allaient opérer de nuit.

Freddy, l'homme de main qui devait assurer le bon fonctionnement de l'opération, était assis dans la voiture à la position du mort, et cela ne lui plaisait qu'à moitié. Surnommé "La Gorgone", il n'avait accepté de tremper dans ce coup qu'à la condition expresse qu'aucun de ses acolytes ne cherche à connaître son identité. C'était une des raisons pour lesquelles il portait des lunettes noires. Malgré la récompense qu'on avait bien voulu lui faire miroiter, il se méfiait du compère qui lui avait été attribué et qui, cet après-midi là, devait lui donner les premières indications. Car en effet, pour l'instant, il ne savait même pas où tout cela devait avoir lieu.

L'homme qui était au volant s'appelait Charly et était bien connu du milieu pour les douze précédentes missions périlleuses qu'il avait menées à bien. Il était plutôt connu sous le nom d'"Hercule", mais cette fois-ci, peut-être par superstition, il avait pris "Persée" pour alias, en souvenir de l'arrière-petit-fils d'Hercule. Bien que d'apparence massive, ce roc cachait une grande érudition doublée d'une expression synthétique qu'il aimait retrouver chez les autres. Il se souvenait d'ailleurs avec nostalgie de son précédent collaborateur, Bobby, mort depuis, et avec qui il avait assuré sa précédente mission au même endroit.

Son attention captivée par la circulation l'empêchait de détailler les maisons qu'ils longeaient actuellement; pourtant, il était sûr qu'elle était dans cette rue, sur la droite. Aussi, il demanda à son acolyte de lui décrire les maisons qu'il voyait à travers la vitre.

Charly: "Dis Freddy, et celle-là, elle est comment?, Qu'est-ce que tu vois?"

Freddy: "Ce que je vois? Un ensemble de petits points de couleur blanc cassé, noir, rouge brique, vert clair ou vert foncé".

Charly: "Mais encore?"

Freddy: "Deux régions, l'une trapézoïdale, l'autre triangulaire, toutes les deux d'un rouge brique et ayant une frontière commune. Chacune possède une frontière en commun avec l'un des deux parallélogrammes blanc cassé qui ont eux-même une frontière en commun. L'une des régions blanc cassé contient deux petits parallélogrammes noirs. Toutes ces régions s'inscrivent sur un fond d'un vert uniforme, limité par une bordure circulaire d'un vert plus foncé".

Charly: "Mais encore?"

Freddy: "Deux surfaces planes rouge brique, l'une trapézoïdale, l'autre triangulaire, adjacentes et dont l'arête commune est convexe. Chacune possède une arête convexe en commun avec l'une des deux surfaces planes (en forme de parallélogramme) blanc cassé qui ont eux-même une arête convexe en commun. L'une des surfaces planes blanc cassé présente deux surfaces noires (en forme de parallélogramme) qui semblent en retrait. Toutes ces surfaces s'inscrivent sur un fond plan d'un vert uniforme, mais les seules ayant une arête commune avec ce fond sont les surfaces blanc cassé. Le fond est bordé par une surface gauche d'un vert plus foncé, avec laquelle il possède une frontière circulaire en commun".

Charly: "Mais encore?"

Freddy: "Un prisme rouge posé sur un parallélepipède blanc cassé. Des cavités intérieures apparaissent sur l'un des cotés du parallélépipède. Le parallélépipède est posé sur une surface plane et uniformément verte. La frontière de ces espaces est marquée par un volume vertical (intérieur d'un cylindre) d'un vert plus foncé".

Charly: "Mais encore?"

Freddy: "Une villa avec des fenêtres sur l'un de ses cotés, entourée par un jardin limité par une haie de thuyas, arrangés en cercle"

Charly: "C'est bon, c'est celle-là. Je vais repasser plusieurs fois devant afin que tu t'imprègnes bien de la situation. Mais tout de même, tu aurais pu me le dire tout de suite plutôt que de me gratifier de toutes ces descriptions intermédiaires qui n'avaient aucun sens pour moi. Comment aurais-tu fait si je t'avais demandé de me décrire ton visage?"

Freddy: "Mais la même chose, exactement la même chose... On peut essayer d'ailleurs si tu le veux".

Et sur ces paroles, il ôta ses lunettes, abaissa le pare-soleil qui était devant lui et se mit en devoir de décrire l'image que lui renvoyait le miroir de courtoisie qui y était fixé. Et c'est alors qu'il fut paralysé, figé comme une statue de pierre, car il avait simplement oublié pourquoi on le surnommait "la Gorgone".

Et c'est ainsi que Persée terrassa la Gorgone, rendant par là même un grand service à l'humanité.

Mais cela, l'histoire non plus ne le dit pas ...

Figure A.13: Compréhension informatique d'une scène

le fonctionnement du système visuel humain, mais aussi à guider l'élaboration d'un modèle informatique de la vision capable d'effectuer une compréhension de scène". C'est la raison pour laquelle toutes les caractéristiques auxquelles nous venons de faire allusion, si elles sont valables pour le système humain, le sont dans le cas d'un système de vision anthropomorphe, c'est-à-dire dans le cas d'un système qui analyse des images qui sont des illustrations planes de scènes tridimensionnelles. **Nous considérons ainsi le système visuel humain non pas en tant que modèle à réaliser, ce qui serait utopique avec la technologie actuelle, mais comme une source d'inspiration.** Ce qui veut aussi dire que ni le bien-fondé, ni la valeur des méthodes de vision spécifiquement informatiques (méthodes actives de formation d'image 3-D) ne sont remis en question. Dans ce dernier cas, si les analogies avec le système visuel humain ne s'appliquent effectivement pas toutes, il n'en n'est pas moins vrai que les méthodes concernées doivent elles aussi être capables, entre autres choses, de maîtriser la grande quantité d'informations numériques qu'elles manipulent pour aboutir à une description de scène "intelligible".

C'est un fait certain, et de nombreuses expérimentations le montrent, il est possible de distinguer plusieurs niveaux de représentation dans le processus visuel humain. Mais **vouloir justifier de l'existence de niveaux en vision par ordinateur par la seule existence de niveaux dans le processus visuel humain ne peut suffire.** Et même **si on admet l'existence informatique de différents niveaux de représentation, rien ne permet d'affirmer a priori que ces niveaux sont identiques à ceux du processus visuel humain.** Dans les deux cas, une égalité des fins ne justifie pas une égalité des moyens. Ce point très important est illustré par les autres justifications à l'existence et à la spécification de niveaux informatiques que nous décrivons maintenant. Ces nouvelles justifications ne sont en rien liées à un quelconque anthropomorphisme, même si, par souci de clarté, nous continuons à utiliser des analogies avec le système visuel humain. Ce qui est le plus important sans doute, c'est que toutes ces justifications concourent à montrer que des niveaux intermédiaires existent, et qu'ils ne sont pas le résultat d'un "découpage" arbitraire.

1.2 Un pont pour la maîtrise de la complexité

Tout comme la vision humaine, la vision artificielle est confrontée au **traitement d'une information massive et redondante.** Dans une image rétinienne, ce sont des millions de cônes et de bâtonnets qui captent l'information. Dans le cas de la vision par ordinateur, c'est une matrice de pixels. Et il suffit de se placer dans le simple cas d'une unique image noir et blanc digitalisée sur 16 niveaux de gris en 512 x 512 pixels (ce qui correspond aux valeurs de notre première expérimentation) pour se rendre compte qu'un objectif comme celui de reconnaître "une maison" à partir de ces 512 x 512 informations entières semble a priori démesuré. La complexité du phénomène provient du fait que de nombreux facteurs interviennent simultanément dans la composition d'une seule image, et donc a fortiori pour sa compréhension. L'apparence d'un objet dans une image est influencée par la nature de la texture

de l'objet, l'angle de la caméra et ses caractéristiques, la nature et l'intensité de la lumière ambiante, l'angle de la source lumineuse, les conditions atmosphériques, et par bien d'autres facteurs encore. Parfois ces facteurs se masquent l'un l'autre, parfois ils se confondent. Tous ces facteurs interagissant en chaque pixel, il est très difficile sinon impossible de déterminer la contribution de chacun d'eux dans la valeur attachée à chaque pixel.

Nous nous trouvons donc face à un problème classique dans les domaines scientifiques : la résolution de problèmes complexes où la combinatoire est élevée. La justification à l'existence des niveaux qui est ici présentée est donc un argument purement scientifique. Une des approches possibles consiste à admettre qu'**une image, comme tout phénomène complexe, ne peut être comprise par une simple extrapolation des propriétés de ses composants élémentaires que sont ses pixels. Il s'agit inversement de suivre une démarche de décomposition de problèmes en sous-problèmes.** Nous avons vu que le problème de compréhension est divisé "en largeur" par l'étude isolée de caractéristiques. De même, il est récursivement séparé "en profondeur" jusqu'à aboutir à des problèmes suffisamment élémentaires pour être préhensibles et pour qu'on leur connaisse des solutions, si imparfaites soient-elles. C'est-à-dire que pour comprendre la scène observée et passer ainsi de la rive IMAGE à la rive SCENE, on édifie un pont à plusieurs arches de pertinence croissante par rapport à l'objectif de description de scène, suivant un plan préétabli mais suffisamment général pour qu'il puisse, dans tous les cas d'étude, tout à la fois maîtriser le flôt des informations initiales et assurer la description de scène finale.

1.3 La représentation de l'Abstraction et de la Décentration

L'image initiale est dépendante de nombreux facteurs, qu'ils relèvent des "objets", de l'"observateur" (notion incluant l'outil de perception, i.e. le "capteur"), ou de l'"environnement d'observation". Au moins trois types de facteurs rentrent donc en jeu, et cependant, seule la scène, au sens des "objets" qui la composent nous intéresse! Il est donc absolument nécessaire que la description finale ne dépende en aucune façon des deux autres facteurs parasites qui gênent l'inférence des objets de la scène observée et de leurs relations : autrement dit, il faut faire "abstraction" (dans son sens courant) de ces facteurs.

Plus précisément, l'élimination de ces facteurs indésirables correspond selon nous à la satisfaction de deux critères, par ailleurs largement étudiés en neuro-physiologie et en psychologie, et que nous reprendrons ici avec des *définitions qui nous sont personnelles* : l'**abstraction** (par rapport à l'"environnement") et la **décentration** (par rapport à l'"observateur"). La justification présentée ici ne relève que du propre phénomène visuel et de son implantation.

1.3.1 La notion d'Abstraction

abstraction : **fait de considérer à part un élément qualitatif ou relationnel d'une représentation, en portant spécialement l'attention sur lui sans s'intéresser aux autres (i.e. indépendamment de l'"environnement**

") **. Par rapport à l'acception habituelle d'abstraction, il s'agit donc d'une abstraction SUR l'élément de représentation POUR lui-même. L'abstraction s'accompagne nécessairement d'un changement de type de représentation.** En inférence de formes à partir de contours telle que nous l'avons abordée pour notre première expérimentation, passer de la notion de STRUCTURES DE LIGNES DROITES extraites d'une image à celle de SURFACES PLANES de l'espace réel pouvant être adjacentes ou s'occulter correspond essentiellement à une abstraction. De même, pour notre seconde expérimentation, passer de la notion de SURFACES CYLINDRIQUES à celle d'OBJETS FILIFORMES correspond aussi à une abstraction.

Intuitivement, une scène peut être décrite à différents niveaux d'"abstraction". Plus la description est évoluée, c'est à dire plus grand a été l'effort de compréhension de la scène étudiée, plus l'être humain a de facilités à reconnaître cette scène. Il est clair que les processus sensoriels de très bas niveau ignorent l'abstraction : en présence d'un objet et des éléments proches, ils ne peuvent pas ne pas appréhender simultanément tout le domaine restreint ainsi délimité, c'est-à-dire qu'ils ne peuvent pas se borner à retenir certains caractères ou éléments de l'objet, en ignorant les autres. Le propre des processus de plus haut niveau est au contraire de choisir, au sein de ce qui est donné, ce qui est nécessaire pour résoudre le problème de raisonnement en jeu, et donc à dépasser le donné.

1.3.2 La notion de Décentration

décentration : **fait de libérer un élément qualitatif ou relationnel d'une représentation de ses liens spatio-temporels avec le sujet percevant (i.e. avec l'"observateur", par opposition à l'"objet" observé). Par rapport à l'acception habituelle de l'abstraction, il s'agit donc d'une abstraction DE l'élément de représentation PAR RAPPORT à l'espace et au temps. La décentration s'accompagne nécessairement d'un changement de type de référentiel.** En stéréovision simple telle que nous l'avons abordée pour notre seconde expérimentation, l'inférence d'OBJETS FILIFORMES est surtout liée à l'exécution d'un processus (mise en correspondance et localisation dans l'espace réel) qui correspond essentiellement à une décentration. De même, chez Marr, le passage du niveau CROQUIS 2,5-D (exprimé dans un repère lié à l'observateur) au niveau OBJET (exprimé dans un repère lié aux objets) contient une forte décentration.

Les processus sensoriels de bas niveau sont subordonnés à des conditions limitatives de proximité dans l'espace et dans le temps. Nous l'avons dit pour l'"abstraction", regardant d'un certain point de vue une touffe d'herbe dans un jardin on ne peut pas ne pas percevoir simultanément les touffes voisines; mais d'un point de vue de la "décentration", on ne peut plus voir en même temps un arbre éloigné sur la gauche ou la maison qui est dans son dos. De plus, les éléments perçus simultanément, grâce à leur proximité apparente, entrent en interaction immédiate les uns avec les autres, d'où un ensemble de déformations possibles. Au contraire, les processus raisonnés de plus haut niveau, indépendamment de tout point de vue et donc des distances spatio-temporelles, peuvent rapprocher n'importe quel élément

d'un autre, mais peuvent également dissocier par la pensée les objets voisins et raisonner sur eux en toute indépendance.

1.3.3 Discussion

A ce point de l'étude , il faut remarquer un fait très important : a priori, et tout comme l'abstraction, c'est très progressivement et non pas brusquement que s'effectue la décentration. Pour s'en convaincre, il suffit de prendre le cas des fonctions raisonnées de haut niveau chez l'homme : la fonction symbolique leur permet des comparaisons spatio-temporelles indépendamment du contact perceptif, mais le propre de ces fonctions est précisément de rester subordonnées aux configurations spatiales et par conséquent à certaines conditions de proximité même si elles sont "élargies" : la décentration demeure donc longtemps très relative.

Les deux "mesures" de degré d'abstraction et de degré de décentration permettent selon nous de particulariser un niveau intermédiaire et de justifier son existence. Mais avec la remarque précédente, si on veut pousser jusqu'au bout l'idée de progressivité dans les notions continues d'abstraction de décentration, un problème surgit alors : en effet, le mieux semble être alors, pour représenter ces deux notions, de disposer d'une infinité de niveaux intermédiaires entre le niveau IMAGE et le niveau SCENE. Ceci est, du point de vue de l'implantation informatique, totalement irréalisable : il faudra donc nous contenter, pour représenter ces notions continues, de choisir ceux des niveaux qui sont les plus significatifs.

1.4 Niveaux d'étude et Communication avec d'autres domaines

Les processus sensoriels de bas niveau ne sont absolument pas décentrés : initialement liés à une certaine position du sujet percevant par rapport à l'objet (centration), ils sont a priori strictement individuels et souvent incommunicables. Pourtant, tout comme les processus plus raisonnés dont c'est un trait propre, il leur faut aboutir non seulement à la constitution de connaissances de plus en plus indépendantes des méthodes utilisées et du système de vision lui-même, mais surtout à la constitution de connaissances communicables, c'est-à-dire universalisables. Cela nous amène à tenir compte d'un nouveau point essentiel qui couvre et déborde cette remarque : en tant que sujet d'étude, la vision par ordinateur se suffit à elle-même, mais son intérêt est exacerbé lorsqu'elle est capable de résoudre des problèmes qu'on lui donne et qu'elle peut fournir des résultats à autrui. Dans les deux cas, un problème fort de communication entre systèmes intelligents existe.

Il se dégage donc une triple notion de "niveau de définition du problème", de "niveau d'étude du problème", et de "niveau d'expression des résultats".

1.4.1 Des Niveaux de Communication différents

Que ce soit du point de vue du niveau auquel un problème peut être énoncé au système de vision par ordinateur, ou du niveau d'expression auquel les résultats sont fournis, le problème de communication sous-jacent relève bien sûr des problèmes de communication homme - machine, et plus généralement de la classe des dialogues

entre systèmes intelligents. Et s'il est clair que le langage humain est l'un des moyens visés de communication homme - machine, il n'est pas démontré que le seul moyen de dialogue entre êtres intelligents soit le langage humain! Tenter une quelconque normalisation de ce dialogue semble d'ailleurs bien délicat puisque chaque système possède sa propre structure, ainsi que des compétences souvent spécifiques. Cet état de fait pose d'ailleurs en filigrane le délicat problème de la connaissance d'autrui en univers multi-agents [BESSI-83]. Quoiqu'il en soit, il est clair qu'il faut que des points d'entrée et de sortie pour la communication avec le monde extérieur existent au sein de ces systèmes. **Nous pensons que ces points de communication doivent correspondre à tous les niveaux de représentation inclus dans le système de vision, et qu'inversement, tout niveau de représentation est un potentiel lieu de communication avec le monde extérieur**, que ce soit le niveau de l'image (par exemple via un partage de pixels entre deux systèmes de vision), le niveau de description de la scène observée (par exemple via une relation homme-machine en langue naturelle) ou tout autre niveau intermédiaire.

Un exemple des besoins de communication se trouve tout naturellement dans le domaine des coopérations bras-oeil. Etudiées depuis très longtemps, ne serait-ce que dans le but d'un contrôle visuel pour assurer une meilleure précision des mouvements d'un manipulateur [SHIRA-73], ces applications font intervenir différents systèmes dans des domaines aussi proches mais aussi spécifiques que la Vision par Ordinateur, la Robotique et l'Intelligence Artificielle. On sait qu'un système de vision permet de détecter la présence ou l'absence d'un objet, de qualifier le mouvement des objets, de vérifier des relations symboliques entre objets. Il peut identifier les objets mais aussi les localiser. Vis-à-vis de la très grande variété de tâches qu'il peut (et doit?) assurer, le système de vision doit présenter une structure ouverte permettant un échange d'informations de niveaux d'abstraction différents. Utilisé par exemple dans un contexte d'"aide" à la robotique d'assemblage, la "recherche de préconditions" ou la "vérification de postconditions" peuvent alors concerner autant les "contacts" entre surfaces que l'"accessibilité globale" d'une région de la scène [LUX-85b]. De même pour l'opération particulière de la saisie, une "localisation précise" ou une "localisation approximative" des positions de saisie correspondent à des besoins différents et donc à des demandes d'indices de différents niveaux [TROCC-86].

1.4.2 Des Niveaux d'Etude différents

Comme tout phénomène complexe, une scène visuelle doit pouvoir être étudiée au niveau auquel on souhaite la comprendre. Inversement, les niveaux de définition de problème et d'expression des résultats peuvent conditionner le niveau d'étude d'un problème. Nous illustrons ces (in)dépendances sur un exemple issu de la physique.

Soit un gaz dans une bouteille dont on effectue une étude thermodynamique : les lois de la thermodynamique sont macroscopiques et les propriétés du gaz qui en résultent sont bien éloignées des propriétés des molécules qui le composent. Une description des effets thermodynamiques - température, pression, densité, et relations entre ces facteurs - n'est pas formulée par un grand ensemble d'équations, une pour chacune des particules concernées. De tels effets sont décrits à leur propre

niveau, celui d'une énorme collection de particules; en principe, les descriptions microscopiques et macroscopiques sont consistantes entre elles.

De même qu'en thermodynamique, pour parvenir à la compréhension d'un système aussi compliqué que le système visuel humain où les niveaux de résolution d'un même problème peuvent être différents, il faut envisager différents types d'explication à des niveaux de description différents. Ces niveaux sont, au moins en principe, liés à un tout cohérent, même si la liaison détaillée entre ces niveaux n'est pas une opération toujours immédiate. Notons au passage que cette analogie avec la thermodynamique ne cautionne pas seulement l'existence de niveaux d'étude différents en informatique, mais aussi quelle que soit la discipline dans lequel chaque niveau du problème est étudié : par ce biais, elle cautionne donc aussi le tryptique neurophysiologie - psychologie - informatique.

1.5 Le contrôle de la multiplicité des sources d'information

Tout ce que nous venons de voir reste indépendant des méthodes de Vision par Ordinateur utilisées pour inférer une description de scène. D'un autre côté, nous avons dit que seule la conjonction de plusieurs méthodes permet d'aboutir à une telle description. L'argument que nous développons ici, *peut-être le plus important*, même s'il paraît le plus évident, constitue une nouvelle justification d'ordre visuel : **à partir du moment où il est admis que chaque méthode constitue un élément de base d'un système de vision, il faut qu'il existe des points de partage, d'échange et de fusion d'informations entre méthodes visuelles étudiant des caractéristiques visuelles différentes** (cf. figure A.14). Cette notion n'est pas nouvelle : on la retrouve par exemple comme base du travail de [HANSO-78].

La remarque est bien sûr applicable pour plusieurs méthodes différentes mais l'est aussi parfois au sein d'une même méthode : dans le cas de la stéréoscopie par exemple, pourquoi faudrait-il que la mise en correspondance s'effectue au niveau des pixels des images plutôt qu'au niveau des objets mêmes qui auraient été reconnus dans chacune des deux images? Nous reviendrons au cours de la partie C sur ce problème fondamental en stéréovision, mais en réalité, ces deux types de mise en correspondance sont non seulement utiles mais aussi nécessaires (il en existe même d'autres!) : tous les deux engendrent des résultats différents, même si dans les deux cas, il s'agit toujours d'inférer des indices visuels plus abstraits et/ou plus décentrés, c'est-à-dire d'un niveau de représentation plus élevé.

2 Ce qu'il faut représenter

Qu'avons nous appris jusqu'à présent? D'un point de vue théorique, nous avons montré que différents niveaux de représentation devaient être distingués pour passer de l'IMAGE à la SCENE. En termes informatiques, nous avons successivement montré que les niveaux à distinguer :

- peuvent s'inspirer de la vision humaine sans pouvoir assurer que ce soient les mêmes,

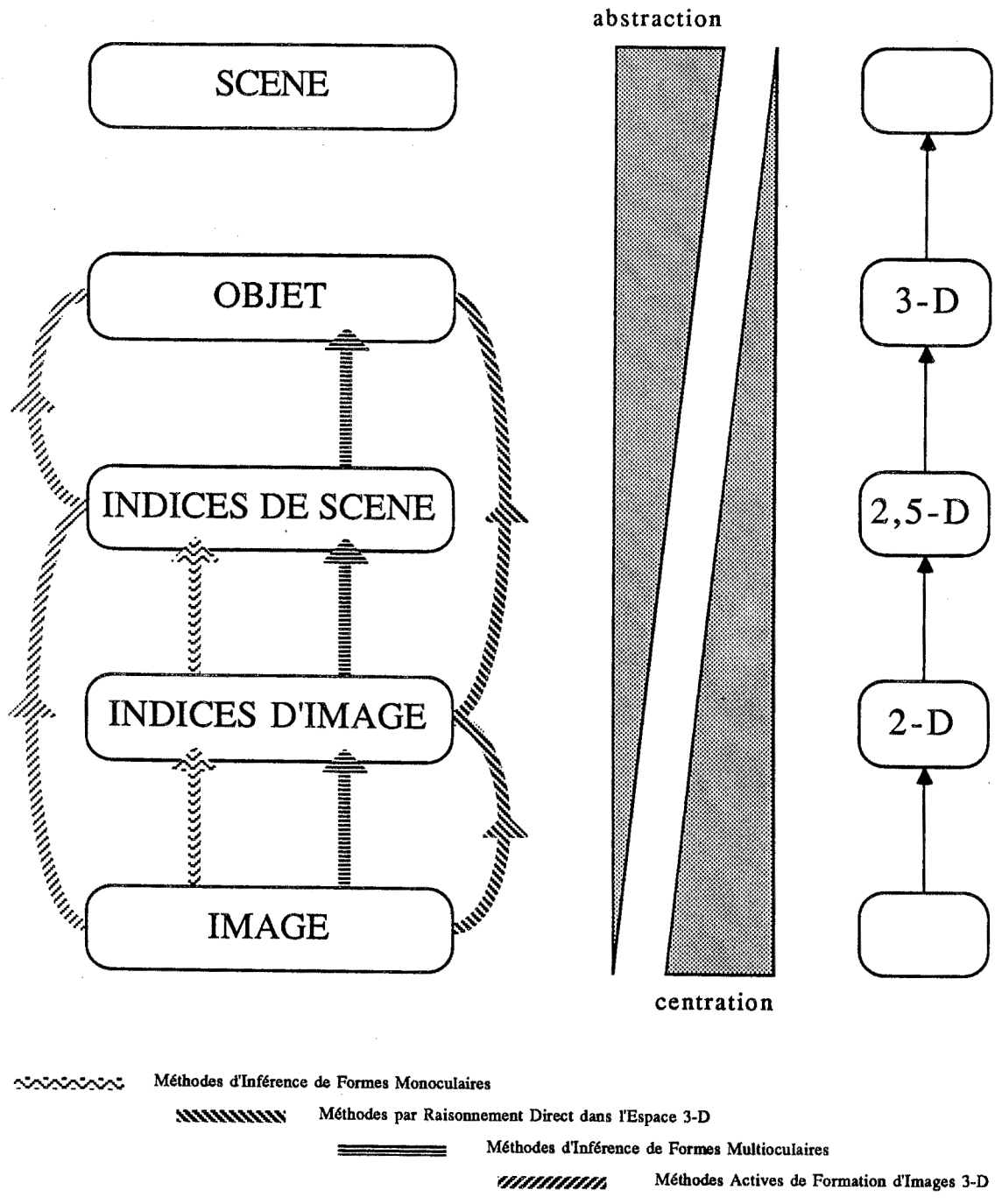


Figure A.14: Organisation des Niveaux et Types de Méthodes

o sont nécessairement en nombre fini d'un point de vue de leur implantation,
o doivent être caractérisés par leur degré d'abstraction et de décentration (ceci se traduisant par des changements de type de modèle de représentation mais aussi de type de référentiel),
o devaient correspondre à des niveaux d'étude et de communication avec l'extérieur,
o devaient permettre la synthèse de résultats obtenus par des méthodes très différentes les unes des autres.

Ayant montré l'existence de différents niveaux de représentation devant satisfaire aux remarques précédentes, il reste cependant deux questions : "COMBIEN faut-il distinguer de niveaux intermédiaires et LESQUELS?" L'expérience que nous avons acquise par nos deux expérimentations et les observations que nous venons d'effectuer dans ce chapitre, nous font distinguer CINQ niveaux de représentation, traduisant le degré atteint dans le déroulement du processus d'interprétation de l'IMAGE en la SCENE (cf. figures A.13 : exemple, A.14 : niveaux et types de méthodes, et A.15 : définition des niveaux) :

les niveaux IMAGE, INDICES D'IMAGE, INDICES DE SCENE, OBJET et SCENE.

2.1 Le Niveau "IMAGE"

"Un ensemble de petits points de couleur blanc cassé, noir, rouge brique, vert clair ou vert foncé".

La première des informations présentes à ce niveau est l'*image digitalisée*, représentée sous la forme d'une matrice de pixels. Chaque pixel de l'image est associé à une valeur discrète (un *niveau de gris*) qui traduit l'intensité de lumière observée en ce point (cas des images noir-et-blanc). A ce niveau, il est possible d'appliquer de très nombreux procédés classiques de prétraitement et d'amélioration d'images, de la correction de l'échelle des niveaux de gris aux techniques de lissage ou de rehaussement de contraste. Parmi ces dernières, se trouve en particulier l'application d'opérateurs linéaires comme celui du *gradient* ou du *Laplacien*. Par exemple, l'application de l'opérateur de gradient (approximation de la dérivée première de l'intensité observée au pixel) en chacun des points de l'image digitalisée, en association avec une valeur de seuillage, permet de construire une *image de contraste*. Nous situons donc l'image de contraste et les informations qu'elle détient au niveau IMAGE. D'autres images "intrinsèques" [BARRO-78], faisant apparaître de façon explicite des caractéristiques de l'image autres que le contraste, figurent aussi au niveau IMAGE.

Quelles que soient les informations intrinsèquement contenues dans l'image, la structure présente à ce niveau est conditionnée par le capteur et par le système de digitalisation utilisés. Il s'agit donc d'une *matrice bidimensionnelle discrète*. La taille de ce tableau est fournie par le système de digitalisation (couramment 256 x 256 ou 512 x 512) tout comme l'échelle des niveaux de gris (généralement au nombre

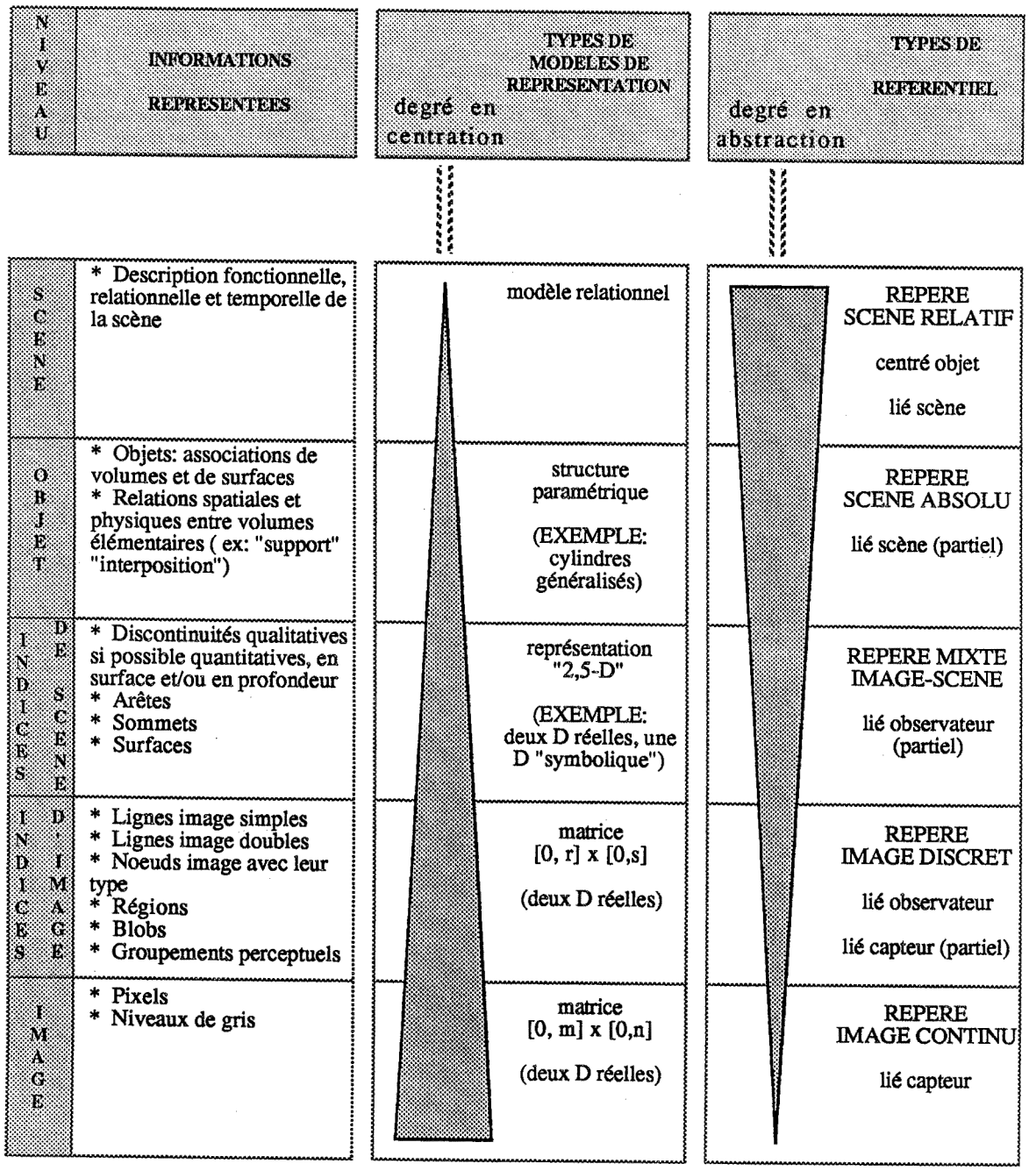


Figure A.15: Les cinq Niveaux et les informations représentées

de 16, 64 ou 256), est donc représentée dans un repère entier *lié à l'observateur et au capteur.*

2.2 Le Niveau "INDICES D'IMAGE"

"Deux régions, l'une trapézoidale, l'autre triangulaire, toutes les deux d'un rouge brique et ayant une frontière commune. Chacune possède une frontière en commun avec l'un des deux parallélogrammes blanc cassé qui ont eux-même une frontière en commun. L'une des régions blanc cassé contient deux petits parallélogrammes noirs. Toutes ces régions s'inscrivent sur un fond d'un vert uniforme, limité par une bordure circulaire d'un vert plus foncé".

Une séquence de points de contraste voisins (par exemple au sens du 4-voisinage ou du 8-voisinage), c'est-à-dire une ligne de contraste, indice de niveau image, peut être interprétée par des *outils analytiques* (par exemple une analyse de courbure) en une *ligne-image*, définie par une équation mathématique plane et dérivable, et limitée dans le cas d'une courbe ouverte par la position des deux *noeuds-image* qui constituent ses extrémités. Une *droite-image* pourra être représentée par ses deux noeuds-image extrémités, ou encore par un angle, une puissance et un intervalle, indépendamment des noeuds-image qui la délimitent. L'information contenue est identique, mais représentée différemment. **Un exemple** d'obtention d'indices d'image est donc l'approximation linéaire de séquences de points de contraste. D'autres indices, tels les *groupements perceptuels* utilisés dans les méthodes par raisonnement direct dans l'espace 3-D, figurent aussi à ce niveau de représentation.

La structure créée à ce niveau contient les noeuds-image et les lignes-image. *Légèrement décentrée par rapport au capteur*, elle est exprimée dans un *système de coordonnées réelles orthonormées, repère bidimensionnel* déduit du repère de l'image digitalisée par des connaissances sur le capteur de l'information visuelle et sur les caractéristiques de digitalisation même.

2.3 Le Niveau "INDICES DE SCENE"

"Deux surfaces planes rouge brique, l'une trapézoidale, l'autre triangulaire, adjacentes et dont l'arête commune est convexe. Chacune possède une arête convexe en commun avec l'une des deux surfaces planes (en forme de parallélogramme) blanc cassé qui ont eux-même une arête convexe en commun. L'une des surfaces planes blanc cassé présente deux surfaces noires (en forme de parallélogramme) qui semblent en retrait. Toutes ces surfaces s'inscrivent sur un fond plan d'un vert uniforme, mais les seules ayant une arête commune avec ce fond sont les surfaces blanc cassé. Le fond est bordé par une surface gauche d'un vert plus foncé, avec laquelle il possède une frontière circulaire en commun".

Nous venons de voir qu'un indice d'image est un élément d'une structure bidimensionnelle, projection d'une partie de l'espace réel. Les indices contenus au niveau INDICES DE SCENE sont au contraire les interprétations des indices d'image en tant qu'*indices du monde physique*, et s'attachent particulièrement à exhiber

la *géométrie des surfaces visibles* dans l'image. Nous parlerons par exemple des indices de scène *arêtes* et *sommets*, qui peuvent être définis "approximativement " (nous verrons que cette interprétation n'est pas une simple bijection) comme les interprétations respectives des droites-image et des noeuds-image.

La structure construite à ce niveau est une *matrice "2,5-D" hétérogène*, car exprimée dans un système de coordonnées qui se trouve à mi-chemin entre un système à deux et trois dimensions. Plus précisément, elle possède deux dimensions numériques et une dimension symbolique. Les trois dimensions sont *liées à l'observateur*, mais parfaitement *indépendantes du capteur*. La structure construite reflète non seulement *l'apparence* (projection) des surfaces visibles des objets en scène, mais aussi des *relations symboliques locales* (en adjacence et/ou en profondeur relatives) entre ces surfaces. Elle constitue, à ce titre, un cas particulier de ce que Marr appelle le croquis 2,5-D.

2.4 Le Niveau "OBJET"

"Un prisme rouge posé sur un parallélépipède blanc cassé. Des cavités intérieures apparaissent sur l'un des cotés du parallélépipède. Le parallélépipède est posé sur une surface plane et uniformément verte. La frontière des ces espaces est marquée par un volume vertical (intérieur d'un cylindre) d'un vert plus foncé".

Il s'agit ici de consigner les résultats de l'interprétation des indices de scènes à un degré d'abstraction plus élevé et dont la *décentration soit totale par rapport au capteur et soit partielle par rapport à l'observateur*. La structure des indices de scène est donc interprétée afin d'obtenir une *description des formes volumiques et surfaciques* et de leur *organisation spatiale*, représentées par une *structure paramétrique* (p.e. les "cylindres généralisés") exprimée dans un *repère scène absolu*. Cette interprétation, qui introduit la notion d'"objet", ne se réduit pas simplement à *un regroupement d'indices de scène*. Par exemple, l'indice de scène correspondant à la projection apparente de la surface gauche d'un cylindre associé à la base visible du cylindre doit être interprétée comme un volume tridimensionnel correspondant au cylindre réel observé.

Moyennant la connaissance de propriétés physiques du monde réel par le système, comme par exemple la notion de pesanteur et donc la notion de verticale, il est possible, dès le niveau OBJET, de travailler sur les formes extraites et de déduire deux types de relations : la *"relation de support"* et la *"relation d'interposition"* entre volumes élémentaires extraits. Notons d'ailleurs que la relation de support peut parfois même être précisée en regard de la stabilité physique de la liaison : un objet suspendu au plafond d'une pièce soit fait partie du plafond lui-même, soit est fixé au plafond. A l'inverse, la seule connaissance de la gravité ne permet pas de préciser plus la liaison qui existe entre deux objets apparemment posés l'un sur l'autre.

Par ailleurs, il est clair au vu de la définition de ce niveau OBJET, que le terme "objet" est à prendre dans le sens où *un objet est une description relationnelle en volumes et/ou surfaces*, et non le résultat d'une désignation. En effet, comme cela

est suggéré dans [LUX-85a], la désignation d'objet est un problème qui ne présente dans le cadre du système que nous présentons qu'un intérêt secondaire si on ne s'attache qu'à associer un nom à une forme extraite ...

2.5 Le Niveau "SCENE"

"Une villa avec des fenêtres sur l'un de ses cotés, entourée d'un jardin limité par une haie de thuyas, arrangés en cercle".

... S'il s'agit par contre d'associer à cette forme tout un contexte visuel, géométrique, historique, fonctionnel, relationnel, conventionnel, etc ..., le tout contribuant alors à cette tâche de désignation, le problème devient alors plus complexe et relève du niveau SCENE: la représentation nécessaire est alors *totalement abstraite* et surtout *parfaitement décentrée*. **Le niveau SCENE nous paraît impossible à définir du point de vue de la seule Vision par Ordinateur, tant il est lié à notre compréhension humaine du monde réel.** Pour arriver à une telle décentration, il s'agit de mettre les objets en relation avec les cadres de référence, de sorte que la centration par rapport au corps propre et la décentration en faveur des relations entre objets prend un sens global et non plus seulement local. Etant parvenu à une "compréhension de scènes", le processus de compréhension est alors achevé.

La structure obtenue, décrite dans un *repère lié à la scène*, reflète *les fonctions des objets* composant la scène et *leurs relations*. Elle peut en particulier faire apparaître des relations entre objets comme la *"probabilité d'occurence des objets"*, les *"positions habituelles et spécifiques des objets"*, ainsi que la *"taille habituelle des objets"*, ces trois notions étant définies par rapport à leur environnement reconnu : c'est ce que pensent certains psychologues [BIEDE-81] et cela reste un sujet très ouvert car peu étudié d'un point de vue du traitement informatique des images.

Remarque 1 : Notre présentation ascendante inclut une construction incrémentale ("mise à jour dynamique") des différents niveaux de représentation, à l'occasion 1/ de changements de "points d'attention", 2/ de modifications de "focalisation", ou 3/ de variations de points de vue de l'observateur. Par contre, la première remarque à cet exposé général est son absence de prise en compte d'un autre phénomène temporel, lié aux changements de scène dus au "mouvement des objets". Il semble bien délicat de prendre en compte cette quatrième dimension alors qu'il est déjà si difficile de comprendre une scène tridimensionnelle statique. Quoiqu'il en soit, tout déplacement relatif des objets de la scène par rapport au capteur engendre une complexité supplémentaire, et nous croyons que ce phénomène temporel particulier doit être répercuté à tous les niveaux de représentation, sans pouvoir dire comment.

Remarque 2 : Une seconde remarque sur la présentation générale de ces niveaux est liée au problème de la reconnaissance. Cela sera en particulier visible dans les parties B et C, la connaissance de modèles, (essentiellement géométriques dans le cas de nos expérimentations), permet d'envisager des heuristiques d'interprétation

adaptées à la reconnaissance des objets connus dans une scène donnée. Cette observation en rejoint une autre : le niveau SCENE et, en moindre mesure, le niveau OBJET, sont non seulement très peu commodes à définir d'un point de vue de la seule vision, mais encore ils sont tout aussi difficiles à préciser en dehors de tout contexte : notre présentation de ces deux niveaux se place dans le contexte de scènes de l'univers réel.

Remarque 3 : Finalement, il nous faut revenir aux réponses à apporter aux questions "COMBIEN de niveaux et LESQUELS?", En effet, **nous ne disposons pas de réponse théorique, mais juste des indications obtenues lors de l'exposé des justifications à l'existence des niveaux, ou issues de nos expérimentations.**

QUELS niveaux? : Les niveaux IMAGE et SCENE sont incontestables puisqu'ils incluent respectivement *la donnée à analyser* et *le but recherché.* Les niveaux INDICES D'IMAGE et OBJET correspondent (leurs dénominations mises à part) à des **consensus** dans la communauté des chercheurs : nous n'en discuterons pas ici. Le niveau INDICES D'IMAGE est relatif à l'état de l'art des techniques utilisées pour extraire des *caractéristiques bidimensionnelles* dans une image. Le niveau OBJET répond plutôt au *lieu habituel de synthèse* des différents types de méthodes en Vision par Ordinateur, mais aussi au niveau de dialogue le plus couramment requis. A l'opposé de l'accord consensuel sur ces quatre niveaux, il reste que le niveau INDICES DE SCENE partage les avis : comme nous l'avons suggéré précédemment, son existence dépend du type des méthodes utilisées. Mais puisqu'il s'agit, *selon nous*, de regrouper des méthodes des quatre différents types (cf. début de la partie A), **le niveau INDICES DE SCENE est donc indispensable, à partir du moment où au moins une des méthodes utilisées est une méthode d'inférence de formes monoculaire ou multioculaire.** C'est cette notion que nous défendons par nos deux expérimentations, l'une fondée sur les contours, l'autre fondée sur la binocularité. Cependant, issue des expérimentations, notre définition du niveau INDICES DE SCENE reste très imprécise.

COMBIEN de niveaux? : la réponse à la question précédente, relative aux seules méthodes de Vision par Ordinateur, nous conduit à affirmer que, de manière générale, quatre niveaux exactement suffisent à passer du niveau IMAGE au niveau OBJET (phase d'Analyse d'Images) : ceux que nous distinguons. Par contre, pour ne pas avoir réellement étudié les niveaux OBJET et SCENE, nous ne pouvons pas affirmer que le passage de l'un à l'autre peut s'effectuer sans autre niveau intermédiaire.

Quelle que soit la validité de la structure hiérarchique que nous introduisons, il nous semble indispensable d'introduire et d'étudier d'autres méthodes de vision par ordinateur par rapport à cette structure. Seule une telle démarche permettra de mieux affiner, comprendre et de formaliser ce que sont exactement les niveaux de représentation, et tout particulièrement les plus élevés.

3 Représentation, Inférence et Contrôle

La distinction entre les cinq niveaux de représentation que nous venons d'expliciter, illustrés par la figure A.15, nous amène, indépendamment du nombre et de la nature de ces niveaux, à choisir des "modèles de représentation" et à distinguer trois types de procédures que nous appelons : "procédures d'enrichissement", "procédures d'inférence" et "procédures de contrôle". Nous définissons dans les paragraphes suivants ces différents types de procédures, tout en les illustrant par une partie du travail que nous avons effectué pour interpréter des indices d'image en indices de scène.

3.1 Les modèles de Représentation

Une représentation est un système formel qui permet de rendre explicites des entités ou types d'information. Comme nous venons de le voir, le "découpage" en cinq niveaux n'est pas arbitraire : chaque niveau correspond à une double spécificité, l'une sur le degré d'abstraction atteint, l'autre sur le degré de décentration atteint. L'existence même de ces deux caractéristiques conditionne la nature de la représentation associée à chaque niveau.

Pour chacun des niveaux que nous avons exhibés, les représentations choisies devront donc être capables de contenir explicitement ou implicitement toutes les informations présentes aux cinq niveaux prédéfinis. Les alternatives de représentation sont nombreuses, et nous nous contentons ici de renvoyer le lecteur aux différents articles, rapports ou bibliographies qui en font un tour d'horizon et une synthèse : [AGIN-76], [BALLA-82], [KANAD-77], [MARR-78] et [NISHI-81]. Notons simplement que deux mesures de qualité d'une représentation semblent très importantes du point de vue de la recherche d'objets qui sont connus du système (reconnaissance) : l'*unicité* et la *continuité*. La première de ces mesures signifie que la modélisation offre une représentation unique de chaque forme du monde physique; dans le cas contraire, se pose inévitablement le problème du choix parmi un large ensemble de représentations apparemment différentes et pourtant identiques. Le critère de continuité traduit le fait que des formes similaires doivent posséder des représentations similaires, et que des formes très différentes ont des représentations très différentes. Il est souhaitable, et tout particulièrement du point de vue de la reconnaissance d'objets, que les modèles de représentation choisis satisfassent à ces deux critères.

3.2 L'enrichissement interne du modèle

Chacun sait que toute représentation particulière extrait une information de manière explicite aux dépens d'autres informations qui sont repoussées plus loin et qui sont alors moins immédiates à récupérer. Ainsi la représentation d'un nombre entier en chiffres arabes rend-elle explicites les différentes valeurs des puissances de dix contenues dans ce nombre. C'est ainsi qu'une information, bien que pouvant être représentée de différentes manières, sera plus explicite dans telle représentation plutôt que dans telle autre.

Mais lorsqu'une multiplication doit être effectuée sur deux entiers, la représentation binaire ou arabe de ces entiers est plus appropriée qu'une représentation en chiffres romains. Cette remarque est bien évidemment applicable au domaine de la Vision par Ordinateur. En effet, le type de représentation présent à un niveau est aussi fortement conditionné par la nature des méthodes avec lesquelles les indices explicitement représentés à ce niveau ont été acquis. La conséquence en est que, sauf si on a beaucoup de chance, cette information n'est pas directement exploitable pour l'interprétation de ces indices en indices de niveau supérieur.

Le but des procédures intra-niveaux appelées "procédures d'enrichissement" est donc, à chaque niveau, sinon d'enrichir réellement la structure brute (résultat de l'interprétation des indices du niveau inférieur), du moins, en en modifiant sa représentation en fonction de son exploitation probable, d'en extraire les informations pertinentes pour l'interprétation des indices en indices du niveau supérieur. Créée de façon explicite à un niveau donné, l'information résultant d'un enrichissement sera alors toujours disponible en accès direct dès qu'elle est sollicitée. Ainsi, dans le cas de l'interprétation des indices d'image en indices de scène, la phase qui consiste à faire l'analyse locale d'un noeud-image par rapport à ses caractéristiques propres et à son environnement local en vue de détermination de son type, est une procédure d'enrichissement.

3.3 l'Inférence : les méthodes et les outils d'Interprétation

Le passage d'un niveau au niveau suivant ne tient pas seulement à la construction de nouveaux indices au sein du niveau inférieur qui puissent rendent possible les inférences effectives du niveau suivant par une simple association, mais surtout à ces "extensions" des structures du niveau courant qui dépassent les informations représentées, en augmentant soit leur degré d'abstraction, soit leur degré de décentration. Nous appelons de telles procédures des "procédures d'inférence". On retrouve déjà ce terme d'"inférence" dans l'expression "méthodes d'inférence de formes", car c'est effectivement de cela qu'il s'agit. Comme nous l'avons suggéré dans l'exposé des différentes méthodes utilisées en Vision par Ordinateur, chaque méthode d'inférence de forme peut être décomposée en inférences successives qu'il s'agit ensuite d'organiser entre elles. Le terme d'inférence est d'autant plus justifié, par rapport à l'usage qui en est habituellement fait en logique, que pour certaines de ces procédures élémentaires, il s'agit effectivement de procédures à base de "règles de réécriture" qui vont être appliquées pour découvrir des indices d'un niveau supérieur, plus abstraits et plus décentrés. C'est ainsi que fonctionnent les procédures qui nous permettent de passer du niveau INDICES D'IMAGE au niveau INDICES DE SCENE dans nos deux expérimentations : des algorithmes d'évaluation de discontinuités locales en surface et/ou en profondeur permettent, en fonction du type du noeud-image, de sélectionner et de construire des interprétations de ce noeud-image qui correspondent à des réalités scéniques du monde observé. Ces procédures, appelées "règles d'interprétation" apparaissent dans un formalisme proche de la logique des prédicats du premier ordre.

Dans tous les cas, les procédures d'inférence sont donc des algorithmes d'in-

terprétation inter-niveaux ascendants ("bottom-up") permettant, à partir d'informations contenues à un niveau donné, d'extraire des informations pour le niveau supérieur. Pour en revenir au problème du type d'un noeud-image, la détermination proprement dite du type (i.e. le fait de regrouper certains arrangements topologiques dans des classes prédéfinies en nombre donné) est déjà presque une inférence, mais elle n'arrive pas à se hisser au niveau INDICES DE SCENE : nous sommes dans ce cas en présence d'une "préinférence".

3.4 Le Contrôle : les méthodes et les outils de Contrôle

Les références traitant des problèmes de contrôle en Vision par Ordinateur sont nombreuses et le lecteur pourra se reporter aux descriptions et aux idées exposées par exemple par [FALK-72] [GARVE-76] [KANAD-77] [LUX-85a] [NAGAO-84] ou encore [SHIRA-75] ou [TSOTS-84] pour, en s'en donnant un simple aperçu, mieux prendre conscience de toute l'étendue du problème.

Il ne s'agit pas ici de discuter de ce qu'est la reconnaissance d'objets et de parler de la famille des systèmes qu'évoque cette interprétation de la notion de contrôle. Ces problèmes font partie de ce que nous appelons "le contrôle global" du processus de reconnaissance. Ce n'est pas le problème qui nous intéresse ici et, pour ne l'avoir entraperçu qu'à travers nos applications industrielles, nous n'en discuterons qu'à l'occasion de notre conclusion. Nous ne souhaitons pas non plus parler de contrôle au sens de l'enchaînement des instructions d'un programme, qui est aussi un tout autre problème.

Ce que nous entendons par "procédures de contrôle" correspond à un point très précis : il peut arriver que les abstractions ou les décentrations globales de niveau supérieur contribuent à diriger les abstractions de niveaux inférieurs. Nous en aurons une illustration toute particulière dans le cadre de notre seconde expérimentation. Mais même dans le simple cas de l'inférence des indices de scène par nos "règles d'interprétation", le maintien de la cohérence de l'interprétation globale de tous les indices image en indices de scène est réalisé par une procédure de contrôle qui n'a rien à voir avec un quelconque processus global de compréhension. La cohérence y est en effet assuré par un algorithme de propagation des contraintes imposées par les inférences locales en chaque noeud-image.

Les modèles de structures de contrôle actuellement existant se situent souvent à un niveau d'abstraction trop élevé. Nous définissons les "procédures de contrôle" comme des procédures inter-niveaux descendants, susceptibles d'être utilisées pour diriger le processus de compréhension de scène dans le sens où elles permettent d'*organiser l'enchaînement local des inférences élémentaires* et éventuellement de l'adapter par les prédictions qu'elles sont capables d'émettre, indépendamment de tout problème de reconnaissance d'objet.

La figure A.16 présente des exemples de ces trois types de procédures qu'il est possible et souhaitable de mettre en oeuvre au sein d'un système présentant différents niveaux de représentation de la connaissance. Au cours des chapitres suivants, nous décrivons, outre les niveaux de représentation utilisés, quelles sont

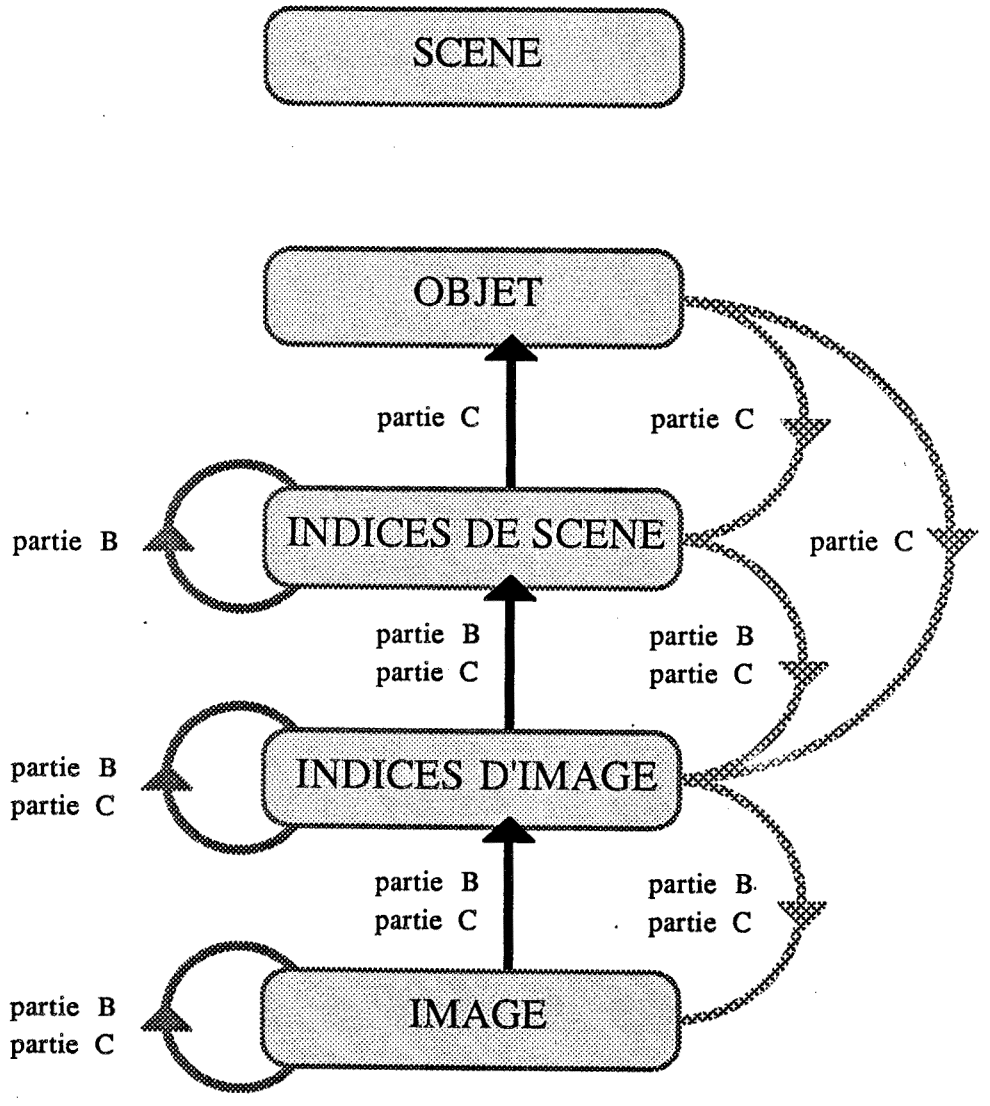


Figure A.16: Exemples de procédures intra- et inter- Niveaux expérimentées

les différentes procédures d'enrichissement, d'inférence, et de contrôle qui ont contribué à mener à bien deux expérimentations. Chacune d'entre elles étudie une caractéristique spécifique du système visuel humain, et chacune a été étudiée dans un domaine différent : l'inférence de formes à partir de contours pour une image monoculaire noir et blanc (partie B) et l'inférence de formes à partir de la stéréovision simple pour des paires stéréoscopiques couleur (partie C).

Partie B

Inférence de Formes à partir des Contours (images noir et blanc - objets solides)

La perception d'objets dans l'espace est un processus que l'on peut définir à partir des propriétés des transformations à trois dimensions et des lois de la nature. En analysant soigneusement ces propriétés, nous avons développé une procédure qui non seulement identifie les objets, mais détermine également leur orientation et leur position dans l'espace.

Larry G. Roberts (1965)

Cette seconde partie décrit une première expérimentation des niveaux proposés. Elle a trait à l'inférence de formes à partir de contours, dans le domaine restreint d'objets solides du "monde des blocs". Traitant d'images noir-et-blanc, elle est tout particulièrement destinée à mettre l'accent sur la distinction entre deux niveaux particuliers : le niveau INDICES D'IMAGE et le niveau INDICES DE SCENE. Suivant notre axe méthodologique, elle montre en outre ce qu'apporte la distinction entre ces niveaux dans le cas concret d'une application industrielle concernant la localisation de paquets sur une palette.

Chapitre B.I

Monde des Blocs et Niveaux de Représentation

1 Le Monde des Blocs et le Monde d'Origami

Dès ses débuts, le domaine privilégié d'expérimentation de la Vision par Ordinateur a été le "monde des blocs". L'étude de la vision dans ce monde restreint, constitué d'objets polyédriques (incluant donc des parallélépipèdes, des pyramides), correspond à une double volonté. La première est celle de découvrir certains des principes de base de la vision (essentiellement dans le sens de la reconnaissance) en se restreignant à un monde bien maîtrisé. Bien que restreint aux objets polyédriques, ce domaine reflète malgré tout une grande partie du monde réel. La seconde est celle de pouvoir rapidement se connecter aux domaines où le "monde des blocs" est aussi un domaine d'expérimentation privilégié. Cette observation est tout à fait notable dans le domaine de la robotique et de la productique, où une collaboration entre un système de vision et un robot passe souvent par la connaissance commune d'un même modèle. Les autres avantages à s'intéresser à ce domaine d'objets solides composés de surfaces planes sont d'ordre informatique. Ils se résument en une seule phrase : une définition du monde des objets réels claire et précise, qui d'un point de vue des modèles de représentation permet de se cantonner à une représentation par sommets et arêtes, et d'un point de vue de l'interprétation des images autorise à se dispenser de vagues heuristiques.

C'est la restriction de ce monde aux seuls objets solides triédriques qui est étudiée par [HUFFM-71], [CLOWE-71], mais aussi par [SHIRA-75] et [WALTZ-75]. Tous ces travaux ont fourni de très nombreux résultats. Ils ont en particulier permis de concentrer de nombreuses caractéristiques (connexité, continuité et convexité par exemple) au sein d'un même étiquetage. Pour schématiser ce qui s'est passé par la suite, on peut dire que chaque chercheur a tenté de construire son propre dictionnaire d'étiquetages possibles des noeuds et des droites, éléments constitutifs des dessins de lignes étudiés, afin d'augmenter le domaine des objets qui pouvaient être interprétés par cette unique technique. A son maximum, un "dictionnaire de jonctions" parfait doit être capable d'étiqueter les objets réels et d'interdire l 'étiquetage d'objets impossibles.

C'est en particulier ce que fait Kanade en introduisant le monde d'Origami. La différence entre le monde des blocs et le monde d'Origami réside dans le fait que le premier permet de décrire des objets volumiques alors que le second, lui, s'intéresse aux objets surfaciques [KANAD-80a]. Les objets décrits par le monde d'Origami sont donc des objets sans épaisseur, une boite évidée de son intérieur par exemple : de façon générale, ce sont ceux que l'on peut fabriquer à l'aide de papier plié, comme le fait la tradition japonaise. Dans les deux mondes pourtant, la restriction très importante reste la limitation à des surfaces planes. Et si la complexité de l 'étude de ces deux domaines est identique (par exemple au niveau des procédures d'étiquetage même ou du contrôle de la cohérence globale des résultats obtenus localement), le dictionnaire de jonctions nécessaire à l'étude du monde d'Origami est nettement plus riche.

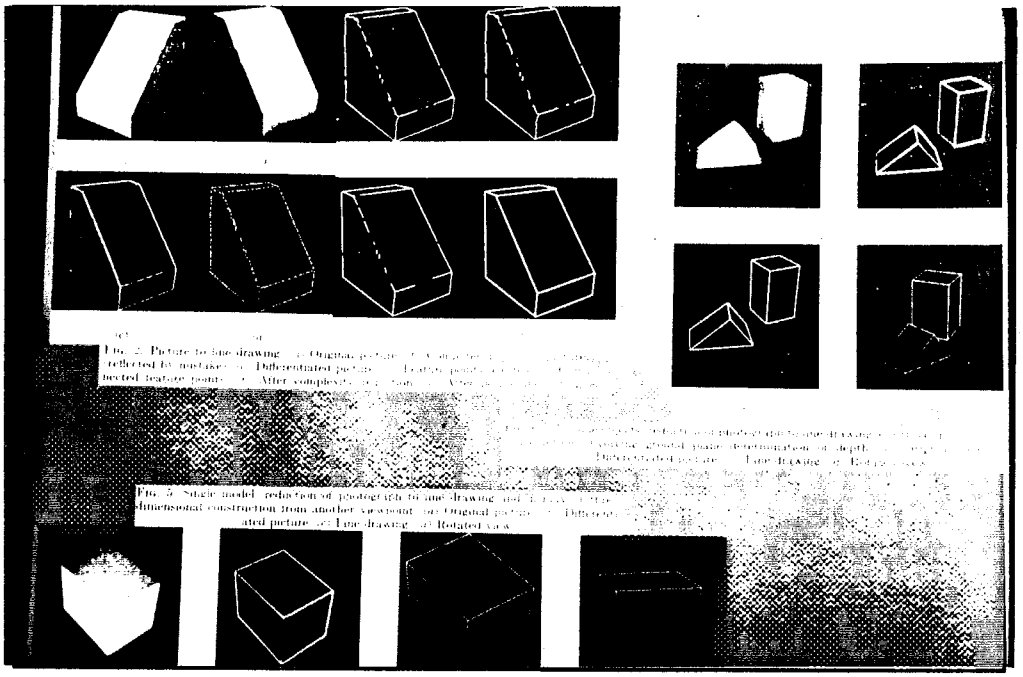
2 Les approches et les Niveaux de Représentation

2.1 L'approche géométrique numérique

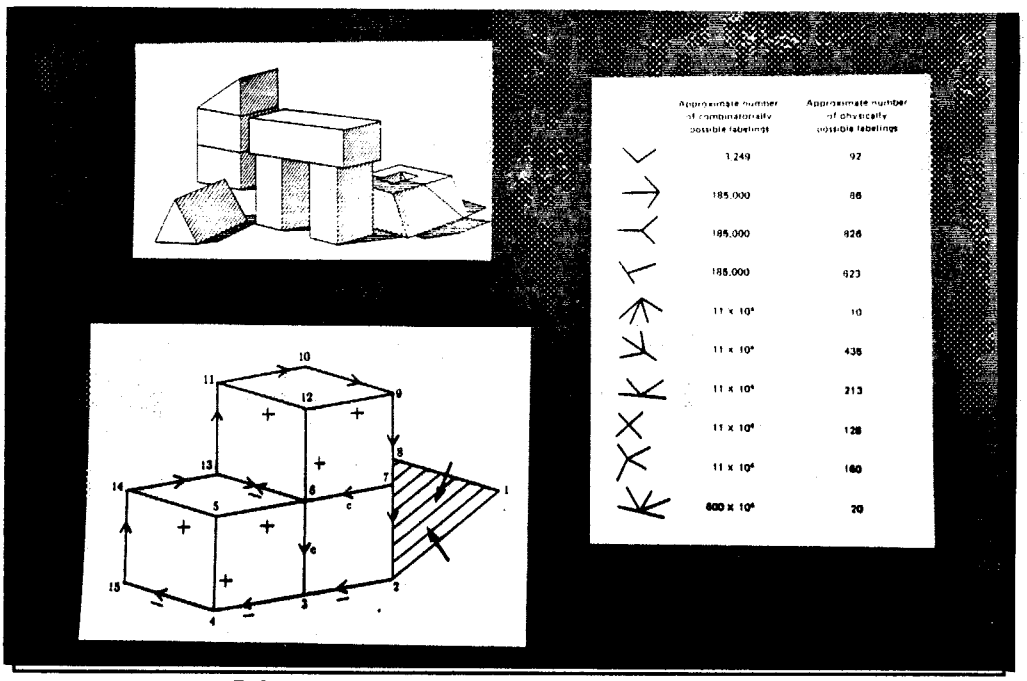
Les premiers travaux effectués en Vision par Ordinateur tridimensionnelle et dans le domaine de la reconnaissance d'objets sont ceux de Roberts. Le trait caractéristique de l'approche de Roberts [ROBER-65] est d'être la prépondérance du numérique, non seulement pour l'extraction des indices de l'image, mais aussi pour le processus même de mise en correspondance.

Dans un premier temps, le système analyse tous les pixels de l'image pour finalement ne retenir que ceux (points de contraste) qui sont susceptibles d'appartenir à des arêtes de l'objet en scène. Ces pixels sont alors regroupés en vue d'obtenir un dessin de traits représentatif de la scène observée (segmentation, cf. figure B.01.a). Ce dessin est constitué de droites obtenues par approximation aux moindres carrés des lieux des points de contraste détectés.

Dans un second temps, ce dessin de traits est mis en correspondance avec des modèles 3-D pour fournir une description quantitative de la scène. Les modèles sont choisis sur des critères simples tels que le nombre de sommets découverts. Après application d'une rotation, d'un changement d'échelle et d'une projection, chaque



B.01.a: *(d'après [ROBER-65-])* Le système de Roberts
approche géométrique numérique



B.01.b: *(d'après [WALTZ-75-])* Le système de Waltz
approche géométrique symbolique

Figure B.01: Les trois grandes approches du monde des blocs (1)

modèle sélectionné est alors comparé au dessin de traits précédemment extrait de l'image.

Finalement, il est possible de visualiser le modèle de la scène reconnue par le système.

Le programme de Roberts progresse séquentiellement du niveau IMAGE au niveau INDICES D'IMAGE, et s'arrête à ce niveau d'interprétation pour, directement et numériquement, mettre en correspondance la structure observée avec les modèles sélectionnés.

L'efficacité d'une telle approche est très fortement conditionnée par la qualité des processus d'extraction des INDICES D'IMAGE. Pratiquement, il est très difficile d'assurer une bonne segmentation. Par exemple, l'utilisation de certains critères fait que certaines droites de faible contraste peuvent être oubliées. au contraire, d'autres droites traduisant par exemple certaines particularités de la texture de l'objet étudié sont extraites.

2.2 L'approche géométrique symbolique

Nous parlerons essentiellement de deux systèmes typiques de l'approche géométrique symbolique : ceux de Waltz [WALTZ-75] et de Shirai [SHIRA-75]. Il ne faut pas cependant perdre de vue que cette approche continue à être étudiée et reste la plus largement répandue, comme en témoignent par exemple les travaux rapportés dans [GUZMA-68] [HUFFM-71] [FALK-72] [MACKW-73] [SUGIH-79] [BARRO-81] ou encore [KANAD-81].

2.2.1 Le système de Waltz

La première partie de l'étude de l'image conduit à construire un réseau de lignes-image correspondant aux lignes de contraste contenues dans cette image. Un étiquetage permet de préciser la nature de ces droites-image suivant qu'elles sont convexes ou concaves, correspondent à des arêtes de contour (arêtes d'ombre ou de contour d'objet) ou non (cf. figure B.01.b). Dans tous les cas, les droites-image et les interprétations qui en sont faites sont confondues dans le système. Les niveaux jusqu'ici étudiés sont donc les niveaux IMAGE et INDICES D'IMAGE. En effet, un étiquetage d'indices d'image n'est qu'une procédure d'enrichissement du niveau INDICES D'IMAGE. Le système de Waltz s'arrête d'ailleurs au niveau INDICES D'IMAGE, aucune description en volumes n'étant réellement effectuée.

L'interprétation de l'image est purement ascendante jusqu'à la phase de reconnaissance elle-même (mise en correspondance entre les arêtes détectées et leurs propriétés, avec une base de modèles décrits de façon identique).

2.2.2 Le système de Shirai

Les niveaux présents dans le système de Shirai sont les mêmes que ceux présents dans le programme de Waltz. La différence essentielle réside dans le fait que la base de modèles des objets connus par le système intervient perpétuellement dans

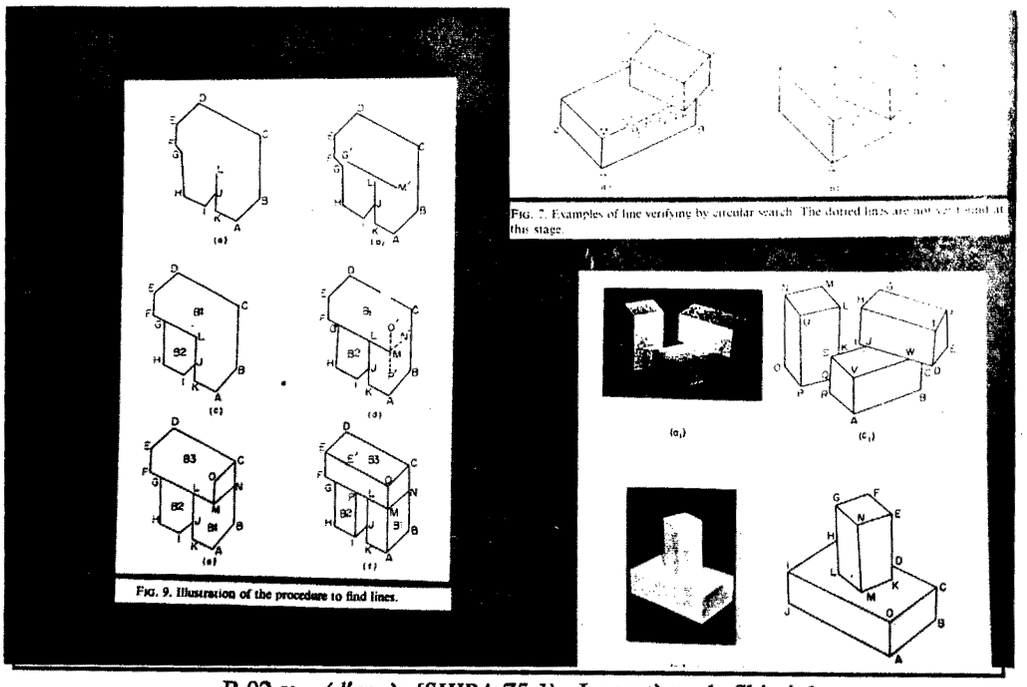
l'extraction et dans l'interprétation de l'information contenue dans l'image (processus descendant). Non seulement elle permet de contrôler les processus d'étiquetage du niveau INDICES D'IMAGE, mais elle intervient aussi dans les extractions d'indices d'image : prédiction de noeud-image par exemple, raffinement d'extraction des droites-image issues d'un même noeud-image. Les procédures de contrôle sont donc plus raffinées que chez Waltz.

Tout comme dans le système de Waltz, les structures des indices d'image et des interprétations auxquelles ils donnent lieu (indices d'image augmentés d'étiquetage de scène) sont confondues (cf. figure B.02.a). Ce dernier choix interdit toute possibilité d'extension à des domaines plus proches du monde réel.

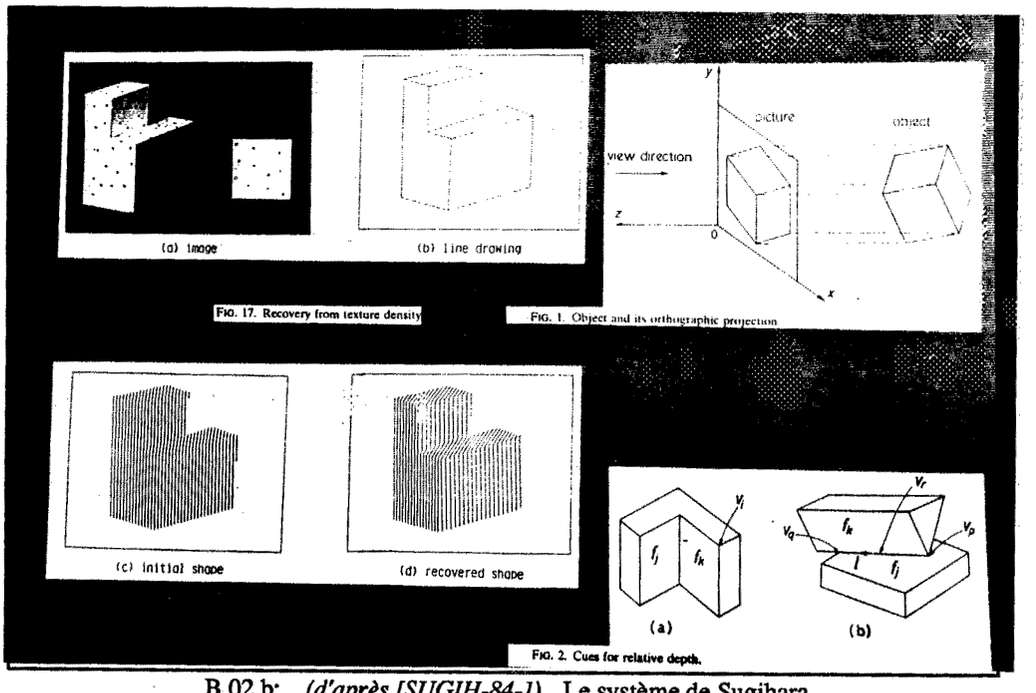
2.3 Les autres approches

Nous avons explicité les différentes approches géométriques car elles sont les plus connexes à notre démarche. Il faut toutefois noter que ces approches ne sont pas les seules et coexistent avec bien d'autres approches, dont par exemple les méthodes syntaxiques de reconnaissance des forme ou encore les approches algébriques.

Dans ces dernières, le problème de l'inférence de formes est formulé comme un problème d'optimisation de contraintes, et la fonction à minimiser peut par exemple être une somme pondérée de mesures d'erreur sur des informations telles que la réflectance ou la texture. La restriction à des objets polyédriques permet en outre de traduire des contraintes algébriques linéaires, là où elles sont en général traduites par l'intermédiaire de l'espace gradient [SUGIH-84] (cf. figure B.02.b).



B.02.a: *(d'après [SHIRA-75-])* Le système de Shirai
approche géométrique symbolique



B.02.b: *(d'après [SUGIH-84-])* Le système de Sugihara
approche algébrique

Figure B.02: Les trois grandes approches du monde des blocs (2)

Chapitre B.II

Distinction entre Indices d'Image et Indices de Scène

Les objets du monde des blocs, facilement descriptibles et engendrant des algorithmes simples à mettre en oeuvre continuent à intéresser bien des personnes, même s'ils sont reconnus comme non satisfaisants. Pourtant, dans la précipitation à obtenir des résultats, les méthodes développées restent souvent très fortement dépendantes de ce monde restreint, et offrent rarement des ramifications vers des domaines plus réels. Cet inconvénient majeur est issu, de notre point de vue, de la confusion qui s'est installée entre les niveaux INDICES D'IMAGE et INDICES DE SCENE : en effet, les objets de ce monde paraissent au prime abord bien trop simples pour que leur interprétation ou leur reconnaissance nécessite un dépassement du niveau INDICES D'IMAGE. A l'inverse, nous pensons que la compréhension de scènes visuelles nécessite une représentation de l'information à différents niveaux pour les études de scènes du monde réel mais aussi pour celles du monde des polyèdres. Expliquer et expliciter cette distinction, c'est ce que nous avons voulu illustrer dans ce court chapitre par différents exemples, sur le cas particulier de la différence entre indices d'image et indices de scènes. Par ailleurs, l'originalité de cette présente partie ne concerne certes pas le fait d'avoir travaillé dans le monde des blocs, mais réside dans la distinction entre les deux niveaux INDICES D'IMAGE et INDICES DE SCENE, et surtout dans la différenciation entre les types de structures créées à ces niveaux, qui reflète des notions différentes.

1 Les limitations de l'étiquetage

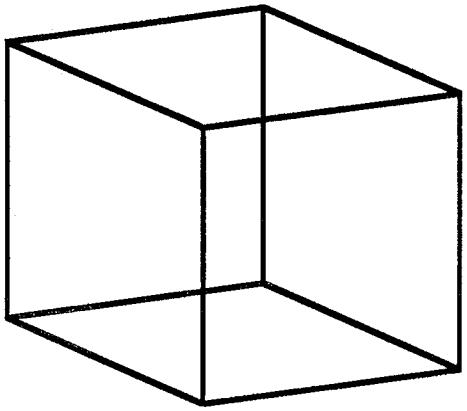
Selon nous, confondre et assimiler les deux structures constitue un choix a priori qui interdit toute possibilité d'extension à des domaines plus proches du monde réel. Et ce qui est plus important, cela interdit aussi tout contrôle du processus d'interprétation des indices d'image en indices de scènes. Ceci est tout particulièrement visible si nous nous retournons vers le processus de l'étiquetage de scènes du monde des blocs, tel qu'il est défini par les travaux de la famille de ceux de [WALTZ-75], ou tel qu'il est très souvent pratiqué par [ROSEN-76].

Rappelons-nous que, d'après nos définitions, seule une évolution en abstraction et/ou en décentration permet un changement de niveau de représentation, et que cette évolution s'accompagne nécessairement soit d'un changement de type de représentation, soit d'un changement de type de référentiel (ce qui n'est pas envisageable pour un simple étiquetage). Or il est remarquable de noter qu'un étiquetage s'effectue par simple adjonction d'un attribut sur des éléments de la structure de niveau INDICES D'IMAGE. C'est l'aveu du fait même qu'un étiquetage d'indices d'image est une procédure d'enrichissement du niveau INDICES D'IMAGE : cela ne peut donc pas nous suffire à atteindre le niveau INDICES DE SCENE, et cela ne nous permet pas de contrôler localement l'étiquetage lui-même, par rapport à des observations de niveau INDICES DE SCENE par exemple (cf. figure A.16).

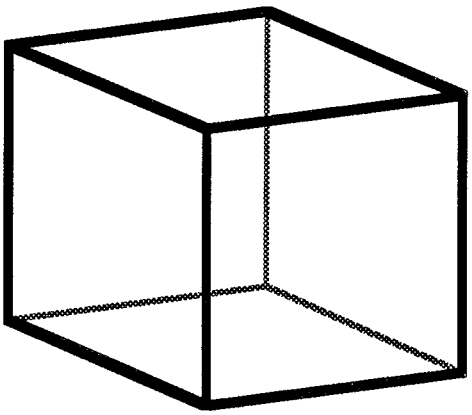
Le défaut qui traduit un manque de contrôle possible en est d'ailleurs accompagné d'un autre, tout aussi grave : en général, les processus d'étiquetage ne fournissent qu'une unique solution au problème de l'interprétation d'une structure d'indices d'image, même si parfois, les autres solutions restent implicites. Cela semble tout naturel, mais est dû en particulier au fait qu'agissant sur la même structure, ils sont donc incapables d'évaluer à nouveau une interprétation locale d'un indice d'image particulier sans annuler d'abord la totalité des interprétations en indices de scène de tous les autres indices d'image, sauf à conserver une incohérence au sein de la double représentation indices d'image / indices de scènes.

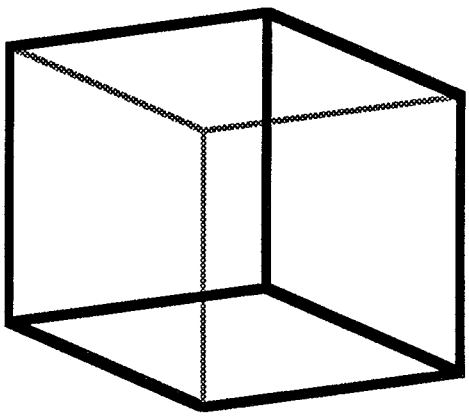
2 Le besoin d'interprétation multiple

Pour surmonter les difficultés que nous venons d'exhiber dans le cas de l'étiquetage, étudions un moment le problème de la reconnaissance du cube transparent illustré par la figure B.03.a, et plus connu sous le nom de "cube de Necker". Ce cube est un outil cher aux psychologues car il montre que l'être humain perçoit alternativement deux formes tridimensionnelles différentes à partir d'une même image bidimensionnelle (cf. figure B.03.b), sans arriver à décider quelle est la bonne interprétation. Et effectivement, cela lui est d'autant plus difficile que les *deux interprétations sont physiquement possibles.* Le cube est décrit sans ambiguité au niveau INDICES D'IMAGE par les segments de droite qui le composent (au nombre de 16 si on admet qu'un noeud-image de type X est la jonction de quatre droites-image) Par opposition, il ne peut être décrit en volumes qu'à partir du moment où une interprétation



B.03.a: Le cube de Necker





B.03.b: Les deux interprétations en indices de scène physiquement possibles

Figure B.03: Le cube de Necker et ses deux interprétations

des noeuds-image de type X a été donnée. Cela revient à dire qu'avant toute interprétation globale de la structure, il faut décider, lorsque deux droites-image se chevauchent, quelle est celle qui se trouve "en avant" de l'autre. Si ce choix n'est pas effectué, deux descriptions en volumes sont réellement possibles : nous pensons qu'il n'y a aucune raison d'en privilégier une par rapport à l'autre.

L'exemple du cube de Necker prouve la nécessité de distinction entre indices d'images et indices de scène et montre qu'il est indispensable que le système de vision, s'il ne possède comme information qu'un réseau d'indices d'image initial comme celui de la figure B.03.a, soit à même de fournir les deux descriptions physiquement possibles.

3 Discussion

L'exemple du cube de Necker est une image où la scène observée est naturellement ambigue : le cube est associé à plusieurs interprétations également valides. C'est aussi le cas des images à sens multiples comme celles de Verbeek (cf. figure A.11).

Mais la cause principale de l'ambiguité d'interprétation d'une image est à chercher dans le fait qu'une image monoculaire, tout comme une image rétinienne pour le système visuel humain, est une illustration *partielle* (seul un "demi-espace" est visible par le capteur) et *bidimensionnelle* (l'image reçue par le capteur est le résultat d'une projection) d'une scène tridimensionnelle. Pour ces raisons ("directionalité " de l'observation de la scène et "projection" de la scène observée), une même image bidimensionnelle peut donc avoir pour origine deux scènes réelles physiques différentes. En général, le système visuel humain interprète l'image d'une seule façon, qui exprime la perception la plus conventionnelle, en fonction d'un contrôle ponctuel qui lui permet, d'après les connaissances de "haut niveau" qu'il peut avoir de la stabilité du monde et des objets, de guider et de préciser les interprétations. C'est d'ailleurs cette réflexion de base qui constitue le fil conducteur des travaux de l'école dont font partie [BARNA-84] [HORAU-85] et [LOWE-85]. Cependant, si le système visuel humain reste "aveugle" à toutes les autres possibilités, cela ne doit pas nous conduire à penser que ces dernières n'existent pas dans la réalité. De la même façon, il n'est pas question d'éliminer de telles possibilités dans le cas d'un système de vision informatique qui peut être l'observateur de telles ambiguités. Comme dans le cas du cube de Necker, le système de vision doit pouvoir fournir les différentes descriptions possibles, quitte à ce que, pour des traitements ultérieurs, il ordonne ces possibilités en fonction d'un certain degré de confiance, ou même qu'il choisisse l'interprétation qui lui convienne le mieux au moment donné.

Cette multiplicité d'interprétations liée à la notion de projection directionnelle d'une scène tridimensionnelle est typique de l'exemple de scène du monde des blocs, illustré par la figure B.04, et que nous expliquons maintenant. Cette scène possède non pas deux, mais au moins trois interprétations possibles qui, dirons-nous, correspondent à des "réalités scéniques". La première décrit un cube qui est "posé sur" une surface de mesure $6S$ si la surface du côté du carré vaut S. La seconde interprétation correspond à deux objets qui sont parfaitement em-

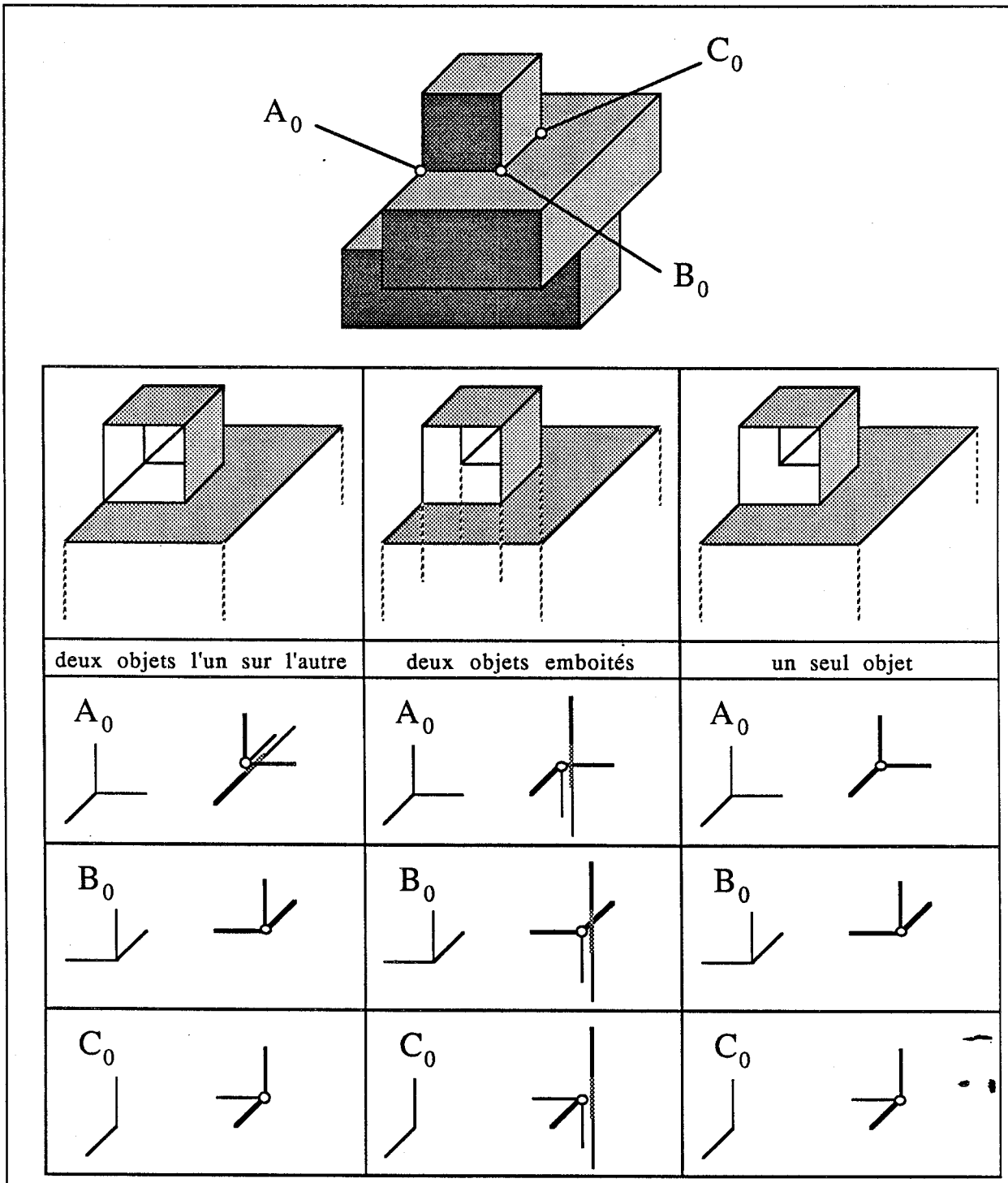


Figure B.04: Différentes interprétations d'une scène du monde des blocs

boités l'un dans l'autre, l'un présentant une surface de valeur $5S$. La troisième correspond à l'interprétation d'un seul objet taillé dans la masse, composé d'au moins sept cubes de surface de côté S. Les trois interprétations sont possibles, et, sans connaissance particulière supplémentaire, aucune décision ne peut être prise (dans un contexte plus général, encore d'autres interprétations sont d'ailleurs possibles). Si par contre, en restreignant le contexte à des objets composé de cinq cubes élémentaires de surface de coté T arrangés de façon planaire ("pentacubes "), on en déduit immédiatement que seule la seconde interprétation est une "réalité scénique contextuelle", et que la valeur S observée correspond à la valeur T réelle. Notons que l'interprétation locale d'un noeud-image peut être identique d'une interprétation globale à l'autre. C'est par exemple le cas pour le noeud-image B_0 dans les premières et troisième interprétations globales. D'autre part, la détermination précise d'une interprétation d'un indice-image donné (resp. du noeud-image "A_0") conditionne les interprétations possibles des noeuds-image immédiatement voisins (resp. du noeud-image "B_0"). Et il est clair que les nouvelles contraintes déduites pour ces noeuds voisins peuvent être propagées récursivement sur leurs propres voisins.

En réponse à ces différences causes d'ambiguité, la possibilité d'engendrer plusieurs interprétations d'une même structure d'indices d'image (ou de toute structure d'un niveau quelconque) constitue l'une des caractéristiques de notre approche. Elle prémunit le système visuel contre toute éventualité de non-reconnaissance au cas où reconnaissance il doit y avoir.

Chapitre B.III

Compréhension d'une Scène du Monde des Blocs

La première des expérimentations que nous avons menées a été effectuée dans le monde des blocs. Il s'agit d'analyser une image noir-et-blanc par inférence de formes à partir de contours pour découvrir quelles sont les formes polyédriques contenues dans cette image. C'est en suivant cet objectif final que nous avons été amenés à travailler directement sur les niveaux IMAGE, INDICES D'IMAGE et INDICES DE SCENE (cf. figure B.05).

D'un point de vue de l'implantation, seuls le niveau IMAGE, les procédures d'enrichissement de ce niveau et certains outils mathématiques nécessaires à l'interprétation en indices d'image sont écrits en FORTRAN sur LSI 11/23 (Système RT11B). Ils ont été réalisés par Augustin Lux au sein du système CAIMAN. Décrits avec beaucoup de précision dans [LUX-83a], ces algorithmes sont donc à la base de la construction du niveau INDICES D'IMAGE : nous y ferons fréquemment allusion en omettant toutefois de les décrire à nouveau.

Les niveaux INDICES D'IMAGE et INDICES DE SCENES ainsi que toutes les autres procédures traitant de ces niveaux ont été écrites en MACLISP sur HB70 (Système Multics).

Le chapitre expose, du niveau IMAGE au niveau INDICES DE SCENE, les outils qui ont été développés pour cette première expérimentation. Nous y présentons une

définition formelle de certains indices d'image et indices de scène. Nous y montrons aussi de façon détaillée les algorithmes d'inférence, de contrôle et d'enrichissement qui touchent à ces niveaux, ainsi que la modélisation sous-jacente utilisée. Une partie de ces travaux est rapportée dans [DEMAZ-82] et [DEMAZ-84a].

1 Interprétation en Indices d'Image

Le système dispose d'indices de contraste (de niveau IMAGE), dont la caractéristique est d'avoir un gradient d'intensité lumineuse supérieur à une certaine valeur de seuil. L'interprétation de ces nuages de points de contraste, se traduit par la création de deux types d'indices d'image, souvent évoqués dans les premiers chapitres du rapport sans jamais avoir été précisément définis : les noeuds-image et les droites-image. Cette phase d'interprétation en indices d'image est aussi appelée "extraction d'indices d'image" ou "segmentation".

1.1 Quelques définitions

Un ensemble de points de contraste *voisins* (au sens du 8-voisinage pour nos expérimentations) et *alignés*, peut-être interprété comme un segment de droite d'un espace à deux dimensions. Il peut donc donner lieu à un calcul d'approximation, et ainsi fournir un indice D composé des extrémités "N_1" et "N_2" de la meilleure droite d'approximation ainsi obtenue.

Les coordonnées des extrémités ne doivent pas être exprimées dans le repère discret de l'image. En effet, elles sont réelles et ne doivent pas être, à ce moment de l'analyse, approchées par des valeurs entières. La structure créée à ce niveau sera donc exprimée dans un système de coordonnées à valeurs réelles, qui de plus est normé. Ce n'est pas le cas du repère du niveau IMAGE qui, lui, est totalement dépendant du matériel : en effet, les dimensions des digitaliseurs de nombreux systèmes ne sont pas normées, et par exemple, la finesse de la discrétisation verticale est plus importante que la discrétisation horizontale.

Notons finalement qu'au caractère incrémental [MERO-75] du logiciel CAIMAN correspond une approche incrémentale de la construction des indices d'image.

Cette introduction nous amène à définir les deux indices d'image utiles pour notre première expérimentation :

droite-image : tout indice image DI, meilleure approximation d'une ou de plusieurs droite(s) de contraste, et maximale : aucun point de contraste découvert ne peut venir allonger la droite-image DI sans croiser une autre droite-image, de sorte qu'aucun point (y compris les extrémités) de tout autre droite-image extraite ne soit intérieure à DI.

noeud-image : toute extrémité NI_1 ou NI_2 d'une ou plusieurs droite(s)-image. Toute droite-image ne contient que deux noeuds-image : ses extrémités.

noeud-image pendant : tout noeud-image extrémité d'une seule droite-image.

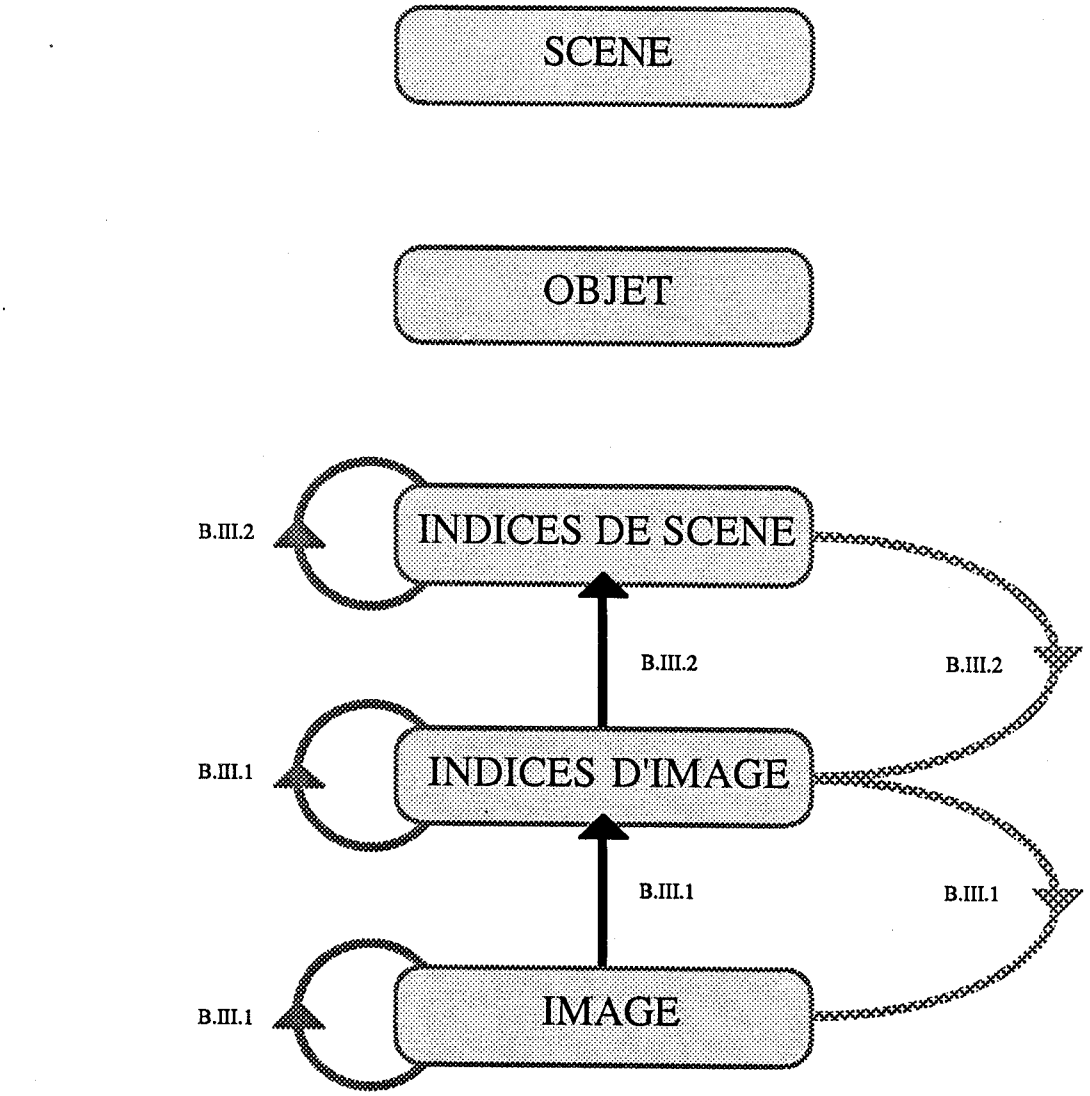


Figure B.05: Utilisation des Niveaux pour notre première expérimentation

droite-image pendante : toute droite-image dont une au moins des extrémités est un noeud-image pendant.

Par rapport à ces définitions, il est clair que les ensembles de noeuds-image et de droites-image évoluent au cours du traitement, selon le degré d'interprétation effectué sur l'image de contraste. C'est-à-dire que la structure d'indices d'image construite à ce niveau est non seulement disponible après une segmentation jugée satisfaisante, mais que sa composition est susceptible de varier dès qu'une demande d'extraction d'un indice d'image supplémentaire est demandée : c'est le contrôle global du processus de compréhension de scène (celui qui assure l'enchaînement des procédures d'inférence, de contrôle et d'enrichissement) qui en décide.

1.2 Procédures d'Inférence et de Contrôle

1.2.1 Algorithme d'Interpretation en Noeud-Image

L'algorithme interprète un point N de l'image en un noeud-image NI. Cette interprétation se fait en trois phases, qui contiennent toutes un part de gestion d'incertitudes, que nous n'évoquons pas ici pour plus de clarté dans l'exposé :

1. Recherche de départs : par appel de la commande AN de CAIMAN en un point de contraste choisi N, on obtient un ensemble de points de contraste $\{C_{iN}\}$, débuts potentiels de droites de contraste (enrichissement de niveau IMAGE) (cf. figure B.06.a)

2. Validation de départs : pour chaque point C_{i_N} découvert, on demande une poursuite de droite de contraste vers l'extérieur de la fenêtre (commande AA) sur un nombre minimal de points. Si cette poursuite aboutit à un succès, le point C_{i_N} est validé en départ d'une droite de contraste DC_{i_N} interprétable en droite-image. Dès qu'un tel point est validé, on sait que le point N est susceptible d'être interprété en un noeud-image NI, s'il ne l'était pas déjà (préinférence vers le niveau INDICES D'IMAGE). Sur la figure B.06.b, seul C_{3_N} n'est pas validé.

3. Interprétation des DC_{i_N} en droites-image $DI_{i_{NI}}$ et du point N en noeud-image NI. Il s'agit, pour chaque point C_{i_N} validé, d'examiner si la droite de contraste DC_{i_N} associée passe par le point N de l'image. Si la droite répond à ce critère, elle est alors interprétée en un "début de droite-image" $DI_{i_{NI}}$ ayant pour extrémités les noeuds-image NI et $CI_{i_{NI}}$, interprétations au niveau INDICES D'IMAGE des points de contraste N et C_{i_N} qui en sont alors déduites. Les $CI_{i_{NI}}$, destinés à évoluer et être précisés ultérieurement, nous les qualifions de "temporaires". La figure B.06.c montre que seules les droites-image $DI_{1_{NI}}$, $DI_{2_{NI}}$ et $DI_{6_{NI}}$ sont créées (inférence vers le niveau INDICES D'IMAGE). Notons que c'est l'inférence d'au moins une droite-image dont l'indice de contraste correspondant passe par le point N qui permet d'interpréter le point N en noeud-image NI (contrôle sur le niveau IMAGE).

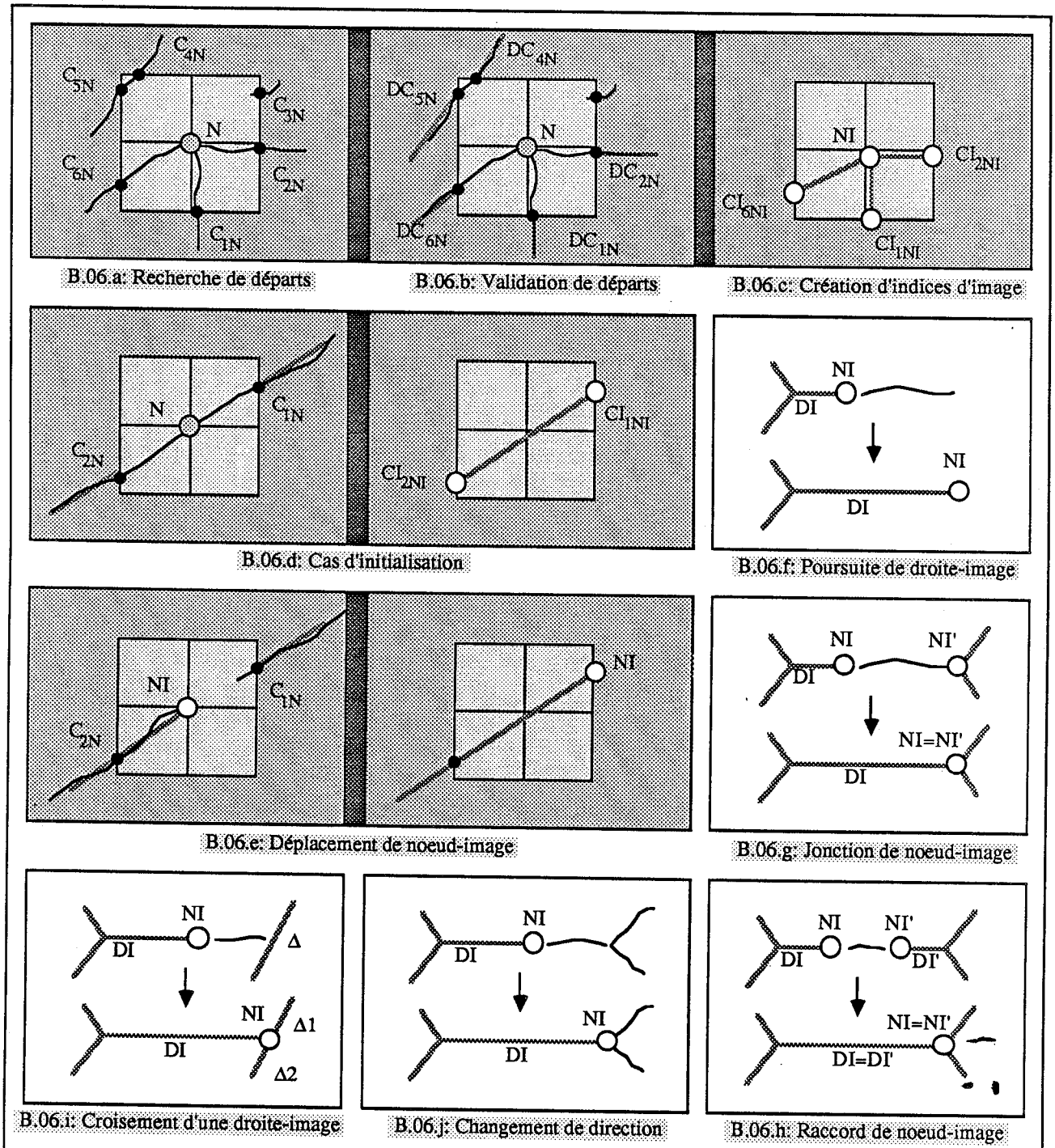


Figure B.06: Extraction incrémentale des indices d'image: fonctions de base

Si aucune des droites de contraste ne peut être interprétée en droite-image, N n'est pas interprété en noeud-image.

Remarque 1 : Le point N peut (suite à une application antérieure de l'algorithme en ce point) déjà avoir été interprété en un noeud-image NI, extrémité d'une ou plusieurs droites-image antérieurement construites. Il est alors nécessaire dans la phase 3/ de l'algorithme, d'empêcher l'interprétation de droites de contraste issues des phases 1/ et 2/ qui ont déjà été interprétées lors d'une analyse précédente. Ainsi modifié, l'algorithme peut alors servir aux deux fins suivantes : essayer d'interpréter un point de l'image en un noeud-image, ou compléter un noeud-image déjà existant. (C'est le contrôle global du processus de compréhension de scènes qui décide du choix de l'utilisation de la procédure, et non une quelconque procédure de contrôle sur le niveau IMAGE telle que nous l'avons définie au cours de la partie A). Dans le second cas, un moyen d'obtenir de nouvelles droites de contraste peut être a / soit de modifier les paramêtres qui définissent le fonctionnement de CAIMAN (par exemple modification du seuil du gradient au-dessus duquel on considère que le point de l'image est un point de contraste), b / soit d'utiliser un algorithme de suivi de droite de contraste différent de celui de CAIMAN, c / soit encore d'utiliser une image intrinsèque du niveau IMAGE différente de l'image de gradient d'intensité lumineuse.

Remarque 2 : Dans le cas particulier où à la fin de la phase 2/, seules deux droites de contraste ont été validées, passant par N en étant alignées (on dit alors que ces droites-image sont "opposées"), ces deux lignes de contraste ne sont interprétées qu'en une seule droite-image. C'est le seul cas où, alors qu'il y a création de droite-image autour de N, il n'est pas possible d'interpréter le point de N en noeud-image (cf. figure B.06.d). Ce cas peut par exemple se produire lors d'une initialisation du système, pour la recherche d'un premier noeud-image. L'initialisation est donc une autre possibilité d'utilisation de l'algorithme. Elle nous en introduit d'ailleurs une dernière.

Remarque 3 : Si le point de l'image N a déjà été interprété en un noeud-image pendant NI (extrémité de DI), et que l'algorithme ne découvre que deux lignes de contraste opposées, passant par N, dont l'une est celle qui a été interprétée avec succès, il les interprète en la même droite-image DI, en ne construisant aucun nouveau noeud-image : il déplace simplement le point de l'image dont NI est l'interprétation, de la valeur N à celle du point de contraste extrémité de la nouvelle droite de contraste détectée (cf. figure B.06.e). Cette particularité traduit le fait qu'une droite-image ne correspond pas obligatoirement à une seule droite de contraste.

Dans tous les cas de figures évoqués dans cet algorithme, les droites-image qui ont été créées sont susceptibles d'être prolongées par application du deuxième algorithme que nous présentons maintenant.

1.2.2 Algorithme de suivi de Droite-Image

Les "débuts de droites-image" $DI_{i_{NI}}$ créées au cours du premier algorithme sont le résultat d'une préinférence, mais il reste que l'une au moins de leurs extrémités (correspondant au noeud-image pendant $CI_{i_{NI}}$) est directement liée au point de contraste correspondant dans l'image. Il s'agit ici d'"étendre" les droites-image, en "tirant" par ce point $CI_{i_{NI}}$ les lignes de contraste dont chaque droite-image est l'interprétation. Cette opération peut, là encore, se faire dans un double but : soit raccorder cette nouvelle droite-image à un noeud-image ou à une droite-image existants, soit découvrir un nouveau noeud-image. L'algorithme se déroule en deux phases :

1. Activation du prolongement de droite : on appelle la commande AA de CAIMAN pour poursuivre l'extraction de la ligne de contraste D associée à la droite-image DI, à partir de son extrémité pendante N associée au noeud-image NI (soit N) (enrichissement de niveau IMAGE)

2. Analyse de la cause d'arrêt de la poursuite de contraste (inférence vers le niveau INDICES D'IMAGE).

   - Soit la poursuite s'est interrompue faute d'un contraste suffisamment élevé : le noeud-image pendant NI est alors déplacé (cf. figure B.06.f).
   - Soit la poursuite s'est interrompue à cause d'un changement de direction du contraste, trois cas se présentent alors (ces cas sont déterminés par une procédure de contrôle sur le niveau IMAGE) :

     o l'analyse s'est interrompue en un point de contraste dont l'interprétation est un noeud-image NI' déjà connu. Les noeuds-image sont alors confondus dans la représentation des indices d'image (cf. figure B.06.g). Si cela s'avère nécessaire, il est possible qu'une ou plusieurs droites-image soient à confondre avec la droite-image DI (cas de DI' sur la figure B.06.h).

     o l'analyse s'est interrompue en un point de contraste dont l'interprétation au niveau INDICES D'IMAGE appartient à une droite-image Δ déjà extraite. La droite-image Δ est alors scindée en deux droites-image Δ_1 et Δ_2, le noeud-image NI devenant extrémité commune aux trois droites-image DI, Δ_1 et Δ_2 (cf. figure B.06.i).

     o l'analyse s'est arrêtée en un point de contraste dont l'interprétation n'est pas un noeud-image connu, et n'appartient à aucune droite-image connue : le noeud-image NI est alors simplement déplacé (cf. figure B.06.j). Le noeud-image NI ainsi repositionné est très fortement susceptible d'être une extrémité commune à plusieurs droites-image qui pourront être détectées par application du premier algorithme au point de l'image de contraste dont NI est l'interprétation directe.

1.3 Procédures d'Enrichissement

Comme nous l'avons dit au cours de la partie A, les procédures présentées ici ne permettent pas de progresser fondamentalement dans l'interprétation de l'image. Elles sont plutôt le reflet d'une autre interprétation des mêmes indices de contraste, qui aurait effectivement pu être effectuée par d'autres méthodes. En effet, rien n'interdit de penser à une procédure d'inférence qui extrairait en premier lieu des lignes-brisées-image, et à des algorithmes d'enrichissement qui construiraient ensuite des noeuds-image et des droites-image à partir des résultats obtenus.

Les procédures d'enrichissement fournissent une base potentielle de travail supplémentaire sur laquelle des algorithmes d'interprétation de l'information visuelle à des niveaux d'abstraction et de décentration plus élevées peuvent être lancés. Par exemple, le calcul de la longueur d'une droite-image est un enrichissement de la droite-image. Toujours à ce sujet, notons finalement que cet enrichissement peut être incrémental, c'est-à-dire effectué à chaque découverte d'un nouveau noeud-image ou d'une nouvelle droite-image. Il peut aussi être global, et n'être activé qu'après détection d'un ensemble d'indices d'image jugé satisfaisant. Ce sera le cas pour le calcul du type d'un noeud-image, procédure d'enrichissement que nous détaillerons. A l'opposé, nous ne ferons qu'évoquer l'algorithme qui extrait des "lignes-brisées-image" qui sont censées traduire les régions de l'image : en effet, nous n'utilisons pas ses résultats pour notre première expérimentation.

1.3.1 Calcul du Type d'un Noeud-Image

Tout noeud-image peut être caractérisé selon 1/ le nombre de droites-image dont il est une extrémité (appelé "valuation" du noeud-image) 2/ le parallélisme éventuel entre ces droites-image (on dira alors que les droites-image sont "opposées") et 3/ la position relative de ces droites-image (relations topologiques). Ces trois critères partitionnent l'ensemble des noeuds-image. Chaque classe de la partition résultante s'appelle un "type".

Par commodité au niveau du langage de description de ces différents types de noeuds-image, il est coutumier de leur associer un nom. C'est ainsi qu'un noeud-image, extrémité commune de trois droites-image exactement, est soit de type "Y ", soit de type "T", ou encore de type "F". Les noms associés aux différents types rappellent les configurations géométriques possibles de plusieurs droites-image ayant une extrémité commune (cas d'une lettre désignatrice) ou sont déduites des combinaisons d'indices de valuation inférieure (cas de plusieurs lettres désignatrices). Outre les facilités de dialogue qu'offre de telles dénominations, elles constituent une base indispensable à l'inférence des indices de scène, du moins de la façon dont nous avons procédé.

La figure B.07 montre les différents types de noeuds-image de valuation inférieure à 5 que *nous* avons distingués. Nous donnons maintenant les définitions plus formelles attachées à ces types, à partir desquelles il est facile d'écrire un algorithme conditionnel de désignation.

o Noeud-image de type **"PENDANT"** : extrémité d'une seule droite-image

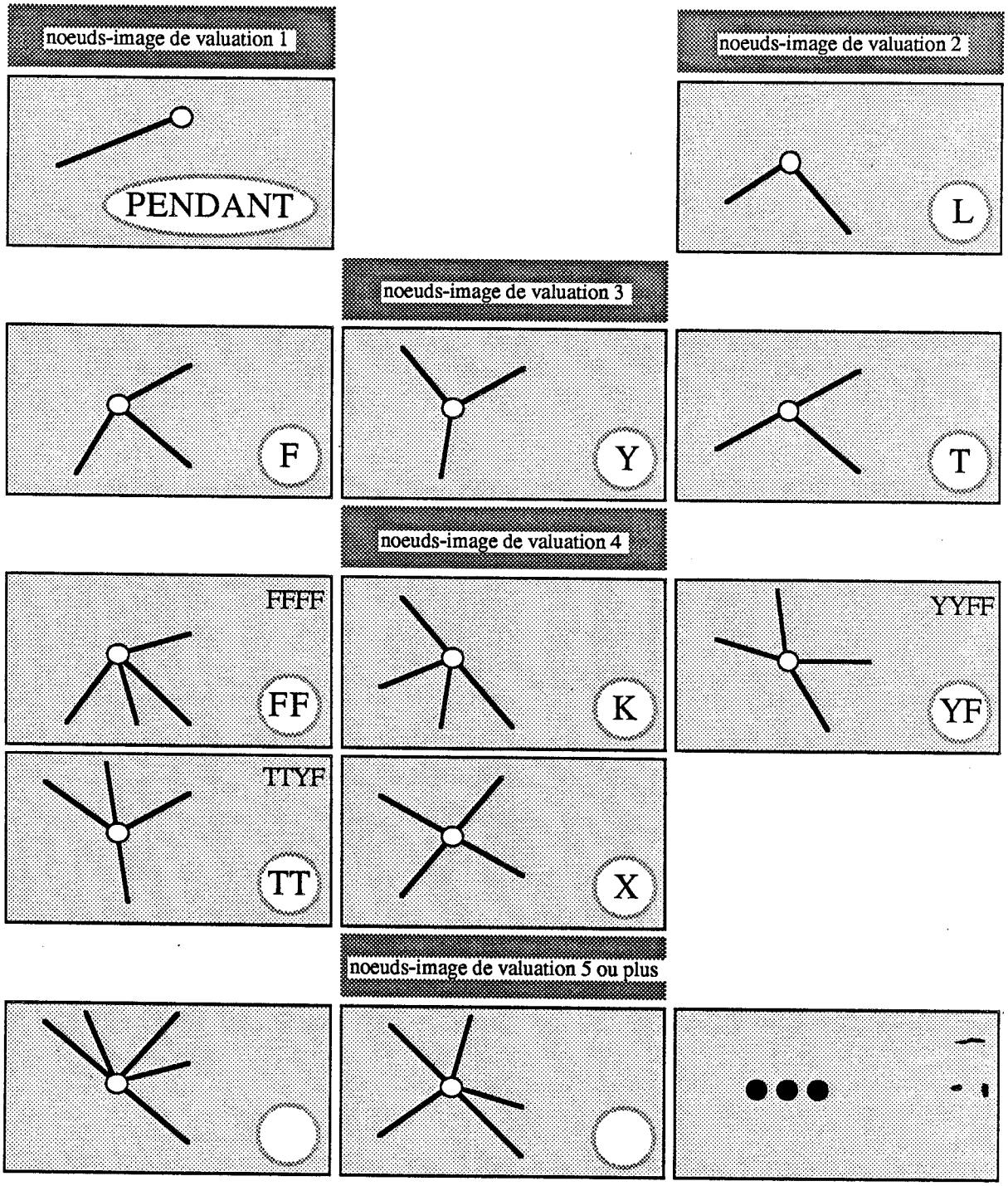


Figure B.07: Les différents types de noeuds-image

o Noeud-image de type "**L**" : extrémité de deux droites-image (obligatoirement non opposées d'après nos définitions)
o Noeud-image de type "**F**" : extrémité de trois droites-image, l'angle absolu maximal entre deux droites-image consécutives étant strictement supérieur à π
o Noeud-image de type "**Y**" : extrémité de trois droites-image non opposées, l'angle absolu maximal entre deux droites-image consécutives étant strictement inférieur à π
o Noeud-image de type "**T**" : extrémité de trois droites-image dont deux sont opposées
o Noeud-image de type "**FF**" : extrémité de quatre droites-image non opposées, l'angle absolu maximal entre deux droites-image consécutives étant strictement supérieur à π
o Noeud-image de type "**TT**" : extrémité de quatre droites-image dont deux seulement sont opposées, l'angle absolu maximal entre deux droites-image consécutives étant strictement inférieur à π
o Noeud-image de type "**K**" : extrémité de quatre droites-image dont deux seulement sont opposées, l'angle absolu maximal entre deux droites-image consécutives étant égal à π
o Noeud-image de type "**X**" : extrémité de quatre droites-image opposées deux à deux
o Noeud-image de type "**YF**" : extrémité de quatre droites-image non opposées, l'angle absolu maximal entre deux droites-image consécutives étant strictement inférieur à π
o Noeud-image de type "**MULTI**" : tout noeud-image de valuation supérieure ou égale à 5.

Remarque : Nous évoquons le type "PENDANT" sans l'avoir explicitement représenté sur la figure B.07 : dans cette première expérimentation du monde des blocs, ce type de noeud-image ne devrait effectivement pas avoir de raison d'exister, les jonctions de l'espace réel ou leurs projections sur une image ayant inévitablement une valuation supérieure ou égale à 2, que ce soit dans le cas d'étude du monde des blocs ou celui du monde d'Origami. Il existe pourtant une raison à leur dénomination, essentielle d'un point de vue de l'implantation : un des effets de l'incrémentalité de nos algorithmes est que toute découverte d'un nouveau noeud-image (par l'algorithme d'interprétation en noeud-image) correspond en général à la création d'au moins un autre noeud-image de type "PENDANT". Vis-à-vis des algorithmes de niveau supérieur, il est donc nécessaire de les désigner de la même façon qu'on le fait pour tous les autres noeuds-image.

1.3.2 Création de Lignes-Brisées-Image

Il est possible, en réorganisant un ensemble de noeuds-image et de droites-image sous la forme de "lignes-brisées-image", de "segmenter en régions" l'image initiale.

Plus formellement, soit NI l'ensemble des noeuds-image, soit DI l'ensemble

des droites-image et soit $TI = \{(N_1, N_2, D) \in NI \times NI \times DI / N_1$ et N_2 sont les noeuds-image extrémités de la droite-image $D\}$, une "ligne-brisée-image" est une suite $LBI = (t_i)_{0<i<n+1}$ de n éléments de TI avec $(t_i = (n_i, n_i^*, d_i)$, vérifiant les propriétés d'être :

continue (les éléments consécutifs de LBI sont connexes) : $(\forall i \in N_{n-1})(n_i^* = n_{i+1})$

d'intérieur topologique vide (son intérieur ne contient aucune droite-image de DI) : $(\forall i \in N_{n-1})(\forall d \in DIn_i^*)$ $(\Theta(d_i, d_{i+1}) \leq \Theta(d_i, d))$ avec $DIn = \{d \in DI / \exists N \in NI$ et $(d, n, N) \in TI\}$ avec $T(D_1, D_2)$ angle orienté au sommet commun des droites-image $D_1 et D_2$

minimale (LBI ne contient pas deux fois le même triplet, même si elle peut contenir deux fois la même droite) : $(\forall (i,j) \in N_n^2)((i \neq /j) \Rightarrow (n_i \neq n_j)$ et $(n_i^* \neq n_j^*))$

maximale : $(\forall (N, N^*, D) \in TI - \{(n_i, n_i^*, d_i) 0 < i < n+1\})$ (l'adjonction de l'élément (N, N^*, D) à la suite LBI provoque la réfutation d'au moins l'une des quatre premières propriétés pour au moins un des éléments de LBI)

Si $LBI = (n_i, n_i^*, d_i)_{0<i<n+1}$ vérifie de plus $n_n^* = n_1$, la ligne-brisée-image est dite "fermée". Dans le cas inverse, la ligne-brisée-image est "ouverte". L'algorithme permettant de construire ces nouveaux indices est assimilable à un algorithme de parcours de graphe avec marquage. L'idée de l'algorithme repose sur le simple fait que toute droite-image appartient exactement à deux lignes-brisées-image. L'algorithme peut être incrémental et être activé au fur et à mesure de la création de nouveaux noeuds-image ou droites-image, ou bien, au contraire, appliqué après extraction de tous ces indices d'image. Dans tous les cas, il fournit le réseau des lignes-brisées-image que constituent les noeuds-image et les droites-image extraits.

Bien entendu, il ne faut pas accorder au terme de ligne-brisée-image une valeur identique à celle qui peut être attribuée aux indices de régions obtenus par des algorithmes classiques d'extension ou de découpage [BALLA-82] [GORDI-83] : en effet, la notion de ligne-brisée-image ne préjuge en rien de l'intérieur d'une région, mais tout au plus de sa frontière. Outre son incapacité à exprimer des caractéristiques (comme la texture, la couleur) propres à la région de l'image qu'elle délimite, une ligne-brisée-image ne peut garantir l'homogénéité de cette région : elle ne permet pas, par exemple, d'affirmer qu'aucun autre indice d'image (et donc aucun autre objet) ne se trouve à l'intérieur du périmètre.

Non utilisée pour notre première expérimentation, nous reviendrons cependant sur l'intérêt de cette notion au cours de la partie C.

2 Des Indices d'Image aux Indices de Scène

Il s'agit ici d'interpréter les indices d'image extraits en *leurs* différentes interprétations, du point de vue de leur forme spatiale et d'établir des relations en adjacence

et/ou en profondeur qu'ils peuvent présenter. Insistons sur ce point particulier, **l'interprétation d'un noeud-image (resp. d'une ligne-image) peut engendrer plusieurs indices de scène qui seront appelés "sommets" (resp. "arêtes ")**; **parfois même ils n'en engendrent aucun mais permettent simplement d'inférer des relations entre indices de scène.** Notons simplement une restriction que nous apportons ici au problème général : nous ne considérons dans notre présentation que des droites-image dont l'interprétation fournit *au moins* une "arête" physiquement réelle.

2.1 Procédures d'inférence et de Contrôle

2.1.1 Définition des Opérateurs d'Interprétation

A chaque type de noeud-image, correspond un ensemble d'"opérateurs d'interprétation" possibles qui correspondent à autant de "réalités scéniques" possibles du monde physique. La figure B.08 représente, pour un noeud-image de type "T ", différentes réalités scéniques et leurs opérateurs d'interprétation correspondant (soient O_{T_1}, O_{T_2}, O_{T_3}, O_{T_4} et O_{T_5}). Nous en verrons d'autres dans le cas très particulier de l'application industrielle de la dépalétisation.

Soit D_{TYPE} l'ensemble des droites-image formant un noeud-image de type TYPE, et soit I_{TYPE} l'ensemble des opérateurs d'interprétation applicables sur un noeud-image de type TYPE. I_{TYPE} est égal à l'ensemble des pseudo-partitions[1] que l'on peut effectuer sur D_{TYPE}, et telles que chacune de ces pseudo-partitions corresponde à une réalité scénique.

C'est ainsi que $\{\{D_1, D_2\}\} \in I_L$, D_1 et D_2 étant les deux droites-image dont l'extrémité commune est un noeud-image de type "L". De la même façon, pour un noeud-image de type T, constitué des trois droites-image DI_1, DI_2 et DI_3 (cf. figure B.08), l'ensemble I_T est constitué des cinq éléments O_{T_1} à O_{T_5}, définis de la manière suivante :

$$O_{T_1} = \{\{D_1\}\{D_2, D_3\}\} \qquad O_{T_2} = \{\{D_1, D_2\}\{D_3, D_1\}\}$$
$$O_{T_3} = \{\{D_1, D_2\}\{D_3\}\} \qquad O_{T_4} = \{\{D_2\}\{D_3, D_1\}\}$$
$$O_{T_5} = \{\{D_1, D_2, D_3\}\}$$

2.1.2 Application des Opérateurs d'Interprétation

D'un point de vue de l'implantation, l'application de chaque opérateur d'interprétation peut être considéré comme une règle de production qui fait partie de la base des connaissances permanentes du système. L'application de tout opérateur O_{TYPE_k} sur un noeud-image N de type TYPE fournit un ensemble $S^N_{TYPE_k}$ structurellement identique à celui qui définit l'opérateur appliqué, obtenu en remplaçant les occurences de D_i par un élément A_i et en réécrivant les éléments de chaque classe de

[1] **pseudo-partition** : on appelle pseudo-partition d'un ensemble E tout sous-ensemble P de l'ensemble des parties de E tel qu'aucun élément de P n'est vide et que l'union des éléments de P soit égal à E. Par rapport à la notion de partition, les sous-ensembles ne sont donc pas forcément disjoints deux à deux.

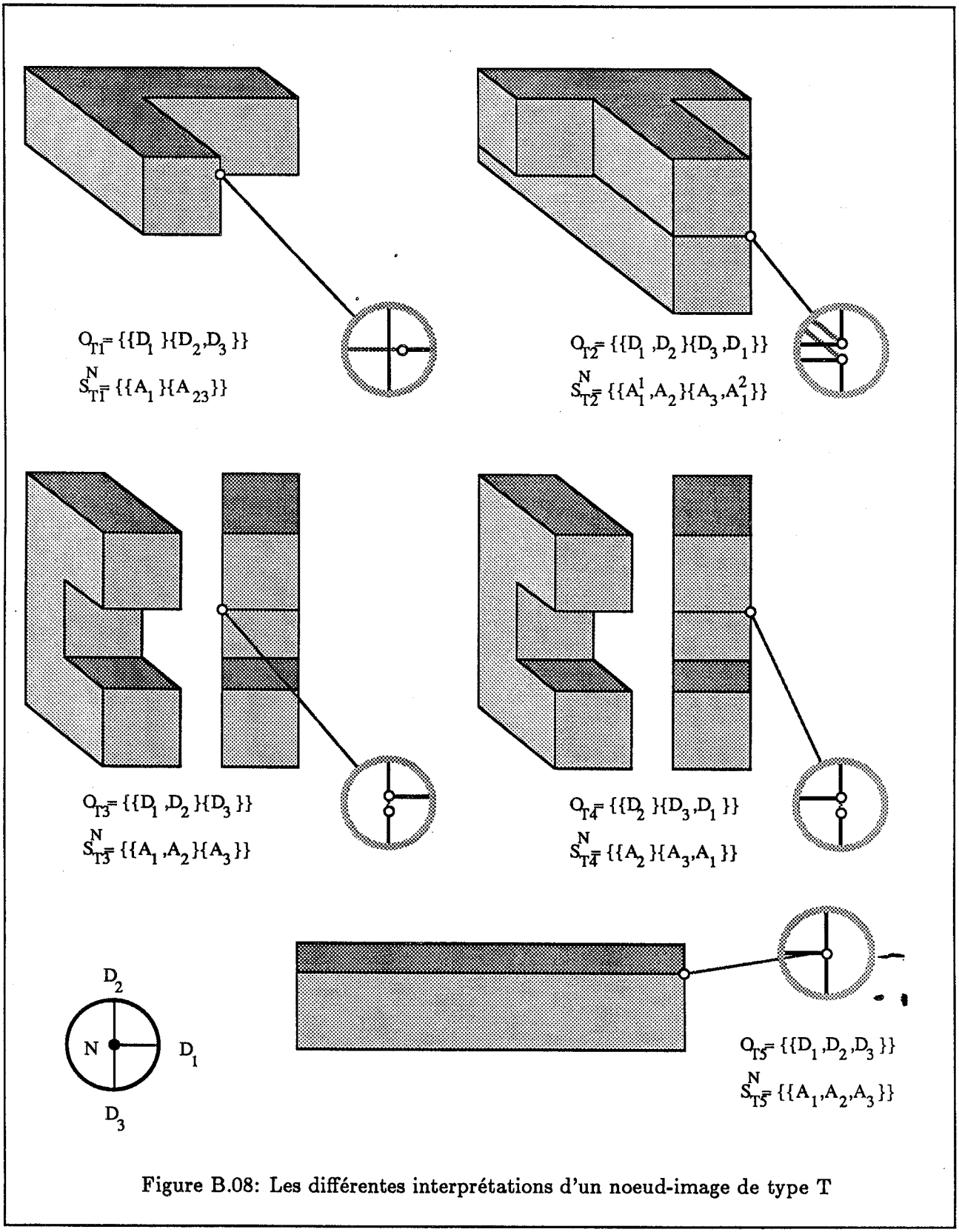


Figure B.08: Les différentes interprétations d'un noeud-image de type T

pseudo-partition moyennant deux règles de transformation :

- Si une droite-image D_i figure p fois parmi les éléments de O_{TYPE_k}, alors elle engendre autant d'éléments A_i^p en lieu et place des occurences de D_i au sein de $S^N_{TYPE_k}$.
- Pour tout élément de O_{TYPE_k} composé de deux droites-image D_i et D_j opposées, la correspondance au sein de $S^N_{TYPE_k}$ se réduit à un unique élément A_{ij}.

L'application du seul opérateur de I_L fournit ainsi l'ensemble $\{\{A_1, A_2\}\}$. De même, l'application des cinq opérateurs de I_T fournit l'ensemble (cf. figure B.08) :

$$S^N_{T_1} = \{\{A_1\}\{A_{23}\}\} \qquad S^N_{T_2} = \{\{A_1^1, A_2\}\{A_3, A_1^2\}\}$$
$$S^N_{T_3} = \{\{A_1, A_2\}\{A_3\}\} \qquad S^N_{T_4} = \{\{A_2\}\{A_3, A_1\}\}$$
$$S^N_{T_5} = \{\{A_1, A_2, A_3\}\}$$

L'exposé de cette définition nous amène à poser les deux définitions suivantes :

Tout élément A_j (resp. A_j^p, A_{ij}) s'appelle une *"arête"* et est une interprétation au niveau INDICES DE SCENE de la droite-image D_j (resp. de la droite-image D_j, des deux droites-image D_i et D_j).

Tout élément de $S^N_{TYPE_k}$, s'il ne se réduit pas un singleton de la forme $\{A_{ij}\}$, s'appelle un *"sommet"* et est une interprétation du noeud-image N.

Remarque 1 : Chaque sommet est défini comme l'extrémité d'une ou plusieurs arêtes (interprétations de droites-image) et est supposé traduire des propriétés d'adjacence réelle entre celles-ci. Mais un sommet n'a pas toujours une signification, et spécialement une position toujours bien précise dans l'image : c'est le cas des sommets qui se réduisent à un seul singleton de la forme $\{A_i\}$. Dans la figure B.08, on peut penser a priori que certaines interprétations devraient engendrer moins de sommets qu'elles ne le font par notre définition. Particulièrement, le cas d'application de l'opérateur O_{T_1} ne correspond à aucun sommet réel alors que la définition constructive que nous venons de donner en fournit un. Il faut donc se méfier de la notion de sommet telle qu'elle est définie : un sommet S dans la structure d'indices de scène ne représente pas obligatoirement un "sommet de l'espace réel" dont l'indice (ou l'un des indices) correspondant(s) au niveau INDICES DE SCENE se situe(nt) au même emplacement dans l'image que le noeud-image NI qu'interprète partiellement S. Prenons par exemple le cas du sommet engendré par application de l'opérateur O_{T_1}. Pouvant être considéré comme un "pseudo-sommet", il est, plus que tout autre, décentré de l'image initiale. En effet, il signifie simplement, moyennant l'occultation provoquée par l'interprétation de D_2 et D_3 en l'unique arête A_{23}, que l'arête A_1 possède une extrémité suivant la direction définie par A_1, sans pouvoir le localiser précisément. Mais, ce faisant, nous disposons de plus d'une information symbolique en profondeur ...

Remarque 2 : ... En effet, la connaissance du fait que l'arête A_{23} occulte l'arête A_1 se traduit par l'inférence de la relation symbolique A_1 EST DERRIERE A_{23}.

Tout comme l'adjacence de plusieurs arêtes ayant une extrémité commune, l'information de profondeur ici découverte traduit bien toute l'abstraction effectuée sur la structure indices d'image. D'un point de vue du modèle de représentation du niveau INDICES DE SCENE, il est donc nécessaire de faire figurer deux dimensions numériques, ainsi qu'une dimension symbolique qui traduit les relations en adjacence et en profondeur relatives entre indices de scène.

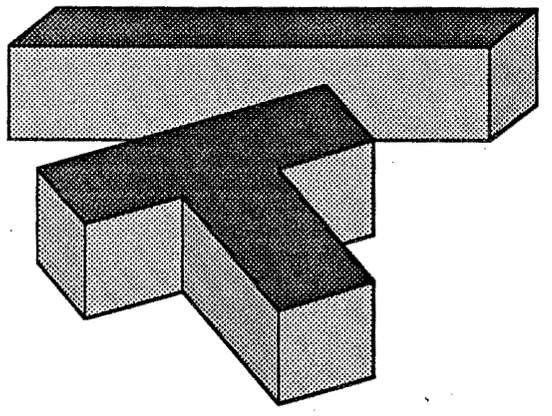
2.1.3 Algorithmes d'Inférence

Nous venons de voir comment il était possible d'obtenir *localement* des interprétations de noeuds-image et de droites-image. Pourtant, l'étude de ces inférences doit aller plus loin pour tenir compte d'une réalité géométrique : une ligne-image possède deux extrémités. Les interprétations de la droite-image et de ses noeuds-image extrémités doivent être cohérentes, ce qui signifie que la réalité scénique associée à l'interprétation de la droite ne peut varier le long de cette droite-image. D'autre part, il est possible d'améliorer notablement les inférences dans le sens d'une réduction de la combinatoire par applications de "règles contextuelles" qui permettent de diminuer le nombre, voire d'ordonner les interprétations possibles d'un noeud-image NI de type "TYPE".

L'algorithme que nous utilisons pour l'interprétation des indices d'image en indices de scène peut être décrit sous la forme de quatre étapes successives, dont trois sont propres à l'interprétation des noeuds-image en sommets. Cet algorithme est illustré par la figure B.09.a, où la scène expérimentale est constituée des deux pentacubes "T" et "I", dont les positions relatives provoquent un "alignement accidentel " :

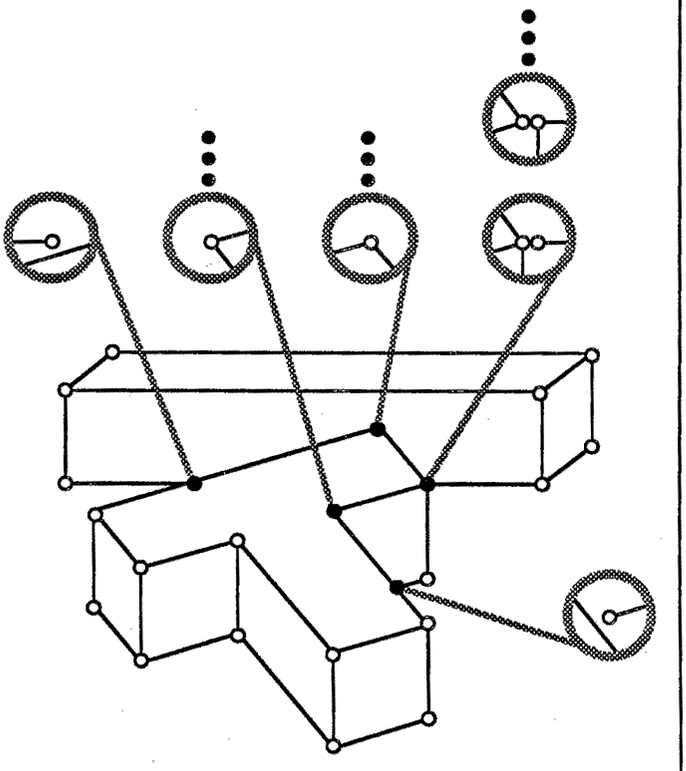
1. Chaque noeud-image est a priori sujet à l'application de tous les opérateurs d'interprétation déduits de son type, correspondant donc à des réalités scéniques (soient $O_1, ...O_5$ pour le noeud-image de type "T" sur la figure). C'est ce que nous avons introduit au paragraphe précédent.

2. L'étude de l'environnement local de ce noeud-image permet de réduire cet ensemble d'opérateurs I_{TYPE} en un ensemble IR_{TYPE} de cardinal inférieur. La démarche que nous avons voulu suivre consiste à utiliser des règles de voisinage très simples (appelées "règles contextuelles") en chacun des noeuds-image présents, règles dont l'application suffit à permettre la compréhension de la forme dans l'espace de la plupart de ces noeuds-image. Bien sûr, ces réductions d'interprétations n'ont aucun caractère de généralité, et correspondent le plus souvent à un contexte particulier. De la même façon qu'il est possible de définir plusieurs jeux d'étiquetage pour un même domaine d'expérimentation, il est possible de définir différents ensembles de règles contextuelles. Tout comme dans le cas de l'étiquetage, et sans ignorer toute la dimension du problème, nous ne discutons pas ici de la validité de tel ou tel ensemble de règles contextuelles. Nous en donnons simplement un exemple, adaptation simplifiée de la base sur laquelle [FALK-72] interprète un dessin de lignes, en supposant que

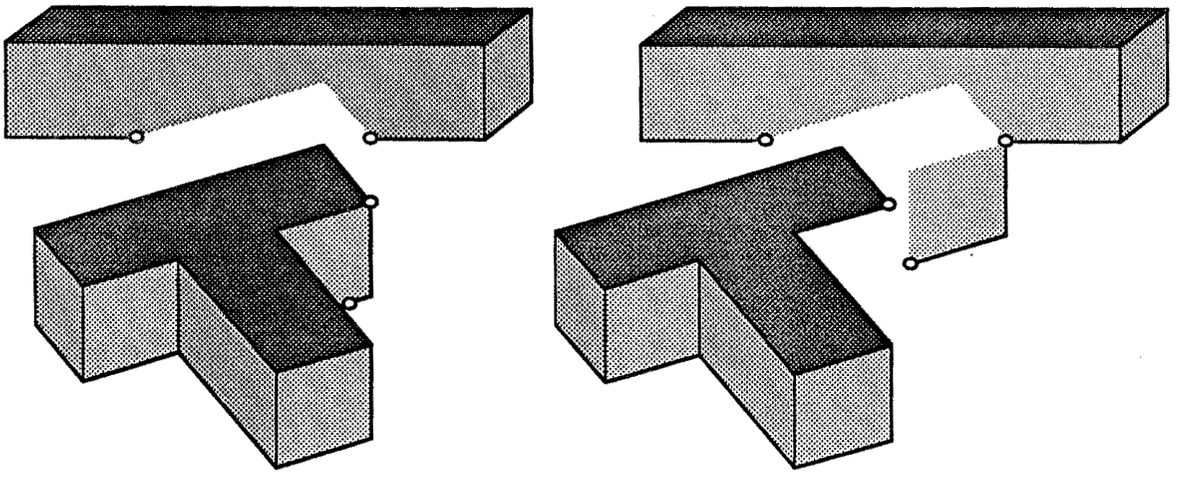


B.09.a: Structure d'indices d'image initiale

L'application de règles contextuelles très simples en chacun des noeuds-image présents permet la compréhension de la plupart des noeuds-image. L'application de l'algorithme de propagation de contraintes, grâce aux connexités offertes par les droites-image, permet de préciser l'interprétation des derniers noeuds-image, et fournit les différentes structures d'indices de scène qui sont les interprétations possibles de la structure d'indices d'image initiale



B.09.b: Interprétation de noeuds-image



B.09.c: Les deux interprétations de la structure d'indices d'image correspondant à des réalités scéniques

Figure B.09: Propagation des interprétations en indices de scène

le système possède la connaissance supplémentaire de savoir ce qu'est et où se trouve le FOND de la scène.

L'exemple se réfère aux figures B.07 et B.08. Une interprétation est dite "UNIQUE" lorsqu'elle permet de réduire l'ensemble des opérateurs d'interprétation possibles à un singleton, correspondant à l'interprétation la plus répandue dans le monde des blocs :

- Noeud-image N de type "**T**" : l'interprétation est UNIQUE (application de $O_{T_1}, N.IR_T = \{\{\{A_1\}\{A_{23}\}\}\}$ sauf si la région opposée au pied du "T" est le FOND de la scène (application de O_{T_2} à $O_{T_5}, N.IR_T = I_T - \{\{\{A_1\}\{A_{23}\}\}\}$), si le pied du "T" est l'une des droites-image opposées composant un noeud-image de type "K" ou si le pied du "T" est une des droites-image non opposées d'un noeud-image de type "X" (application de $O_{T_2}, N.IR_T = \{\{\{A_1^1, A_2\}\{A_3, A_1^2\}\}\}$).
- Noeud-image N de type "**F**" : l'interprétation est UNIQUE ($N.IR_F = \{\{\{A_1, A_2, A_3\}\}\}$) sauf si la droite-image centrale D_2 est aussi l'une des droites-image opposées formant un noeud-image de type "K" ($N.IR_F = \{\{\{A_1, A_2^1\}\{A_2^2, A_3\}\}\}$, ou si l'une des régions adjacentes à la droite-image centrale est le FOND de la scène ($IR_F = I_F$)
- Noeud-image N de type "**Y**" : l'interprétation est UNIQUE ($N.IR_Y = \{\{\{A_1, A_2, A_3\}\}\}$) si au moins l'une des droites-image est aussi la droite centrale d'un noeud-image de type "F" qui est correct, sinon $IR_Y = I_Y$.
- Noeud-image N de type "**TT**" : l'interprétation est UNIQUE ($N.IR_{TT} = \{\{\{A_4^1, A_1, A_2^1\}\{A_2^2, A_3, A_4^2\}\}\}$), sauf si l'une des droites-image non opposées D_2 ou D_4 est aussi le pied d'un noeud-image de type "T" ($IR_{TT} = I_{TT}$).
- Pour **tous les autres** noeuds-image N, ($IR_{TYPE} = I_{TYPE}$).

3. L'application au noeud-image donné des opérateurs d'interprétation subsistant à cette réduction, fournit ses interprétations possibles, ainsi que celles des droites-image dont il est une extrémité (inférence vers le niveau INDICES DE SCENE, cf. figure B.09.b).

4. Il reste alors, par un algorithme de propagation de contraintes, à vérifier la compatibilité de ces nouvelles interprétations par rapport à celles déjà effectuées sur les noeuds-image voisins et réciproquement, afin de ne conserver que des interprétations des noeuds-image voisins qui puissent coexister avec les interprétations retenues pour NI (contrôle sur le niveau INDICE D'IMAGE). Ce type d'algorithme à récursivité croisée entre les différents noeuds-image, directement issu de celui de [WALTZ-75] et que nous ne reprenons pas ici en détail, est fondée sur l'observation des connexités apparentes entre droites-image en vérifiant la cohérence globale des interprétations. L'application de l'algorithme suffit en général à lever le voile sur l'interprétation des derniers noeuds-image qui n'auraient pas encore été interprétés.

Remarque 1 : Il est clair que les noeuds-image de type "PENDANT" n'ont aucune signification physique, et que leur interprétation au niveau INDICES DE SCENE est très ambigue, puisqu'ils correspondent à un des défauts de connaissance aux niveau IMAGE. Néanmoins, quelle est la signification au niveau INDICES DE SCENE d'un noeud-image de type PENDANT? C'est qu'il existe, dans la direction de ce noeud pendant, un sommet de l'espace réel. Le seul moyen de les interpréter correctement, pour rester cohérent avec l'ensemble de la procédure exposée, est donc de les interpréter en un ou plusieurs singleton(s) de type $\{A_i\}$ (donc comme des pseudo-sommets) car les droites-image auxquelles ils sont liés, elles, participent à l'interprétation des autres noeuds-image. Par rapport au cas général des pseudo-sommets, le défaut d'interprétation n'est plus ici dû à une occultation quelconque, mais à un manque d'information au niveau INDICES D'IMAGE. Le caractère incrémental des algorithmes pose un problème identique lorsqu'il s'agit d'interpréter un noeud-image dont la valuation observée n'est pas celle qu'elle devrait avoir par rapport à la projection de la réalité scénique observée (cas d'une mauvaise interprétation en indices d'image de la structure image). Mais là, nous ne pouvons pas décider quels sont ces noeuds-image à problèmes. Cependant, nous ne pouvons pas nous dispenser de les interpréter : le risque est alors trop grand d'utiliser des algorithmes qui, disposant de trop peu de connaissances suffisamment fiables, seraient incapables d'une quelconque interprétation pour l'inférence des niveaux supérieurs ("effet d'horizon").

Remarque 2 : Plus la valuation d'un noeud-image est élevée, plus le nombre d'interprétations correspondant à des réalités scéniques est élevé. La raison majeure à s'en tenir à des noeuds-image de valuation inférieure ou égale à 4 est liée au contexte de notre expérimentation dans le monde des blocs. Au-delà de cette valeur (cas de noeuds-image de type MULTI), la connaissance explicite des différentes interprétations en indices de scène nous paraît inutile vis-à-vis de l'explosion combinatoire qu'elle engendre, d'autant que pour les objets étudiés, de tels noeuds-image correspondent à des alignements accidentels qu'il est en général possible de préciser par le simple algorithme de propagation de contraintes sur les interprétations des noeuds-image voisins.

2.2 Procédure d'Enrichissement

Les interprétations des noeuds-image et droites-image étant construites, il est alors possible de construire d'autres indices de niveau INDICES DE SCENE. C'est par exemple le cas pour les "faces", qui correspondent aux surfaces de la scène réelle qui sont observées. L'idée de l'algorithme qui assure cet enrichissement est un parcours du graphe des indices de scène, avec exploitation systématique des propriétés entre arêtes et de leurs compatibilités d'interprétation par rapport aux relations de profondeur découvertes. L'application d'un tel algorithme permet de découvrir les différentes structures surfaciques qui correspondent à l'une des réalités scéniques possibles de la scène observée, et qui constituent les interprétations possibles de la structure indices d'image initiale (cf. figure B.09.c).

Remarque 1 : L'algorithme de propagation de contraintes utilisé pour l'inférence des indices de scène est une procédure de contrôle entre le niveau INDICES D'IMAGE et INDICES DE SCENE, au sens où nous l'avons défini dans la partie A. N'y interviennent que des connaissances contenues en ces niveaux. Toujours suivant nos définitions, le choix de telle ou telle structure d'indices de scène parmi celles qui sont physiquement possible n'est pas le fait d'une quelconque procédure de contrôle entre ces deux mêmes et seuls niveaux : la résolution du problème de choix ne peut être réalisée que grâce aux inférences ultérieures où grâce à des informations issues des niveaux OBJET et/ou SCENE.

Remarque 2 : Dans le sens où la méthode utilisée est incapable d'interpréter des images où trop de noeuds-image auraient une valuation supérieure ou égale à 5, *la méthode utilisée n'est pas complète.* En revanche, dans la mesure où elle permet d'interpréter des réalités scéniques et éventuellement de réfuter les interprétations qui n'en sont pas [HUFFM-71], *la méthode employée est logiquement correcte*, même s'il est indiscutable que ce résultat reste à prouver. Bien que nous introduisions ces notions par rapport à la Déduction Logique, il est tout aussi vrai que nous ne nous attachons pas ici à définir des propriétés de systèmes formels, mais à établir des relations entre réalité physique et un système formel.

3 Modélisation et Identification des Objets du Monde des Blocs

Les modèles de représentation utilisés sont fortement dépendants des algorithmes et du domaine étudié. De façon générale, toutes les structures relèvent d'une modélisation par sommets et arêtes, héritée de celle des objets qu'ils étudient, et qu'illustrent parfaitement les pentacubes sur lesquelles la plupart des expérimentations ont été effectuées (cf. figure B.10). Au niveau INDICES D'IMAGE et INDICES DE SCENE, tout indice extrait contient des informations qui lui sont propres (par exemple au niveau INDICES D'IMAGE, les coordonnées dans l'image d'un noeud-image), mais aussi les informations qui le relient aux autres indices de même niveau (par exemple au niveau INDICES DE SCENE, les faces auxquelles appartient un sommet). Les informations issues de l'application des procédures d'enrichissement correspond à une redondance explicite de l'information.

Si l'on dispose de modèles d'objets qu'il s'agit de distinguer dans la scène sur le simple critère de forme, stopper le processus de compréhension au niveau INDICES DE SCENE suffit, sans aucun calibrage, à effectuer la reconnaissance des objets présents dans la scène. Pour permettre l'aboutissement d'une quelconque procédure de mise en correspondance assurant ce but, plusieurs propriétés sont utilisés : le principe de l'unicité de mise en correspondance avec les faces réelles de l'objet, le respect de la valuation apparente possible des sommets réels, l'utilisation de la conservation par projection de certaines propriétés géométriques, comme la connexité entre arêtes, le parallélisme, et les règles de déformation projective des

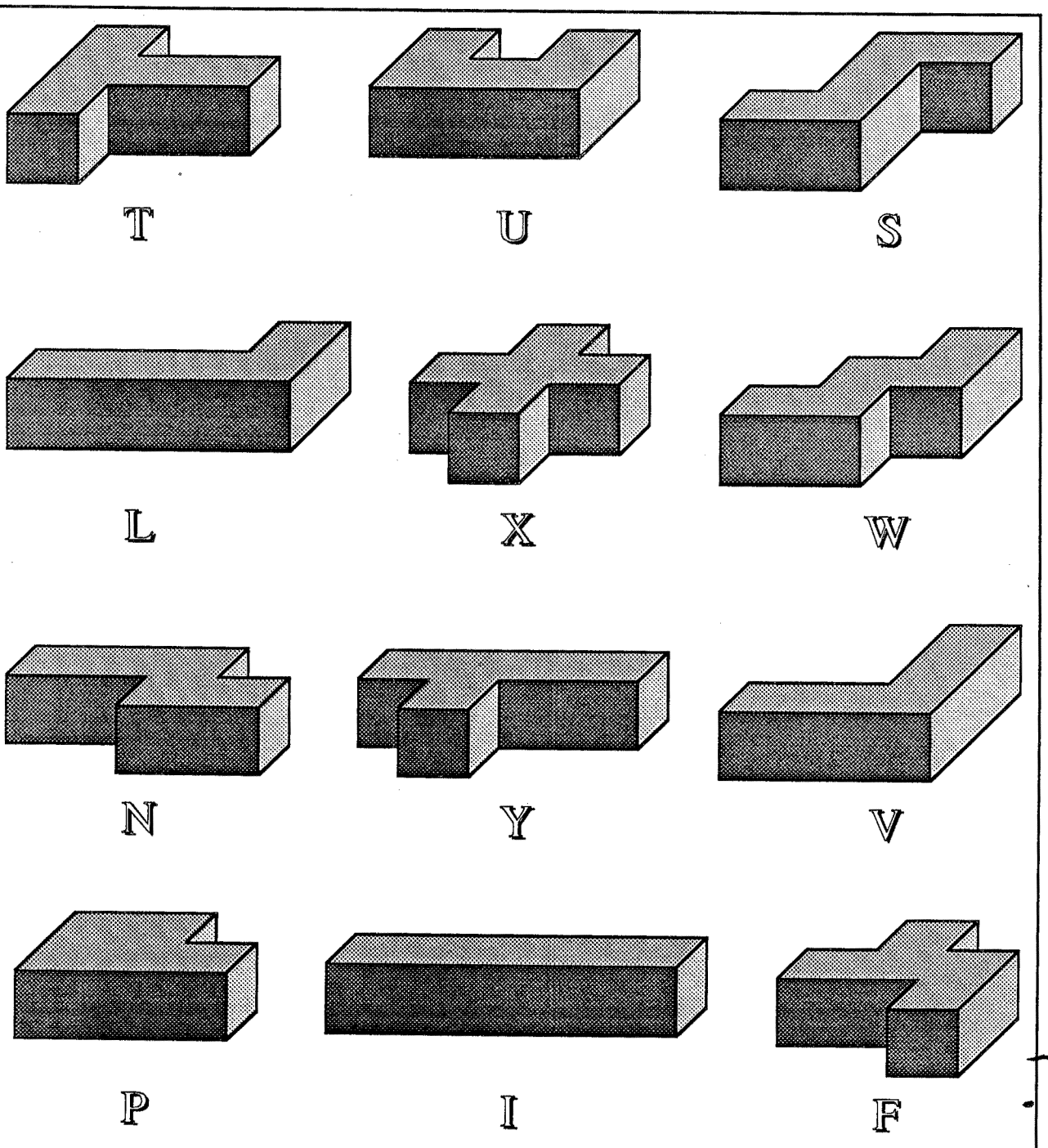


Figure B.10: Modélisation du Monde des Blocs par sommets et arêtes

longueurs et des angles pour les surfaces planes [DEMAZ-82]. D'autres connaissances peuvent bien sûr être utilisées. C'est le cas par exemple de la symétrie oblique, qu'utilise d'ailleurs [KANAD-80b], non seulement dans un but de reconnaissance, mais aussi en tant que connaissance contextuelle pour inférer des indices similaires à nos indices de scène.

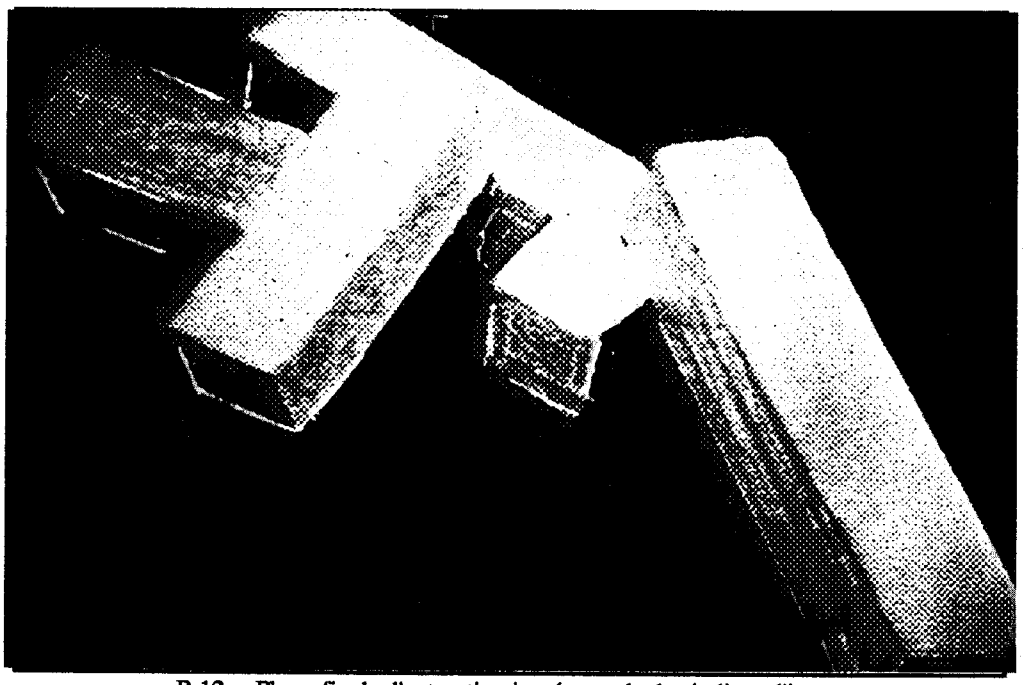
4 Quelques résultats

Les deux algorithmes d'interprétation des indices de niveau IMAGE en indices-image (suivi de droite-image et interprétation en noeud-image) donnent la potentialité nécessaire à un contrôle global extrêmement simple d'extraire tous les noeuds-image et droites-image connexes à l'interprétation d'un point de contraste initial. Il suffit en effet d'enchaîner les deux algorithmes sur l'ensemble des noeuds-image pendants et d'arrêter l'itération lorsque l'ensemble des indices d'image est stabilisé. Les figures B.11 et B.12.a montrent le résultat de l'application d'un tel contrôle grâce à trois initialisations successives de l'image. La figure B.12.b visualise la structure des indices d'image finalement obtenus.

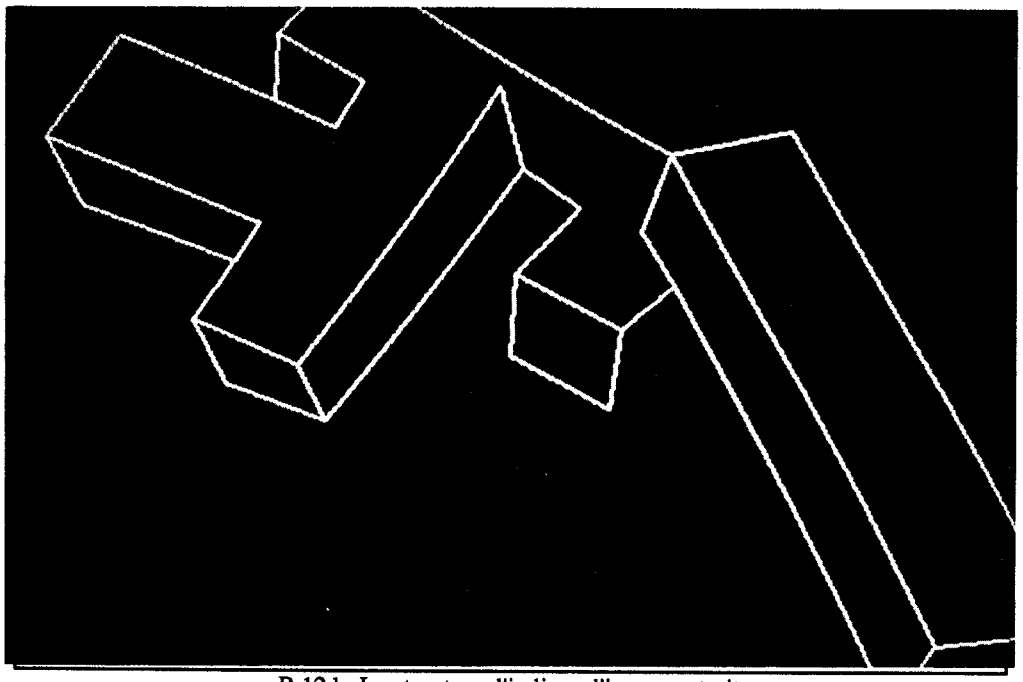
B.11.a: Phase intermédiaire d'extraction incrémentale des indices d'image

B.11.b: Phase intermédiaire d'extraction incrémentale des indices d'image

Figure B.11: Extraction des indices d'image: résultats expérimentaux (1)



B.12.a: Phase finale d'extraction incrémentale des indices d'image



B.12.b: La structure d'indices d'image extraite

Figure B.12: Extraction des indices d'image: résultats expérimentaux (2)

Chapitre B.IV

Application à la Localisation de Paquets sur une Palette

Consécutivement à notre expérimentation de la distinction entre indices d'image et indices de scène dans le monde des blocs, nous avons eu la chance de pouvoir tester ces idées par le biais d'une étude de faisabilité d'un problème industriel concernant le déchargement de palettes constituées de paquets parallélépipédiques. Cette étude a été effectuée en collaboration avec la société ITMI à laquelle le problème avait été soumis.

1 Le problème

Il s'agit d'intégrer un système de vision au sein d'un ensemble robotisé pour permettre le déchargement de palettes. La définition et les conditions d'exploitation sont caractérisées par plusieurs critères :

- Une palette est constituée de paquets de taille identique et arrangés suivant un plan appelé "plan de palétisation". Plusieurs couches de paquets peuvent figurer sur une même palette : le schéma de palétisation, même s'il reste connu du système, peut varier d'une couche à l'autre. Quoiqu'il en soit, tous les paquets sont non seulement de formes identiques, mais encore posés sur la palette sur la même face.

- Le dispositif imaginé pour décharger les paquets est un bras manipulateur muni de ventouses qui doit venir prendre chaque paquet au centre de sa face de dessus, lorsque cette face est libérée de la présence des autres paquets qui éventuellement la recouvrent et empêchent donc sa manipulation. La manipulation d'un paquet peut s'effectuer sans que la disposition des autres subisse une altération. L'analyse d'une seule image doit donc suffire à comprendre la scène observée.

- Le vrai problème se résume en fait en une seule phrase : par rapport au schéma de palétisation, les paquets chargés sur la palette peuvent avoir glissé. Ce déplacement, souvent parallèle aux deux directions théoriques des côtés des paquets, diminue de façon importante le crédit qu'on peut accorder au schéma même de palétisation (cf. haut de la figure B.13).

Une première analyse du problème consiste tout naturellement, vue la simplicité apparente du problème, à se tourner du côté de la vision bidimensionnelle et même à n'utiliser que la vision binaire. En effet, disposant du schéma de palétisation, et pouvant connaître la hauteur de chaque couche a priori, on peut penser qu'il suffit d'une caméra calibrée dans un plan bidimensionnel : chaque nouvelle couche est amenée par élévation de la palette, de sorte que le plan d'affleurement des surfaces de dessus des cartons de la couche la plus élevée de la palette vienne coincider avec le plan de calibrage, le plan jusqu'auquel le bras manipulateur vient saisir les paquets. La solution de la vision binaire n'est pas adéquate pour plusieurs raisons :

- Les paquets présentent sur leurs faces et de façon plus ou moins ordonnée des inscriptions et aussi des étiquettes qui peuvent elles-même être remplies d'inscriptions. D'autre part, les surfaces à observer sont fortement texturées : ce phénomène inévitable est dû à l'utilisation d'emballages déformables qui, typiquement, sont en carton. Finalement, il est possible qu'une configuration quelconque présente un manque de contraste entre surfaces carton - carton. Dans tous les cas, il s'agit d'éliminer ces défauts artificiellement. Chacun connaît les limitations de la vision binaire, en particulier vis-à-vis des conditions qu'elle requiert du point de vue de l'éclairage.

- Chacun connaît aussi la difficulté de la vision binaire à expliciter de façon non ambigue des indices de région qui soient tout à la fois précis dans leur définition et qui ne se recouvrent pas. Cela n'est possible dans le contexte que nous étudions que si les paquets se déplacent franchement jusqu'à se séparer les uns des autres. Malgré l'existence des déplacements des paquets, on ne peut pourtant pas supposer une telle propriété a priori.

Une seconde analyse permet, sinon de proscrire une analyse bidimensionnelle guidée par le but du type de celle de [SOUVI-83], du moins d'en montrer les limites. Les paquets sont tous identiques, et cela permet effectivement de penser à un algorithme de recherche dans l'image d'une structure composée d'indices d'image organisés en parallélogramme. Profitant du schéma de palétisation, la reconnaissance permet, une fois un paquet découvert, d'en déduire la position des autres.

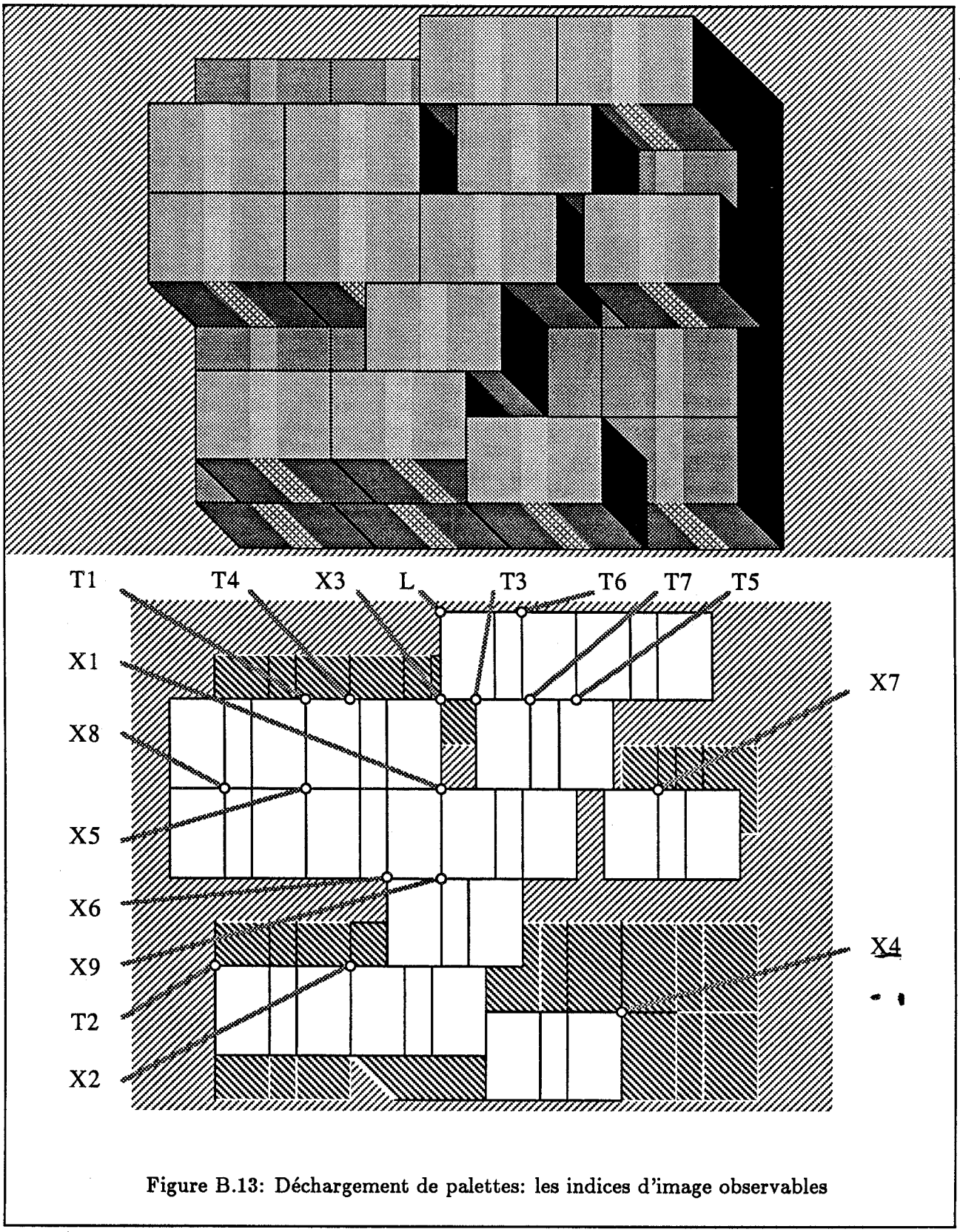


Figure B.13: Déchargement de palettes: les indices d'image observables

Ceci est vrai théoriquement cependant les paquets peuvent avoir glissé dans les deux directions de palétisation, d'où une inefficacité de la procédure. Inversement, si le schéma de palétisation n'est pas trop modifié, la procédure se heurte à un second problème : de larges bandes adhésives recouvrent les paquets pour les fermer. La détection des bords de ces bandes, placées au même endroit sur tous les paquets, peut provoquer l'éclosion d'un grand nombre de correspondances illicites (cf. haut de la figure B.13), et donc de faire exploser la combinatoire du processus de prédiction - vérification.

Tous ces problèmes, ajoutés au fait que si la prise de vue est orthogonale au plan des surfaces observées, les droites-image visibles de l'image n'appartiennent pas forcément aux projections des paquets de la couche supérieure, font que nous n'avons gardé de ces analyses que trois points positifs : 1/ l'analyse d'une seule image permet la compréhension de la géométrie de la couche supérieure de la palette à décharger 2/ la possibilité de connaître l'élévation des paquets à décharger, et qu'il suffit donc de décrire la surface des paquets qui servira de face de préhension pour le manipulateur de déchargement, 3/ la connaissance, d'après le schéma de palétisation, de deux directions privilégiées Δ_1 et Δ_2 qui sont définies par les axes médians théoriques des surfaces supérieures des paquets à décharger.

2 Notre approche

La méthode que nous avons suivie est celle que nous avons présentée au cours du chapitre précédent, dans un domaine d'expérimentation encore plus contraint que celui du monde des blocs. Pour nous affranchir du problème ... :

- ... posé par toute procédure orientée vers la détection d'un parallélogramme et pour résoudre le problème du glissement, nous utilisons une méthode d'analyse ascendante.

- ... des étiquettes, nous employons une méthode incrémentale d'extraction de contours et non de régions. S'agissant d'éviter le problème d'interprétation de la texture des emballages, nous segmentons l'image en droites-image et noeuds-image.

- ... des lignes de contour issues des objets des couches inférieures, nous choisissons une prise de vue oblique qui permettra de discriminer facilement les indices d'image issus des paquets de la couche à décharger.

- ... de l'interprétation des indices d'image de sorte qu'ils traduisent tous une réalité scénique, nous utilisons des opérateurs d'interprétation.

- ... de l'extraction des indices d'image qui sont utiles à la localisation, nous nous référons à des règles d'interprétation contextuelles de voisinage.

3 Le type et l'extraction des indices d'image

Les seuls types d'indices d'image à considérer, car utiles à *l'atteinte d'un but prédéfini à l'avance*, sont les noeuds de type L, T, et X. Tous les autres noeuds-image extraits au fur et à mesure des deux procédures "d'interprétation en noeud-image" et de "poursuite de droite-image", voient leur développement enfermés par la seule poursuite des droites-image parallèles aux directions Δ_1 et Δ_2, et par la suppression pure et simple, dans la structure des indices d'image extraits, des droites-image qui ne suivent pas cette direction.

En outre, nous avons rajouté à l'algorithme de poursuite de droite-image un détecteur de départs potentiels de droites-image : il parcourt les points de l'image de part et d'autre de la droite-image et signale la présence d'un fort contraste suivant une des deux directions Δ_1 et Δ_2.

Le problème de l'initialisation de la procédure d'extraction des indices d'image n'en n'est pas réellement un : en effet, la prise de vue que nous avons choisie permet de connaître dans quel demi-écran on peut trouver avec une très forte probabilité une droite-image correspondant à une arête d'un paquet de la couche supérieure, en partant du bord de l'image et en suivant une direction centripète. Les cas défavorables paraissent tout d'abord poser un problème particulier, mais ils sont rapidement résolus grâce aux règles d'interprétation contextuelle. Dans tous les cas, au fur et à mesure de la découverte et de l'interprétation de nouveaux noeuds-image, on arrive rapidement à détecter et à isoler celles qui, malgré qu'elles suivent les directions privilégiées, ne doivent pas être poursuivies. Le bas de la figure B.13 montre toutes les configurations possibles dans lesquelles il est possible de découvrir un noeud-image de type "L", "T" ou "X".

4 L'interprétation en Indices de Scène

La connaissance très précise du contexte dans lequel le système évolue permet non seulement de ne retenir que les interprétations qui correspondent à des réalités scéniques, mais encore précisent les conditions contextuelles dans lesquelles elles sont applicables. La figure B.14 montre, pour un "contexte local" donné et un type de noeud donné, quelle est l'unique interprétation scénique correspondante au niveau INDICES DE SCENE *dans le seul plan des surfaces visibles de la couche supérieure*. Cette restriction est apportée pour ne disposer, à l'issue du processus d'interprétation des indices d'image, que des indices de scène correspondant aux objets recherchés. Par "contexte local" d'un noeud-image NI, nous entendons ici la connaissance de la nature des régions qui, par leur adjacence apparente, ont engendré la naissance de NI. Les différentes natures de région considérées sont 1/ le fond de la scène, 2/ un emballage de couche supérieure, 3/ une bande adhésive de couche supérieure, ou 4/ un emballage de couche inférieure : le trop faible contraste observable dans le cas d'une bande adhésive de couche inférieure nous permet de ne pas tenir compte de cette dernière nature de région. La figure B.14 montre le contexte local de chaque règle, représenté dans ses prémisses.

Lorsqu'une règle d'inférence est applicable sur un noeud-image, elle engendre

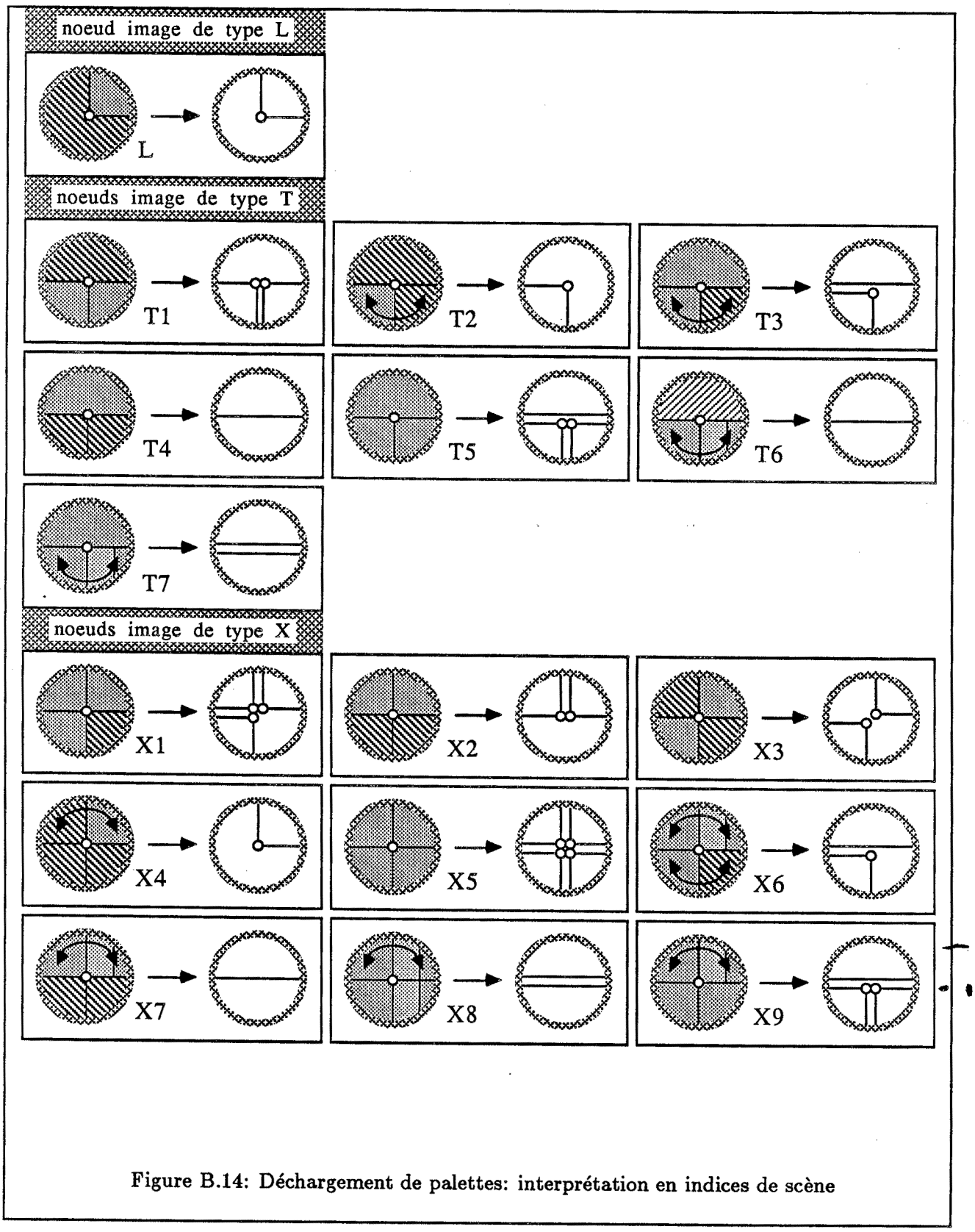


Figure B.14: Déchargement de palettes: interprétation en indices de scène

autant de sommets et d'arêtes qu'il est nécessaire, ainsi que nous l'avons décrit dans le cas général. Dans le cas où le contexte local d'un noeud-image n'est connu que partiellement, on conserve bien sûr toutes les interprétations possibles qui peuvent être inférées : cette réduction des interprétations, bien qu'elle ne soit pas totale, rend néanmoins possible l'application d'une propagation de contraintes partielle sur les noeuds-image environnants. A la stabilisation du processus d'inférence, si le contexte global a été découvert, il existe alors une seule interprétation possible de la structure d'indices d'image extraite. Dans les autres cas, plusieurs interprétations sont possibles.

5 Les résultats obtenus

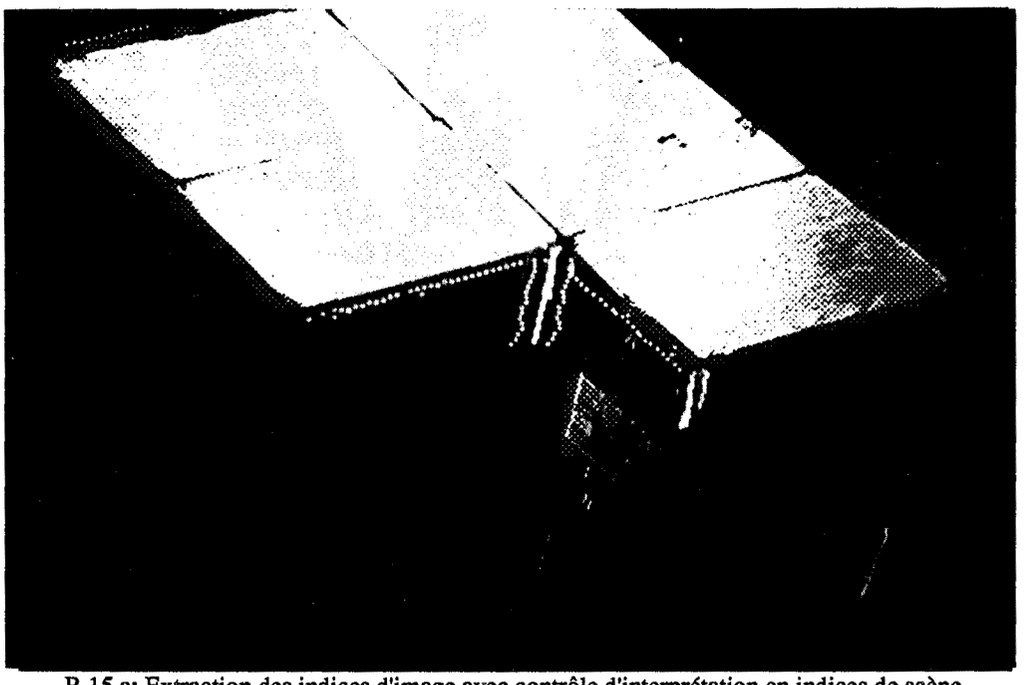
Les figures B.15.a et B.15.b montrent deux phases intermédiaires et la phase finale d'extraction des indices d'image en montrant explicitement une partie du contrôle global. Notons sur ces images, qu'à la limite, le détecteur de départs potentiels de droites-image à partir d'une droite-image poursuivie permet souvent à lui seul, (s'il est suffisamment sévère quant à la détection de contraste qu'il fournit) de s'affranchir de nombreuses règles d'interprétation : celles qui font intervenir le problème des bandes adhésives, ou même celles qui font intervenir le problème des droites-image correspondant à des indices d'image de surfaces des couches inférieures. La figure B.16.a montre le résultat final de l'extraction des indices d'image, alors que la figure B.16.b illustre le réseau des indices d'image extraits et retenus, ainsi que leur interprétation en indices de scène sous-jacente.

Dans le cas d'une mauvaise extraction des indices d'image, comme cela est le cas dans l'exemple étudié, il est nécessaire qu'un contrôle global décide de pousser plus loin l'analyse de certains noeuds-image pour détecter précisément les paquets recherchés. cela est possible puisque le système, grâce au schéma de palétisation dont il dispose, connaît le nombre de paquets à décharger et leur localisation approximative.

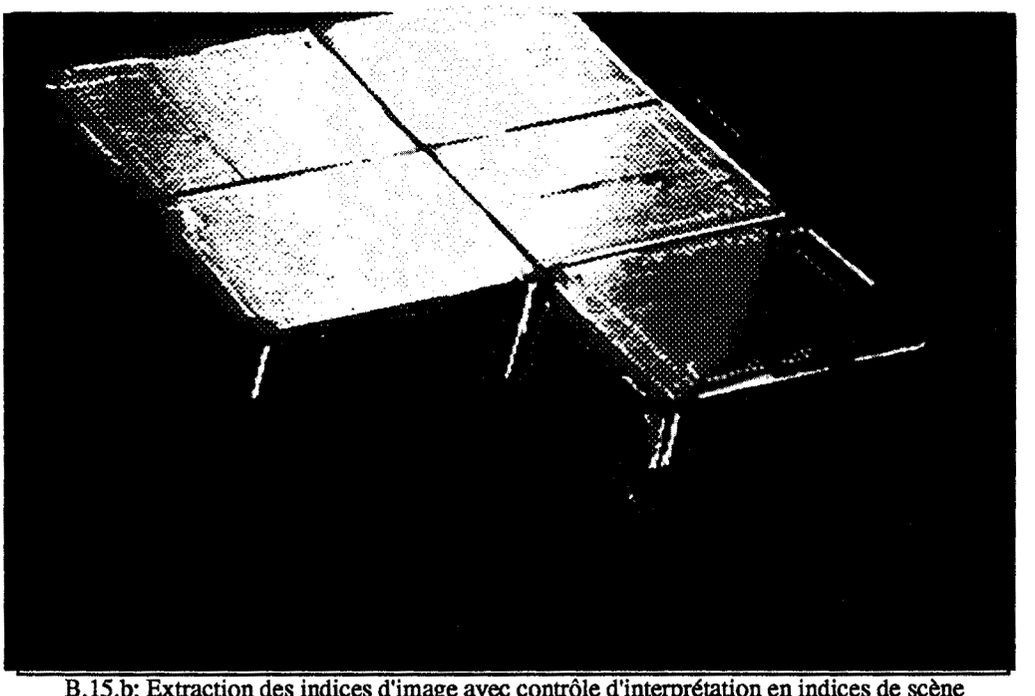
Notons finalement un avantage important à l'incrémentalité de notre approche : elle permet de découvrir au niveau INDICES DE SCENE ce qu'il est permis de considérer comme une surface d'un des paquets recherchés sans qu'il soit nécessaire que toute la structure au niveau INDICES D'IMAGE ait été extraite de l'image étudiée. Cette propriété est particulièrement intéressante d'un point de vue industriel, puisque la manipulation des premiers paquets reconnus peut donc s'effectuer en parallèle avec la fin du processus de compréhension de la scène observée.

Bien que l'étude effectuée ait démontré la faisabilité d'une réalisation en distinguant entre indices d'image et indices de scène, une restriction d'usage et une limitation sont cependant à porter au débit de notre approche.

1. La limitation est liée aux éventuels déplacements d'un ou plusieurs paquets. De la façon dont nous avons présenté ici l'algorithme d'interprétation de l'image en indices d'image, nous avons supposé que toutes les droites-image appartiennent à un même réseau connexe de lignes-brisées-image. Dans le cas

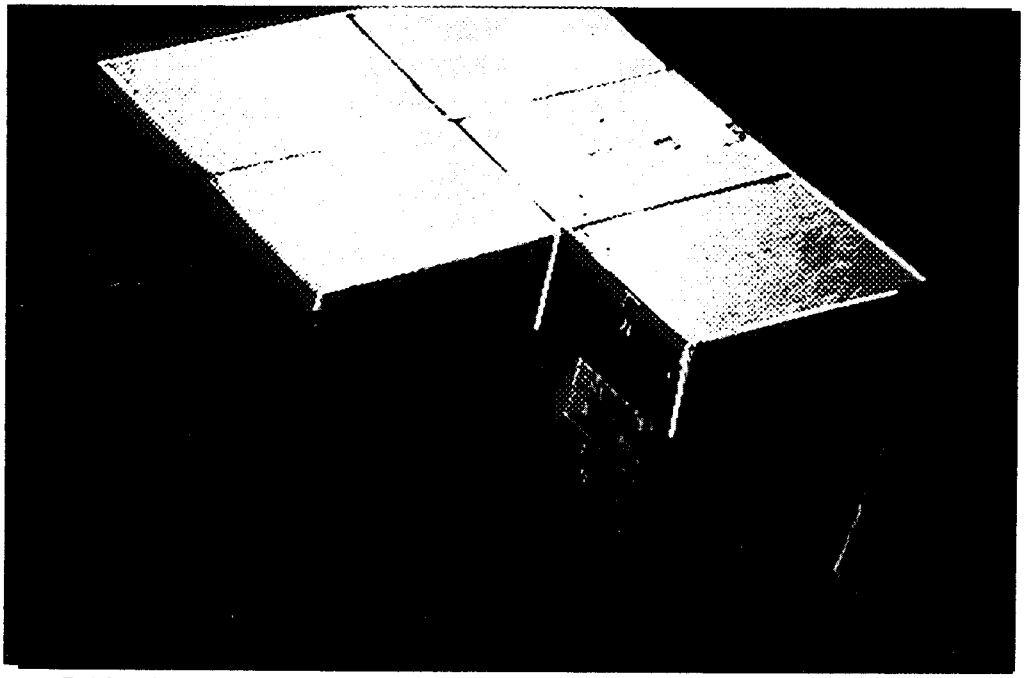


B.15.a: Extraction des indices d'image avec contrôle d'interprétation en indices de scène (phase intermédiaire)

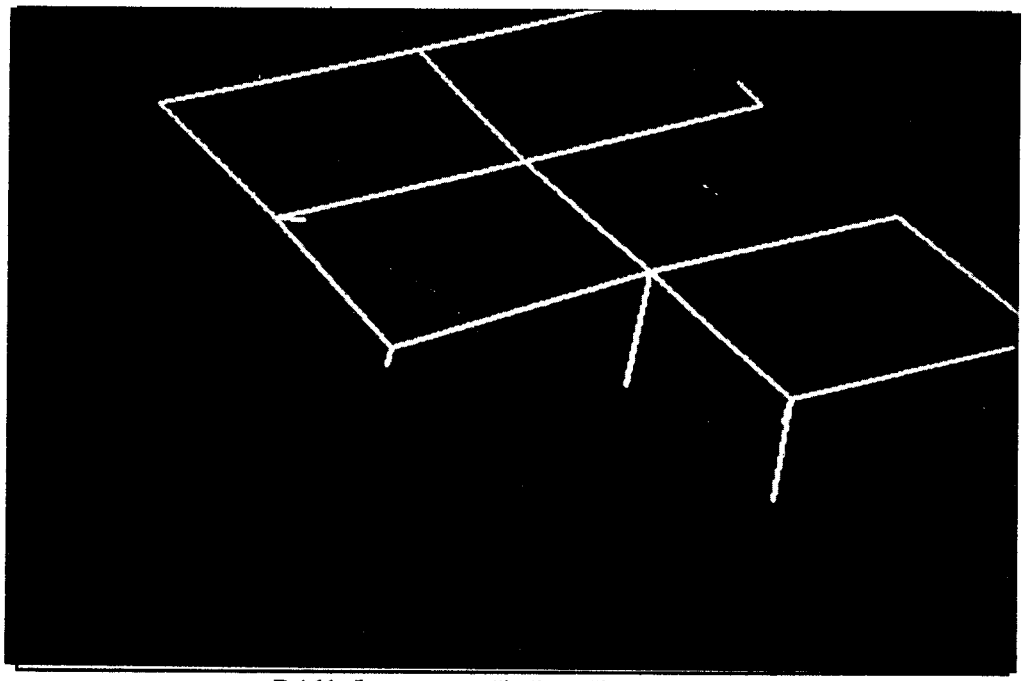


B.15.b: Extraction des indices d'image avec contrôle d'interprétation en indices de scène (phase intermédiaire)

Figure B.15: Déchargement de palettes: résultats expérimentaux (1)



B.16.a: Extraction des indices d'image avec contrôle d'interprétation en indices de scène (phase finale)



B.16.b: La structure d'indices d'image extraite et son interprétation sous-jacente en indices de scène

Figure B.16: Déchargement de palettes: résultats expérimentaux (2)

où l'observation de cas d'exception s'avère trop fréquente, il est nécessaire de ne pas oublier de paquets sur une couche avant de passer à la couche suivante. Tout comme dans l'exemple décrit par les figures B.11 et B.12, il suffit alors d'initialiser plusieurs fois le processus d'extraction d'indices d'image dans les régions inexplorées de l'image.

2. La restriction d'usage est que, d'après les indices que nous avons utilisés, les paquets ne sont "autorisés" qu'à se déplacer suivant les seuls axes du plan de palétisation de chaque couche. Cela reste malgré tout cohérent avec le fait que la topographie des surfaces visibles des paquets correspond théoriquement à un quadrillage assez solidaire et régulier, et que les déplacements d'ensemble ou locaux sont plus fréquents dans les directions du schéma de palétisation que dans les autres directions. Quand bien même ce ne serait pas le cas, nous avons vu dans le chapitre précédent qu'il était toujours possible de modifier l'ensemble des règles contextuelles (en tenant compte par exemple de nouveaux types de noeuds-image) sans modifier aucunement les procédures d'inférence et de contrôle.

Partie C

Inférence de Formes à partir de la Stéréoscopie Simple (images couleur - objets flexibles)

Percevant les branches enchevêtrées d'un arbre mort, je ne puis d'abord distinguer si la branche "a" est en avant ou en arrière de la branche "*b* " jusqu'au moment où, atteignant leur point d'intersection, je vois passer "a" sur "*b*" et "*b*" sous "a" : cette relation "a *sur b*" acquiert alors le rôle d'un indice dont l'utilisation me permet de structurer immédiatement l'ensemble des positions relatives des autres segments de "a" et de "*b*". Mais cet indice n'est ainsi qu'une partie ou un aspect de l'ensemble total constituant le signifié. Et il n'en consiste qu'une partie interchangeable, car j'aurais pu d'abord percevoir que "a" est plus proche de moi que "*b* ", ce qui m'aurait conduit à anticiper, en cherchant le point d'intersection, le passage de "a" sur "*b*" : en ce cas, l'évaluation globale des distances m'eût servi d'indice ou de signifiant perceptif et la position de "a" sur "*b*" en leur intersection eût été par exemple éclairée par cet indice, si elle avait été peu visible (devenant ainsi "signifiée").

Jean Piaget (1961)

Cette troisième partie décrit une nouvelle expérimentation des niveaux proposés. Elle a trait à l'inférence de formes à partir de la stéréovision simple, dans le domaine restreint des objets flexibles filiformes. Elle montre comment la notion des niveaux s'intègre pour cette nouvelle technique, mais aussi quels sont les apports engendrés par l'utilisation d'images couleur. Du niveau IMAGE au niveau OBJET, elle présente les différentes méthodes développées. Là aussi, cette partie s'achève sur la description d'une application industrielle : la localisation et l'identification de fils électriques, dans un contexte d'automatisation de la production d'ensembles câbles-connecteurs.

Chapitre C.I

Couleur, Stéréovision, et Objets Filiformes

1 Couleur et Niveaux de Représentation

La tradition en Vision par Ordinateur veut que l'on se concentre sur le traitement d'images noir-et-blanc. Nous voyons cependant un double intérêt à utiliser des images colorées. L'intérêt le plus immédiat est lié à la possibilité d'adjoindre des attributs chromatiques à tout indice d'un quelconque niveau de représentation. Le second avantage, qui découle de la plus grande richesse des informations initiales, se traduit par une meilleure qualité des - et une plus grande quantité en - indices qui peuvent être extraits, dès le niveau INDICES D'IMAGE. C'est la préférence accordée à tel ou tel de ces avantages qui détermine, selon nous, les deux approches informatiques de la couleur.

1.1 Les différentes approches informatiques de la Couleur

1.1.1 Images Couleur et Systèmes de Coordonnées

Une image couleur peut être acquise par digitalisation d'une scène observée par une caméra couleur d'un type quelconque. Si on ne dispose que d'une image noir-et-blanc, l'image couleur peut être obtenue soit à l'aide de filtres colorés (soustraction de lumière), soit à l'aide d'une illumination artificielle de la scène (addition de

lumière) avec les trois couleurs primaires R(ouge), V(ert), B(leu). Directement fournies par le capteur et mémorisées sous cette forme au niveau des calculateurs, les images RVB sont les images colorées les plus fréquemment rencontrées en Vision par Ordinateur. Disposant d'images RVB, où chaque pixel possède donc non plus une unique valeur d'intensité lumineuse mais trois, il est possible de transformer les valeurs RVB en d'autres "coordonnées colorées", par des opérations linéaires ou non. Ceci peut être vu comme la sélection d'un système de coordonnées et d'un ensemble d'axes de l'espace tridimensionnel de la couleur. Un certain nombre de référentiels couleur sont maintenant couramment utilisés. Dans tous les cas, le passage bijectif d'un système de coordonnées à un autre est une procédure d'enrichissement destinée à traduire certaines propriétés et à ***rendre explicites*** un certain nombre d'informations utiles à la segmentation des images couleur.

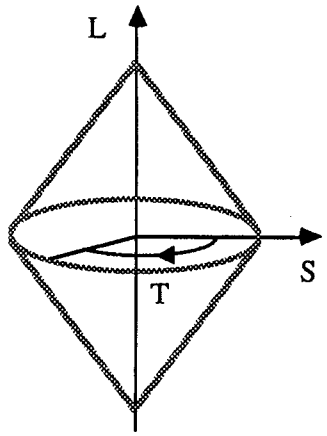
C'est ainsi que le système "TSL" (mis pour Teinte, Saturation, et Luminance) est celui qu'utilisent les médecins pour mesurer la perception humaine de la couleur. C'est aussi le repère dans lequel les informaticiens de la couleur travaillent le plus souvent[1](cf. figure C.01.a). La Teinte se réfère au nom de la couleur (par exemple bleu, rouge, orange). La Saturation traduit le degré de pureté, c'est-à-dire le grisé d'une couleur (par exemple, le rouge est une couleur saturée; le rose, par contre, est de la même teinte, mais est d'une saturation moindre). La Luminance mesure la luminosité d'une couleur, c'est-à-dire son degré de densité.

De nombreux autres systèmes de coordonnées existent. Le lecteur pourra se reporter à [JUDD-75], [OHTA-80] ou encore [BROWN-82] pour une présentation et une analyse détaillées de ces référentiels. Nous ne donnons ici que les noms de quelques-uns et certaines de leurs particularités :

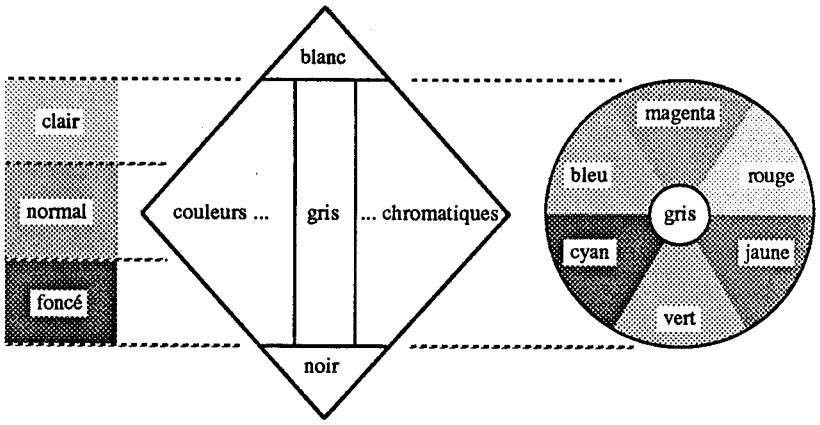
- le système "rvb" (système des couleurs normalisées qui décrivent les informations chromatiques indépendamment de la luminance),
- le système "YIQ" (système de coordonnées pour les signaux de télévision. La coordonnée Y, très proche de la luminance, est le seul signal visible sur un téléviseur noir-et-blanc. Les autres coordonnées, I et Q, sont des mesures chromatiques),
- le système "XYZ" (système de coordonnées colorimétriques, pour la mesure des couleurs de surfaces),
- le système "UVW" (système des couleurs uniformes, adopté par les médecins, et dans lequel une distance euclidienne correspond assez bien aux perceptions humaines).

[1] **système TSL** : système de coordonnées cylindriques dans l'espace des couleurs, où la luminance parcourt l'axe du cylindre, où la saturation est la distance radiale à l'axe du cylindre, et où la teinte est donnée par le déplacement angulaire à partir d'une certaine couleur standard. Une définition possible du système TSL est la suivante :

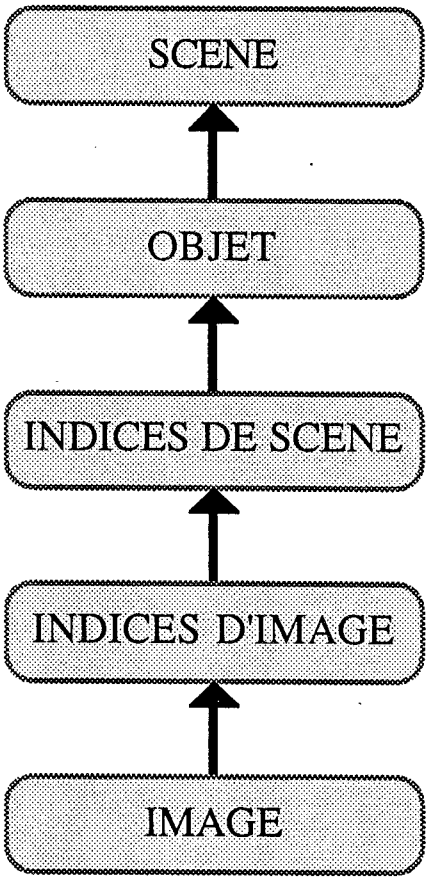
$$T = \arctan\left(\frac{\sqrt{3}(V-B)}{2R-V-B}\right) \qquad S = 1 - 3\frac{\min(R,V,B)}{R+V+B} \qquad L = \frac{R+B+V}{3}$$

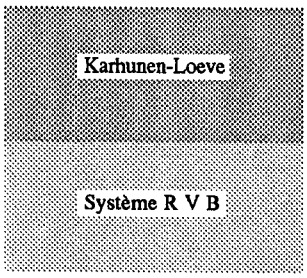


C.01.a: Le système de coordonnées Teinte - Saturation - Luminance

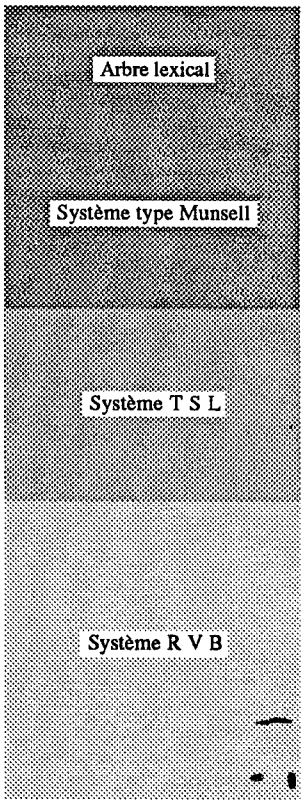


C.01.b: Le système de type Munsell utilisé





Pour l'extraction d'indices



Pour la qualification d'indices

C.01.c: L'utilisation des modèles selon les niveaux de représentation atteints

Figure C.01: Couleur et Niveaux de Représentation

Chaque combinaison hybride de trois coordonnées, dès qu'elle constitue une base dans l'espace des couleurs, est bien sûr possible et constitue un nouveau référentiel dans lequel il est possible de travailler. Les combinaisons sont multiples, mais finalement, les questions de savoir comment utiliser l'information chromatique et quels sont les systèmes de coordonnées les plus appropriés n'ont été qu'assez peu étudiées.

1.1.2 L'approche Noir-et-Blanc

L'approche noir-et-blanc de la couleur est liée à la *segmentation* (ou "inférence d'indices d'image"). Dès lors qu'un système de représentation de l'information de couleur a été déterminé, il est possible d'effectuer une segmentation des images couleur identique à la segmentation des images noir-et-blanc. Il suffit de prendre chaque nouvel axe de l'espace tridimensionnel des couleurs comme nouvelle base de la segmentation, tout comme l'échelle des niveaux de gris (i.e. l'intensité lumineuse) le sont dans le cas des images noir-et-blanc. Plus généralement, il s'agit de déterminer un scalaire $S(R,V,B)$ à partir duquel il est possible d'appliquer les algorithmes classiques de segmentation des images noir et blanc.

1.1.3 L'approche spécifique Couleur

Dans cette seconde approche, il s'agit de conserver simultanément les nouveaux axes de représentation de l'espace des couleurs. C'est le cas par exemple de l'algorithme de "seuillage dans un espace multidimensionnel" décrit dans [OHTA-80]. L'utilisation de plusieurs systèmes de coordonnées correspond à la volonté d'expliciter les coordonnées colorées les plus discriminantes possibles, en effectuant le minimum de calculs à partir des trois couleurs de base RVB. Notons que cette approche est souvent liée au besoin de *qualification* (par attribut de couleur) des indices de quelque niveau que ce soit. D'un point de vue de la représentation de ces attributs, un modèle de type Munsell, représenté par la figure C.01.b, est bien adapté à la dénomination de la couleur des objets en langage usuel.

1.2 Notre approche et l'adéquation de la Couleur aux Niveaux

Comme nous l'avons dit dès le début de ce paragraphe, nous distinguons deux problèmes distincts par rapport à leur insertion au sein des cinq niveaux de représentation que nous distinguons : l'extraction des indices et la qualifications d'indices (cf. figure C.01.c). Différents travaux de recherche menés par [GORDI-83] [GORDI-85], essentiellement fondés sur des études comparatives de différents systèmes de représentation pour la couleur, ont permis d'établir des résultats qui, outre l'analyse succinte du problème de la couleur telle que nous venons de la présenter, nous ont guidé dans notre approche.

1.2.1 Le point de vue de la Segmentation en Indices d'Image

Il est clair que l'utilité d'une coordonnée colorée est grandement influencée par la structure des scènes colorées à analyser. D'un point de vue de la segmentation, il

faut donc adapter les coordonnées colorées aux images à étudier. Par ailleurs, les coordonnées colorées à variance élevée sont les plus utiles à la segmentation d'une image de couleur. C'est ainsi que l'utilisation des coordonnées RVB et XYZ ne donne pas de très bons résultats pour la segmentation, car ces coordonnées possèdent un fort facteur d'intensité et sont fortement corrélées entre elles. Le système HLS donne de bien meilleurs résultats. **Mais c'est encore le système obtenu par la transformée de Karhunen-Loeve (détermination des axes principaux d'inertie de la matrice de covariance RVB) qui reste le meilleur choix, du fait de son adaptation maximale à l'image étudiée** [KENDE-76] (cf. partie gauche de la figure C.01.c).

1.2.2 Le point de vue de la Qualification des Indices

L'information chromatique n'est pas toujours importante pour le processus de segmentation, même dans le cas de scènes très colorées qui possèdent une variance étendue pour ses composantes chromatiques, et à condition, bien sûr, que l'information de luminance soit assez importante. Nous avons choisi, au niveau INDICES DE SCENE, de rendre l'information noir-et-blanc explicite. D'autre part, la teinte est indépendante de tout point de vue de l'observateur mais aussi de l'intensité de la source d'illumination ambiante. Pour ces deux raisons, **nous utilisons au niveau INDICES DE SCENE le système TSL plutôt que le système hérité RVB qui, lui, nous sert de support de représentation au niveau INDICES D'IMAGE. Au niveaux OBJET et SCENE, ce sont respectivement une représentation de type Munsell et un arbre lexical [GORDI-86] qui sont utilisés, dans le but de permettre une communication efficace avec le monde extérieur** (cf. partie droite de la figure C.01.c).

1.2.3 Discussion

Jusqu'à présent, les programmes de compréhension d'image ont utilisé différentes distances métriques aussi bien que différents espaces de couleur. Si la distance euclidienne reste la plus courante, d'autres, comme le maximum de la différence dans chaque coordonnée colorée, et la somme des différences dans toutes les coordonnées colorées, ont été appliquées avec succès. Nous verrons, pour le problème de la stéréovision tel que nous l'avons abordé, comment la différence dans chaque coordonnée T, S et L nous permet de restreindre la combinatoire du processus de mise en correspondance entre indices de scène.

Notons finalement que dans certains cas, il n'existe pas de différences significatives suivant l'ensemble des coordonnées utilisées. Cela semble être dû au fait que l'information de couleur pour les scènes naturelles est essentiellement bidimensionnelle (la coordonnée de luminance plus une coordonnée chromatique) : cela peut être vérifié expérimentalement par la synthèse d'une image couleur à partir de seulement deux coordonnées, comme par exemple I_1 et I_2, deux des axes d'inertie moyens découverts par Ohta, et qui, pris dans cet ordre, véhiculent respectivement 80% et 17% de l'information de couleur. Cela sera aussi tout à fait notable dans

les résultats de notre expérimentation : la coordonnée S, tout comme I_3, ne nous permet que peu de mises en correspondance effectives entre indices de scène[2].

2 Stéréovision Simple et Niveaux de Représentation

Bien qu'étudiée en informatique depuis très longtemps, au contraire de la couleur, la stéréovision simple n'a toujours été étudiée qu'implicitement par rapport au problème des niveaux de représentation de l'information binoculaire. Pourquoi? Il nous semble que la raison principale en est l'approche même du problème de la stéréovision, c'est-à-dire son découpage en quatre sous-problèmes. A l'opposé, nous distinguons deux problèmes à résoudre, l'un (la "localisation") par rapport à la décentration, l'autre (l'"identification") par rapport à l'abstraction.

2.1 Les problèmes soulevés

Contrairement à la couleur, nous ne reviendrons sur le détail des principes de la stéréovision, puisque nous nous y sommes déjà attardés dès le début de ce manuscrit. Le lecteur, là encore, pourra se référer aux nombreux travaux effectués dans le domaine, et en particulier à [FRISB-79] [MARR-79] [GRIMS-81] [BRADY-82] ou [TSUKI-85]. Rappelons simplement que les quatre phases qui constituent le problème sont les suivantes :

- recherche des indices intéressants dans les deux images,
- mise en correspondance entre ces indices,
- calibrage du montage stéréoscopique,
- localisation des objets dans l'univers réel

Les niveaux de représentation présents dans les systèmes de vision qui suivent ce schéma de décomposition ne font que refléter implicitement l'ordonnancement dans le temps suivant lequel les différents problèmes sont résolus. **Pourtant, l'existence de ces niveaux peut être justifiée : les deux premiers problèmes ont trait à des abstractions SUR l'objet : il s'agit donc, d'après nos définitions, d'un problème d'"abstraction". De même, les deux derniers problèmes relèvent beaucoup plus d'une abstraction des objets PAR RAPPORT à l'espace, et donc, toujours d'après nos définitions, plutôt d'une "décentration".** Dans les deux cas, il apparaît bien que la stéréovision puisse être directement reliée aux niveaux de représentation de façon explicite.

[2]**système** $I_1I_2I_3$: système de coordonnées déterminé expérimentalement par Ohta à partir de statistiques sur la transformée de Karhunen-Loeve :

$$I_1 = \frac{R+V+B}{3} \qquad I_2 = \frac{R-B}{2} ou \frac{B-R}{2} \qquad I_3 = \frac{2V-R-B}{4}$$

La qualité et la nature des indices à mettre en correspondance dépend énormément de la nature de l'observateur, et plus précisément de la géométrie du capteur stéré-oscopique utilisé. C'est pourquoi, avant de nous attacher aux problèmes d'abstraction et donc de déterminer quels seront ces indices, nous nous attachons tout d'abord à résoudre les problèmes de décentration.

2.2 Notre approche générale de la Stéréovision Simple

Avant toute chose, il nous apparaît important de souligner que *la stéréovision ne sert pas à localiser précisément les objets d'une scène quelconque.* Il est souvent reproché à la stéréovision d'être imprécise. Cela n'est vrai qu'à partir du moment où c'est un calcul de "disparité" (notion géométrique que nous avons déjà introduite et qui correspond effectivement chez l'homme à la notion de disparité rétinienne) qui permet la localisation des objets de la scène observée. En effet, les algorithmes qui sont alors appliqués correspondent chez l'homme à une "vision de loin" pour une perception imprécise du relief.

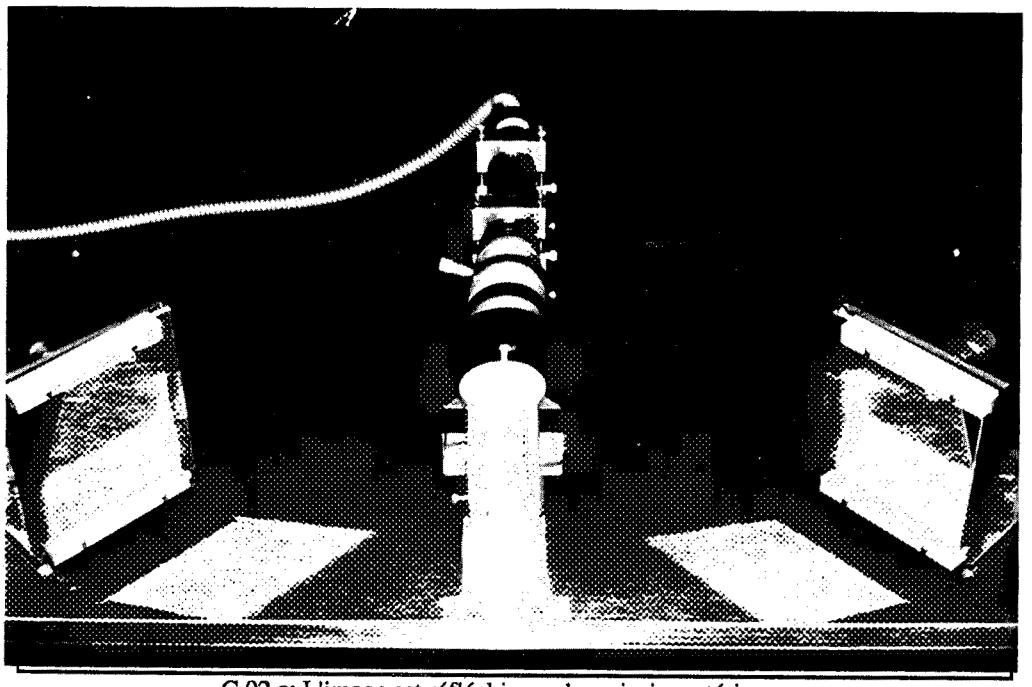
A partir de quelle distance avons-nous surtout besoin d'une très bonne précision? Chez l'homme, cela correspond à des besoins précis de coopération bras-oeil. Nous pensons qu'il en est de même pour la Vision par Ordinateur, c'est-à-dire que **le but essentiel de la stéréovision est d'acquérir une information précise dans un espace restreint et proche du capteur.** Cette notion n'est pas nouvelle, et nous la retrouvons déjà chez [SHIRA-73] où l'"oeil" est utilisé pour contrôler et préciser les mouvements d'un bras manipulateur. On retrouve cette même idée plus récemment chez [INOUE-84], pour un traitement assez identique à celui que nous effectuons pour la manipulation de fils électriques, sujet de notre seconde application à un problème industriel.

En conséquence de tout ce qui précède, **nous nous plaçons donc dans le cas d'une stéréovision dite "stéréovision large"**, où l'angle rapporté entre les deux axes focaux des caméras (réelles ou virtuelles) est un angle "important". Nous plaçant en outre **dans un contexte de la stéréovision simple**, nous connaissons la géométrie du capteur stéréoscopique avant toute analyse d'image : c'est pourquoi. nous pourrons aussi utiliser la contrainte de la ligne épipolaire au cours de notre analyse.

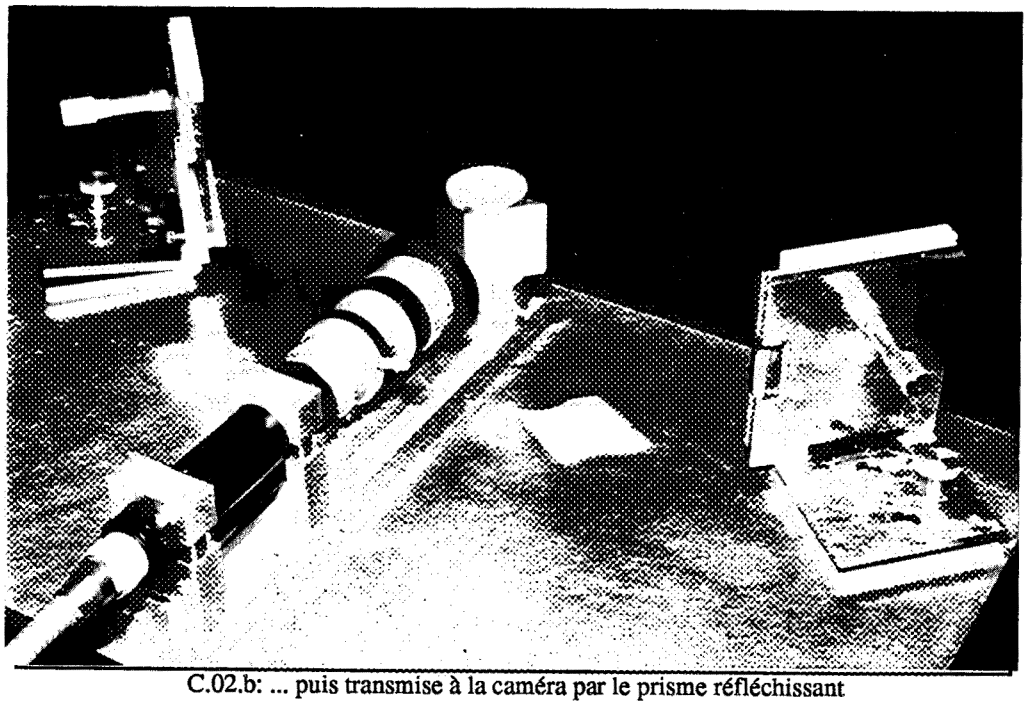
2.3 Notre approche du Calibrage et de la Localisation

2.3.1 Description du Capteur Stéréoscopique

Pour étudier le problème de l'inférence de formes à partir de la stéréovision simple, nous avons été amenés, dans un premier temps, à concevoir un capteur adéquat à la stéréovision large, et donc capable de fournir des "paires" stéréoscopiques présentant une très forte disparité (cf. figure C.02). Le capteur que nous avons réalisé [DEMAZ-85b] est cependant à rapprocher du modèle du système binoculaire humain proposé par Longuet-Higgins [LONGU-82], dans le sens où les deux plans médians horizontaux des deux yeux sont confondus.



C.02.a: L'image est réfléchie par les miroirs extérieurs ...



C.02.b: ... puis transmise à la caméra par le prisme réfléchissant

Figure C.02: Notre capteur stéréoscopique

1. Le capteur stéréoscopique est composé d'un prisme réfléchissant et de deux miroirs rotatifs attachés sur le même support, et sur lequel une caméra peut être fixée. Dans cet assemblage essentiellement usiné en dural, le prisme est fixé dans l'alignement de l'axe optique de la caméra, axe avec lequel les deux faces réfléchissantes forment un angle de 45^o. Les deux miroirs sont à égale distance du prisme, et peuvent être mus en rotation à l'aide de vis micrométriques.

2. L'assemblage ainsi constitué (prisme, miroir et caméra unique) simule deux caméras coplanaires aux mêmes paramètres internes, et dont les axes optiques forment un angle Θ qui peut prendre des valeurs entre 0^o et 90^o, selon la position des miroirs. Un tel assemblage permet de fournir simultanément deux images (une paire stéréoscopique) du même objet de deux points de vue différents, de sorte que la partie gauche (resp. droite) de l'image vidéo contient l'image de la caméra virtuelle gauche (resp. droite).

3. Une propriété intrinsèque intéressante du capteur réside dans sa prise en compte de l'horizontalité de la contrainte de la ligne épipolaire. De sorte que, sous de "bonnes" conditions d'utilisation, et si un point (X, Y, Z) de l'espace réel est visible dans les deux images aux pixels $P_g(I_g, J_g)$ et $P_d(I_d, J_d)$, alors par construction les deux coordonnées J_g et J_d sont égales à une même valeur, soit J.

2.3.2 Comparaison avec d'autres capteurs

Il est clair que le dessin de notre capteur ne présente pas de caractéristique révolutionnaire. Mais tout comme on n'écrit pas de la même manière avec un stylo à plume et avec un stylo à bille, l'une des qualités propres à notre capteur réside dans sa très grande souplesse d'utilisation. De même, les simplifications déduites du modèle de calibrage que nous verrons à la fin de cette partie, permettent d'envisager une coopération bras-oeil précise, rapide et efficace. Dans ces directions, le capteur conçu apparaît comme particulier.

De fait, notre capteur est conceptuellement différent de celui de Teoh et Zhang [TEOH-84], dont l'angle Θ est fixé à 0^o et où le prisme est remplacé par un miroir rotatif, de sorte qu'une caméra unique puisse acquérir en séquence les deux images de la paire stéréoscopique.

La première de ces contraintes est en désaccord avec l'opinion qui est la notre et selon laquelle le but essentiel de la stéréovision est d'obtenir une information dans l'optique d'une coopération bras-oeil précise. C'est la raison pour laquelle nous utilisons généralement notre capteur avec des valeurs de Θ comprises entre 60^o et 80^o.

La seconde contrainte du capteur de Teoh et Zhang interdit les études sur le mouvement, dans ce sens bien précis où il ne permet pas d'acquérir deux images simultanées d'un objet en mouvement. Dans ce sens, la conception de notre capteur est plutôt à rapprocher de celle du capteur de Geschke [GESCH-79]. Le dessin de ce dernier, s'il substitue deux miroirs fixes au prisme de notre capteur (montages

équivalents), préserve la qualité de simultanéité de l'acquisition des deux images. Là encore, cette qualité nous paraît essentielle dans le cadre d'une coopération bras-oeil.

2.3.3 Le Calibrage et la Localisation

Disposant de l'outil nécessaire à une stéréovision large, il s'agit maintenant de se donner les moyens de résoudre les problèmes de décentration. Le calibrage d'un capteur de vision consiste à établir les relations analytiques entre "les coordonnées de l'espace réel" et "les coordonnées de l'image". Le "calibrage direct" fournit les coordonnées image en fonction des coordonnées réelles. Le "calibrage inverse", comme son nom l'indique, fournit ... l'inverse.

Prenons tout d'abord le cas d'une caméra unique : une première approche du problème est de déterminer un modèle géométrique général d'une caméra qui tienne compte de ses paramètres internes et externes, et qui tienne aussi compte du phénomène de perspective [DUDA-73] [GENNE-79]. Une fois ce modèle géométrique découvert, on en déduit immédiatement le modèle analytique de calibrage qui lui est associé. Nous montrons, au cours de l'exposé de nos méthodes, que la détermination du modèle analytique est assez aisé. Les équations de calibrage, fournissant les coordonnées (I, J) d'un point de l'image en fonction de (X, Y, Z), coordonnées réelles du point observé, sont de la forme :

$$I = \frac{a_{01}X + a_{02}Y + a_{03}Z + a_{04}}{a_{09}X + a_{10}Y + a_{11}Z + 1}$$

$$J = \frac{a_{05}X + a_{06}Y + a_{07}Z + a_{08}}{a_{09}X + a_{10}Y + a_{11}Z + 1}$$

les "a_i" étant exprimés en fonction des paramètres de la caméra [DUDA-73].

Bien qu'il soit facile en principe de mesurer les différents paramètres de la caméra (sa longueur focale par exemple), en pratique, on se retourne vers une seconde approche dans laquelle c'est la caméra elle-même qui joue le rôle d'appareil de mesure [BORIA-84]. Cette opération consiste à observer des points dans l'image, points dont les coordonnées dans l'espace réel sont parfaitement connues du système (appelés "points de contrôle"). Chaque observation fournit deux équations linéaires, d'inconnues les "a_i". Il suffit donc de cinq correspondances et demi, et d'un algorithme de résolution de système linéaires (du genre "Gauss" ou "Moindres Carrés ", selon le nombre de correspondances retenues), pour déterminer parfaitement le modèle de calibrage direct de la caméra. Notons aussi qu'il est impossible, sauf connaissances supplémentaires, de déterminer un calibrage inverse dans l'espace tridimensionnel pour une caméra unique : le problème du calibrage inverse est sous-déterminé.

Dans le cas très particulier de la stéréovision simple, nous connaissons la géométrie du montage stéréoscopique. Cette aide précieuse, au sens où elle permet de tenir compte de la contrainte de la ligne épipolaire, conduit donc souvent à calculer le modèle de calibrage des caméras avant tout autre traitement [PRAZD-82]. Un moyen simple de réaliser cet objectif est de calculer le modèle de calibrage direct stéréoscopique comme la conjonction de ceux de chacune des deux caméras (virtuelles ou réelles selon le montage) avec les formules que nous venons de donner. En conséquence, le problème du calibrage inverse est alors surdéterminé.

Nous verrons cependant que dans notre cas particulier de stéréovision large, nous avons réussi à simplifier le calibrage, grâce à la connaissance de la géométrie du montage que nous avons dessiné, et grâce à l'utilisation que nous avons fait de ce capteur. Ce faisant, les modèles de calibrage que nous avons déterminé sont involutifs.

2.4 Notre approche des Indices et de l'Identification

2.4.1 Le processus de mise en correspondance

La mise en correspondance est un processus de recherche. A ce titre, elle est composée de deux éléments : une mesure de différence et une stratégie de recherche.

Mesure de Différence Chaque type de mesure de différence est appropriée selon le type d'indices d'image à mettre en correspondance. La qualité majeure recherchée pour une mesure est son insensibilité aux changements de contraste et de luminosité entre les images. Ces différences peuvent apparaître lorsque les deux vues du même objet sont acquises différemment, lorsque les images sont prises séparément sous des conditions de lumière différentes, ou encore parce que la luminosité apparente d'une surface change avec l'angle selon lequel elle est vue. Au niveau INDICES D'IMAGE, la "corrélation normalisée" est une mesure de la similarité entre régions qui est insensible au contraste et à la luminosité. Les droites-image (indices linéaires) sont similaires si elles ont la même orientation et si le gradient d'intensité traversant ces droites est le même, mais deux arêtes mises en correspondance peuvent présenter des régions adjacentes d'apparence dissemblable.

Stratégies de Recherche Trouver des indices de correspondance entre deux images est le problème le plus important de l'interprétation des images stéréoscopiques. Même le meilleur des processus de détection d'indices sélectionnera encore des indices d'une image qui peuvent être mis en correspondance avec plus d'un indice de l'autre image. Quelques remarques viennent cependant spécifier le problème de mise en correspondance et en donner les limites, sans toutefois le résoudre :

- La mise en correspondance est un problème fortement combinatoire.
- Deux propriétés de la matière, la cohésion et l'opacité, peuvent aider à la résolution des mises en correspondance ambigues. Grâce à la cohésion de la matière, les distances de la caméra aux objets qui apparaissent proches les uns

des autres dans l'image tendent à être les mêmes. A cause de la coutumière opacité de la matière, chaque point de l'image est associé à une mesure de profondeur unique.

- Nous savons que, connaissant les paramètres de la caméra, la mise en correspondance peut être facilitée en tirant avantage du fait que la correspondance se situe uniquement le long des lignes épipolaires. Cela est spécialement important pour les modèles informatiques de la vision stéréoscopique humaine, dans lesquels la recherche est limitée le long de lignes horizontales pour une même position verticale dans les deux images. A cette contrainte de la ligne épipolaire s'ajoutent deux heuristiques souvent utilisées : a/ des indices proches dans une image s'apparient avec des indices proches dans l'autre image, b/ l'ordre des indices mis en correspondance est préservé le long des lignes épipolaires.

- A cause des occultations possibles entre objets d'une même scène, certains indices d'une image ne sont pas visibles dans l'autre; il est donc impossible de trouver une mise en correspondance entre ces indices.

- Si la scène observée comporte la répétition d'un même motif (par exemple les fenêtres d'un immeuble), la mise en correspondance est inévitablement ambigue.

2.4.2 Les Indices à rechercher et l'Identification

Nous avons vu comment calibrer un système de vision stéréoscopique et ainsi assurer la décentration nécessaire à l'expression des coordonnées réelles d'un point (ou plus généralement d'un indice) observé dans les deux images. Il nous faut maintenant répondre à une autre question : sachant les qualités requises par un processus de mise en correspondance, quelle est la nature des indices qui doivent être mis en correspondance? A cette question, [FRISB-79] répond qu'il est possible d'utiliser des indices d'un niveau quelconque. Nous pensons nous aussi qu'il n'y a aucune impossibilité à faire cette correspondance entre indices d'un niveau quelconque, à partir du moment où l'un des objectifs suivis est de trouver des indices d'une image qui peuvent être mis en correspondance dans l'autre image de façon non ambigue. Il suffit, pour s'en convaincre d'observer les travaux antérieurement réalisés (cf. figure C.03).

Au niveau IMAGE Ne pouvant détecter aucun indice monoculaire dans un stéréogramme aléatoire de points ou de droites, les travaux respectifs de [GRIMS-81] et de [NINIO-81] correspondent à une mise en correspondance au niveau IMAGE. Bien sûr, la correspondance ne s'y fait pas au niveau pixel, car le fait que deux pixels de deux images possèdent les mêmes niveaux de gris ne préjuge que peu du fait qu'ils puissent être les reflets du même objet de la scène. C'est ainsi que la proposition de Marr est fondée sur l'observation des "passages à zéro" dans l'image Laplacienne, après un filtrage Gaussien passe-bas. Les "passages à zéro" sont les

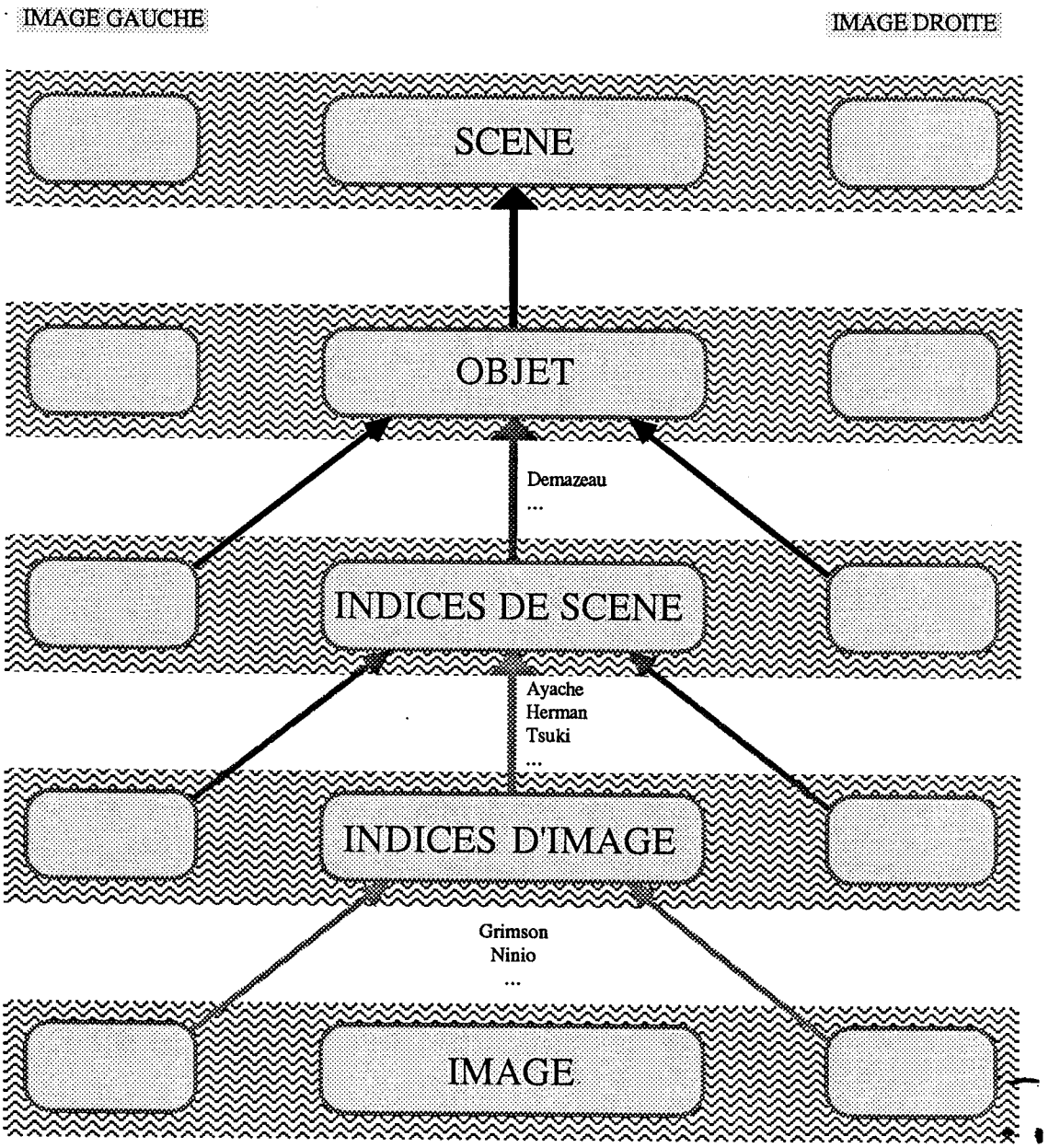


Figure C.03: Stéréovision et Niveaux de Représentation

points d'une image auxquels les dérivées secondes changent de signe [MARR-79]. Thorpe, quant à lui, s'intéresse à des points de contraste [THORP-83b].

Au niveau INDICES D'IMAGE Vis-à-vis de la combinatoire, il est donc plus sage de considérer des indices qui véhiculent une grande information, ou une forte variance. Une seconde approche consiste à considérer les régions et les lignes, et donc à travailler au niveau INDICES D'IMAGE. C'est d'ailleurs à ce niveau que se concentrent la plus grande partie des travaux actuellement effectués en stéréovision, comme le montrent par exemple les travaux de [AYACH-85] ou de [HERMA-84]. D'autres travaux, analysant l'évolution probable d'ensembles d'étiquettes d'une image à l'autre, comme ceux de [THORP-83a] ou de [TSUKI-85] relèvent aussi du niveau INDICES D'IMAGE.

Au niveau INDICES DE SCENE Il reste qu'un processus de mise en correspondance fondé sur un calcul de disparité, et donc nécessairement appliqué à des caractéristiques des niveaux IMAGE ou INDICES D'IMAGE, devient d'autant plus inefficace que l'on se rapproche plus d'une stéréovision dite large. Quels sont, pour de telles configurations, les indices à extraire? Selon [BRADY-82], les problèmes de la mauvaise précision et de la forte combinatoire du processus de mise en correspondance sont liés au fait qu'à une très forte disparité entre les images devrait correspondre des indices à mettre en correspondance qui relèvent de niveaux d'abstraction de plus en plus élevés. C'est dans cet esprit, et surtout bénéficiant des résultats de notre première expérimentation, que nous avons été amenés à effectuer une mise en correspondance au niveau INDICES DE SCENE [DEMAZ-85a]. C'est-à-dire que, pour les indices que nous mettons en correspondance, et outre le lieu dans l'image où ces indices peuvent se trouver, nous connaissons aussi des relations en adjacence et/ou en profondeur relative. Disposant toujours de la parfaite connaissance de la géométrie du capteur stéréoscopique, il est clair que de tels indices, par leur richesse, peuvent par exemple guider le processus de mise en correspondance en effectuant des prédictions qui vont bien au-delà des effets de la traditionnelle contrainte de la ligne épipolaire. Nous reviendrons, lors de l'exposé des méthodes que nous avons développées, sur la nature de la "mesure de différence" et sur celle du "processus de recherche" que nous avons utilisés.

3 Les objets gauches, flexibles et filiformes

Pour nous évader du domaine des polyèdres, pour lesquels le passage du niveau INDICES DE SCENE au niveau OBJET est "trop simple" du fait que chaque surface observée est une surface que l'on sait être plane, mais aussi parce que nous n'étions pas encore parvenus au niveau OBJET par notre première expérimentation, nous nous sommes tournés vers des objets gauches et flexibles.

3.1 Généralités sur les objets gauches et flexibles

Passer d'un monde uniquement constitué de faces planes à un monde composées de surfaces courbes dont on n'observe que des projections, n'est pas le fait d'un prolongement immédiat. Nous n'avons malheureusement pas pu étudier ces objets en dehors de la sous-classe des objets filiformes. C'est pourquoi nous n'avons que quelques remarques générales à formuler ici, et qui peuvent être regroupées sous une seule observation : à tous les niveaux de représentation, l'introduction des objets gauches et flexibles engendre des problèmes de natures diverses mais tous aussi difficiles à résoudre.

Au niveau INDICE D'IMAGE Un tout premier intérêt pour les objets gauches se retrouve chez [HUFFM-71], pour une tentative d'étiquetage de ces objets. La technique de l'étiquetage est déjà très discutée pour des scènes du monde des blocs, à cause de sa tendance à l'explosion combinatoire. Cette méfiance est exacerbée lorsqu'il s'agit d'objets gauches, car l'analyse de ces objets introduit des types de noeuds supplémentaires sans en supprimer. Le problème de l'étiquetage semble actuellement être dans une impasse, et ceci malgré les travaux importants de [CHAKR-82] et plus récemment de [MALIK-85]. Un autre problème non immédiat est lié au fait que pour les objets gauches, les indices d'images linéaires à extraire ne sont plus des lignes droites, mais des lignes qui peuvent être courbes ("lignes-image "), et pour lesquelles l'état de l'art des processus de segmentation ne propose que peu de solutions.

Au niveau INDICES DE SCENE Une ligne-image observée ne correspond plus nécessairement à une arête physique du monde réel, mais à une simple projection d'une surface tangente à la ligne de vue de l'observateur! La présence de ce phénomène engendre donc de nouveaux indices de scène qui sont des pseudo-arêtes, avec une signification identique à celle des pseudo-sommets que nous avions déjà rencontré dans le monde des blocs. Finalement, par rapport aux opérateurs d'interprétation que nous avons introduits pour l'étude de scènes composées d'objets du monde des blocs, la description de scènes physiques du monde réel nécessite, au minimum, l'élaboration de règles contextuelles supplémentaires pour l'interprétation de noeuds-image gauches (intersections de plusieurs lignes-images qui ne sont pas toutes des droites). Nous pensons malgré tout que le problème de ces "extensions " est de complexité moindre que la création de jeux d'étiquetage gauches, tout à la fois corrects et cohérents.

Au niveau OBJET Flexibles, les objets à inférer ne bénéficient pas d'une définition rigide. Par ailleurs, et indépendamment de la flexibilité, les surfaces extraites étant gauches, elles ne sont pas parfaitement déterminées par la simple connaissance de leur contour. De façon générale, des méthodes d'inférence de formes comme la réflectance ou la texture sont bien mieux adaptées à l'analyse des objets gauches que la méthode d'inférence de formes à partir des contours. C'est pourtant cette dernière que nous utilisons conjointement avec la stéréovision. Mais c'est aussi une

des raisons pour lesquelles nous nous sommes restreints à une sous-classe des objets gauches et flexibles : celles des objets filiformes.

3.2 Notre approche : le cas des Objets Filiformes

3.2.1 Définition des Objets Filiformes

Il n'existe pas de définition universellement reconnue de la qualité de filiformité : elle peut par exemple être définie à partir d'un opérateur de contraste qui squelettise les objets, ou bien encore par la détection de deux bords parallèles. De notre côté, nous définissons empiriquement les objets filiformes comme une classe d'objets possédant les caractéristiques communes suivantes :

- Chacun de ces objets est assimilable à - et représentable par - un unique cylindre généralisé[3].
- A tout instant, le point de la génératrice est un point de symétrie centrale de la section de balayage.
- le rapport de la surface moyenne de la section de balayage sur le carré de la longueur de la génératrice est très faible. Si ce rapport tend vers 0, on préfèrera alors parler d'objets "squelettiques".
- l'apparence de la section de balayage, soit la section de balayage *observée dans l'image* est un indice d'image de taille "relativement" faible par rapport à la taille de l'image totale. Cette dernière partie de la définition prévient du fait que de nombreux objets dits "filiformes" ne le sont que suivant l'échelle selon laquelle ils figurent dans l'image. Ceci est typique par exemple des poteaux télégraphiques suivant la distance à laquelle ils sont observés.

Les trois premières parties de cette définition concernent la géométrie des objets, dans un repère qui leur est propre, (informations de niveau OBJET), à l'opposé de la dernière qui fait référence au repère de l'image (information de niveau INDICES D'IMAGE). L'hétérogénéité de cette définition, qui ne se cantonne donc pas à une simple caractérisation relative au *modèle géométrique* de l'objet mais qui fait aussi appel à *l'apparence* de cet objet dans l'image, correspond de fait aux différents niveaux auxquels il est possible de faire intervenir les informations contextuelles que le système de vision peut posséder sur la scène qu'il étudie.

[3] **cylindres généralisés** : comme leur nom le suggère, les cylindres généralisés constituent une classe d'objets obtenus par extension de la définition d'un cylindre [AGIN-76]. Un cylindre ordinaire est obtenu par un balayage d'un *disque* le long d'un segment de droite passant par son centre. Le disque est maintenu perpendiculaire au segment de droite, qui est l'*axe* du cylindre. Le cylindre peut être "généralisé" par une ou plusieurs extensions : 1/ l'axe peut être courbe, 2/ le rayon du disque peut varier en fonction de sa position le long de l'axe (la fonction curviligne traduisant cette évolution est appelée "règle de balayage") 3/ la section déplacée peut être une figure plane quelconque plutôt qu'un disque 4/ la section déplacée peut être maintenue à un angle quelconque, non orthogonalement à l'axe.

3.2.2 Apport au Niveau Indices d'Image

Les indices d'image linéaires correspondant un objet filiforme sont deux lignes-image qu'il est possible d'extraire incrémentalement et en parallèle (notion de "suivi double "), tout en vérifiant à chaque instant deux propriétés : 1/ la distance entre les points extraits simultanément reste faible par rapport à la longueur des courbes extraites et par rapport à la taille globale de l'image 2/ les deux courbes extraites sont symétriques par rapport à leur axe médian.

3.2.3 Apport au Niveau Indices de Scène

Tous les objets filiformes ne sont constitués que d'un cylindre généralisé. Les interprétations des noeuds-image, si elles ne correspondent pas à une des extrémités réelles des objets observés, correspondent donc à des pseudo-sommets (cf. partie B) engendrés par des occultations entre objets filiformes. De la même façon, chacune des deux lignes-image issue d'un suivi double au niveau INDICES D'IMAGE dit être interprété comme une pseudo-arête, représentant simplement l'apparence d'une surface tangente à la ligne de vue de l'observateur.

3.2.4 Apport au Niveau Objet

La "rétroprojection" dans l'espace tridimensionnel d'un objet filiforme est immédiate. En effet, chaque point de la génératrice étant un point de symétrie de la section de balayage, la projection sur l'image de la génératrice est exactement l'axe médian des deux pseudo-arêtes définissant le contour apparent de l'objet filiforme. Inversement, une extrapolation cylindrique fondée sur les pseudo-arêtes définit parfaitement la surface réelle du cylindre généralisé observé.

Chapitre C.II

Analyse des Systèmes existant

Les objets filiformes constituent une classe d'objets fréquemment présente dans les images du monde réel. Par exemple, dans des scènes d'extérieur, les branches des arbres ou encore les poteaux télégraphiques sont des éléments de cette catégorie. De même, les images aériennes ou les images satellites contiennent des objets filiformes tels les routes ou les rivières [NAGAO-80], voire certains ponts [HAVEN-83]. Dans le domaine de la médecine, les radiographies du corps humain offrent à l'observation des vaisseaux et des artères. Finalement, dans le domaine industriel, la classe des objets filiformes est essentiellement représentée par les câbles et fils électriques. Tous ces domaines, auxquels est lié l'avenir du traitement d'images et de la vision par ordinateur, appellent de plus en plus à une toujours meilleure compréhension des scènes qu'ils étudient.

La compréhension de scènes composées d'objets filiformes est un important domaine d'étude qui est exploité depuis maintenant plus de cinq ans, et a donné lieu à de nombreuses réalisations. Souvent simples et astucieuses, les méthodes suivies ne nécessitent pas de grands moyens matériels (parfois juste une caméra noir-et-blanc), et traduisent par là-même leur vocation de résoudre un problème réel. Dans ce sens, le manque de généralité dont ces systèmes pourraient être taxés n'est souvent imputable qu'aux faits qu'ils fonctionnent dans un but précis. Ceci est tout particulièrement vrai dans le domaine de la robotique où, aucun système plus général n'existant à un coût raisonnable, le problème principal est de déterminer la

position et l'orientation des *extrémités* des fils présents dans l'image.

1 Les approches bidimensionnelles

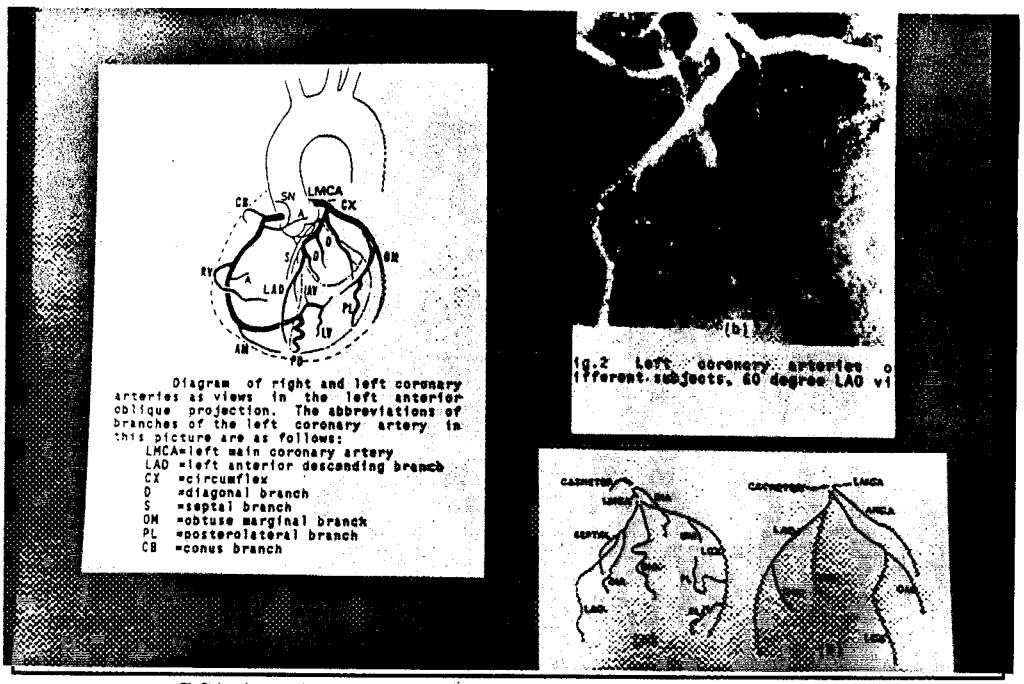
Les systèmes présentés dans ce paragraphe ne nécessitent qu'une approche bidimensionnelle, soit que les scènes à analyser le soient effectivement, soient que les contraintes adoptées sur les scènes et leur environnement permettent de réduire d'une dimension le problème initialement tridimensionnel.

1.1 Le Système de TSUJI et NAKANO

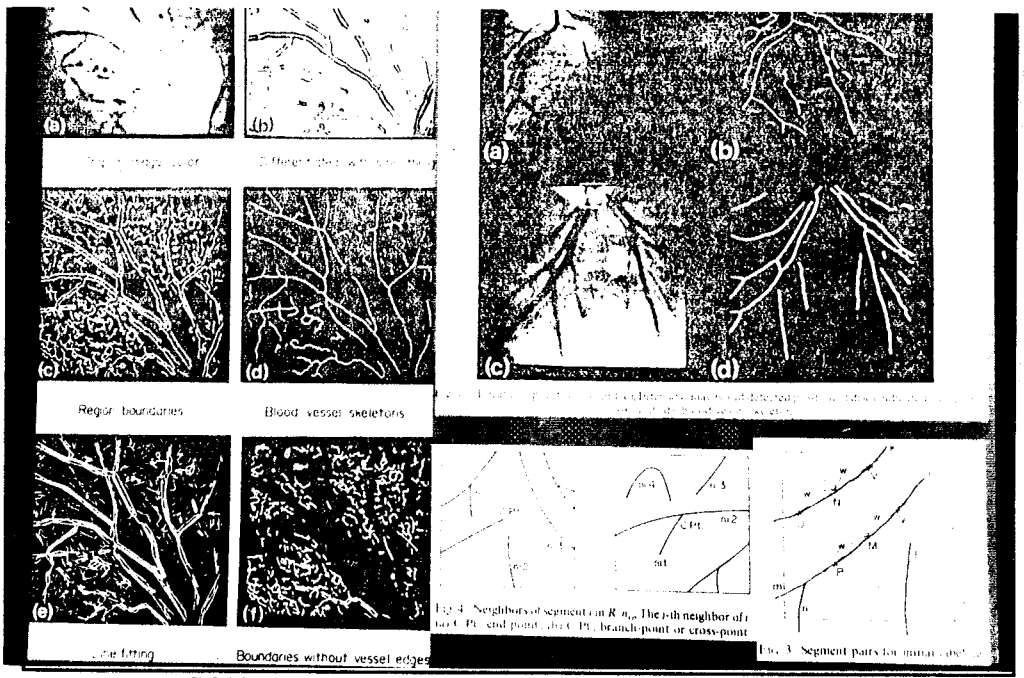
Le système d'interprétation de [TSUJI-81] est destiné à l'interprétation de ciné-angiogrammes (cf. figure C.04.a). Les ciné-angiogrammes dont des images d'un film radiographique du système artériel de la région du coeur. Le but du système est l'identification d'un certain nombre de vaisseaux importants en fonction d'un modèle implanté sous forme de règles. L'analyse est effectuée sur une séquence d'image. L'approche suivie est l'interprétation des lignes de contraste par un système expert. Un prétraitement extrait les lignes de contraste qui sont définies chacune par une séquence de points et une épaisseur moyenne. Le système contient plusieurs ensembles de règles de production. Les *règles d'identification* servent à l'interprétation des lignes, et sont associées avec un facteur de plausibilité. Les *règles de jugement* servent à comparer les interprétations de deux images différentes, et à effectuer un choix en cas de contradiction. Les *règles de contrôle* définissent à chaque instant les sous-ensembles de règles activables. Ces règles traduisent l'ordre dans lequel il faut procéder à l'identification.

1.2 Le Système d'AKITA et KUGA

Le système développé par [AKITA-82] s'attache à analyser des images de fond de l'oeil (cf. figure C.04.b). L'analyse de ces images est dirigée vers l'examen des vaisseaux sanguins (artères et veines) et la détection de deux types de régions caractéristiques. L'analyse de ces quatre types d'"indices d'image" permet la détection et la qualification de maladies chez l'homme comme l'hypertension ou le diabète. La reconnaissance des segments de lignes décrivant les artères et les veines s'effectue sur un réseau de lignes obtenues par un algorithme tenant compte de la forme des objets recherchés. L'interprétation de ces lignes en "artère" ou "veine" est un étiquetage qui s'effectue à partir de trois types de "noeuds-image" : les extrémités, les branchements, et les croisements. Des connaissances du type "les artères (resp. les veines) ne se branchent qu'à partir d'artères (resp. de veines)" ou "les artères (resp. les veines) ne se croisent pas entre elles" ou encore "aucun vaisseau sanguin n'est isolé) permettent d'établir un premier étiquetage. Un algorithme de relaxation permet de compléter cet étiquetage par une analyse de probabilité des relations topologiques entre vaisseaux.



C.04.a: *(d'après [TSUJI-81-])* Le système de Tsuji et Nakano



C.04.b: *(d'après [AKITA-82-])* Le système d'Akita et de Kuga

Figure C.04: Différentes approches des objets filiformes (1)

1.3 Le Système de VERNON

Le système de Vernon [VERNO-84] localise des fils électriques en vue de leur amenée à une machine à sertir. L'approche suivie consiste à contraindre fortement l'espace de travail pour pouvoir se ramener au traitement d'une image binaire bidimensionnelle. De fait, l'espace de travail est conditionné de sorte que la scène présentée à la caméra n'est pas tout à fait aléatoire : les fils électriques présents sont au plus au nombre de deux, et le plateau qui les supporte constitue un fond bien visible et suffisamment contrasté. De plus, les fils sont supposés être "à plat", c'est-à-dire que leur variation spatiale dans la direction orthogonale au plateau est très faible. Il suffit alors de traiter le problème sur les deux dimensions du plateau (xx, yy) la cote z d'un fil à sertir étant connue a priori.

- L'image initiale du système de Vernon (cf. figure C.05.a) est une image à niveaux de gris et la première phase consiste en un seuillage à seuil global en vue de la binarisation de cette image. Comme les nombreux systèmes fondés sur cette technique, le problème essentiel est la détection du seuil de segmentation. Adapté ici à l'image, il est obtenu par un calcul de moyenne sur les niveaux de gris des pixels de l'image qui constituent les "arêtes" détectées par un opérateur adéquat.

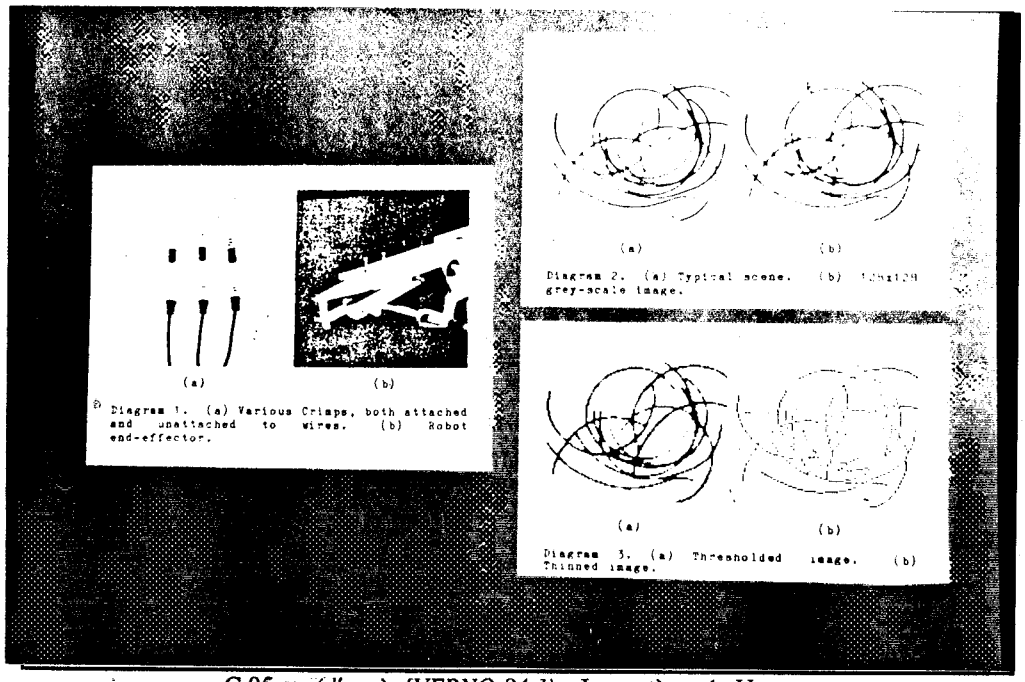
- La seconde phase est une squelétisation des objets détectés dans l'image binaire, de sorte que la largeur des fils observés soit réduite à un pixel. L'algorithme utilisé, respectant la connexité initiale, permet une représentation médiane des fils avec conservation de leur longueur approximative.

- La troisième phase du traitement consiste à isoler les "portions de fil" (assimilables à la notion de "lignes d'image") susceptibles d'être traitées, c'est-à-dire dont au moins l'une des extrémités est libre, et dont la longueur est supérieure à une longueur donnée. Il est alors procédé à la détection d'un point de prise. Là encore des contraintes additionnelles ont été incorporées pour interdire la sélection de fils dont les extrémités se recouvrent.

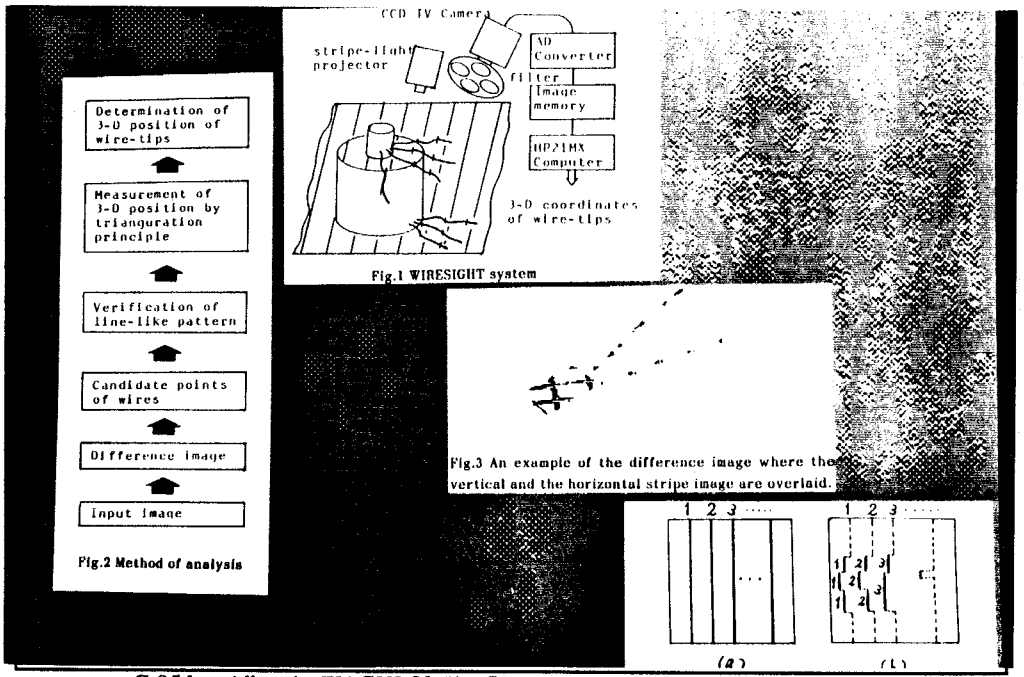
- Les coordonnées tridimensionnelles recherchées sont alors extraites grâce à une modélisation tridimensionnelle de l'environnement, utilisant des polynômes du troisième degré.

D'un point de vue fondamental, l'obtention de résultats fiables par usage de la vision binaire n'est possible que si la complexité de la scène est minimisée au possible : objets suffisamment disjoints pour une bonne séparation, bien illuminés, tels qu'au plus 2 fils se recouvrent pour une bonne interprétation, et un fond de scène bien contrasté.

Finalement, le bon fonctionnement du système est fortement conditionné par l'existence de la table de travail : sans ce plan de travail, ou même si ce plan de travail contient d'autres objets que les fils, le système ne peut rien détecter.



C.05.a: *(d'après [VERNO-84-])* Le système de Vernon



C.05.b: *(d'après [YACHI-82-])* Le système de Yachida, Tsuji et Huang

Figure C.05: Différentes approches des objets filiformes (2)

2 Les approches tridimensionnelles

Il est remarquable de noter que la finalité de la plupart des systèmes suivant une approche tridimensionnelle consiste à localiser des fils électriques en vue de leur saisie par un robot manipulateur. Les systèmes présentés et discutés ici reflètent ce phénomène et leur objet est donc identique à celui de l'application que nous avons développée (détaillée au chapitre C.IV.), mais les méthodes et les conditions d'environnement leur sont souvent spécifiques.

Cependant, les limitations intrinsèques à ces méthodes mises à part, il reste que certaines des contraintes (par exemple quant au nombre de fils pouvant figurer dans les images à étudier) sont directement liées à la non prise en compte de la distinction entre indices d'image et indices de scène.

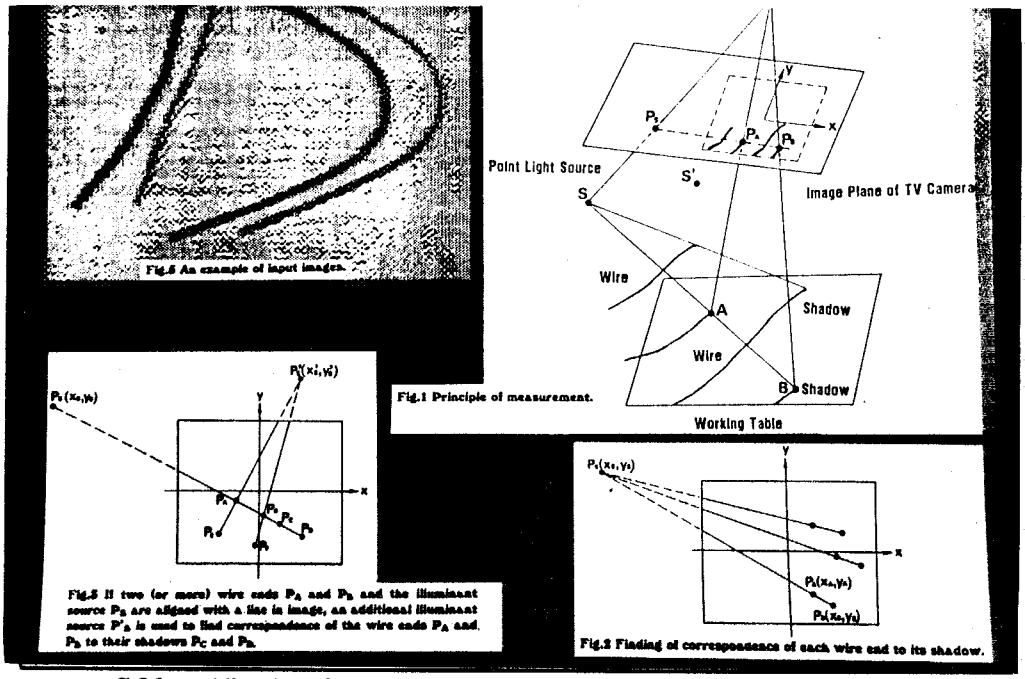
En outre, la description qui suit illustre combien, pour un problème de vision précis (ici la localisation de fils électriques), les méthodes adoptées peuvent être différentes (méthode optique de formation d'images 3-D, inférence de forme à partir des ombres, inférence de forme à partir de la stéréovision simple), tandis que les résultats obtenus sont souvent comparables.

2.1 Le Système de YACHIDA, TSUJI & HUANG

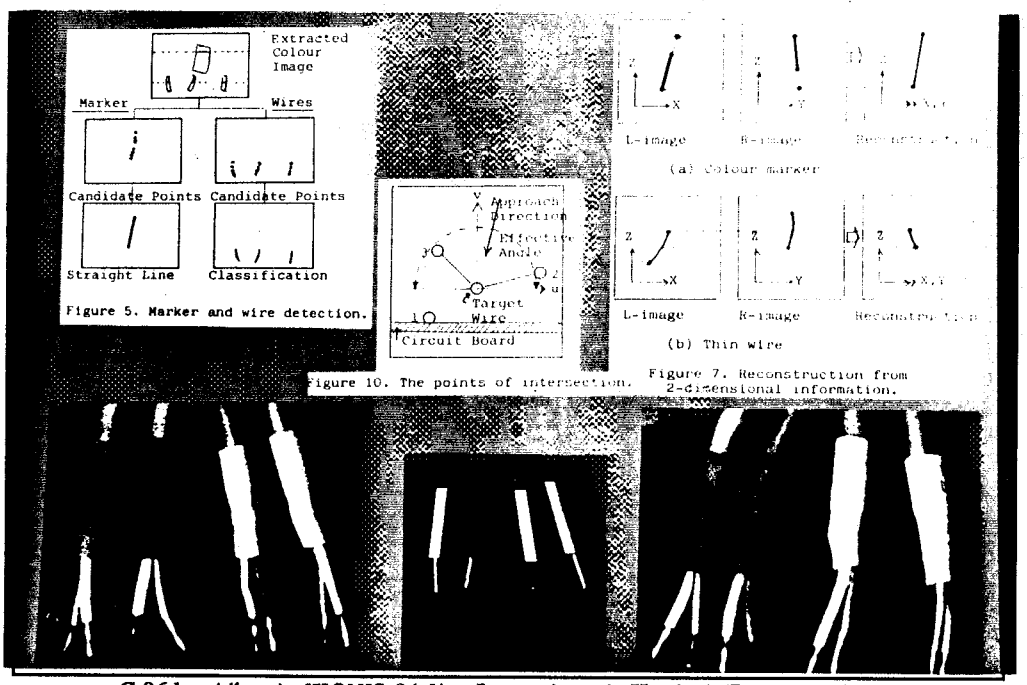
Le système de [YACHI-82] est destiné à localiser des extrémités de fils dans l'espace tridimensionnel, dans le but de l'automatisation du câblage de électrique de pièces comme des moteurs, des transformateurs, etc ... La méthode utilisée est une "méthode active optique de formation d'images tridimensionnelles" (cf. figure C.05.b). Un plan de lumière traverse l'espace observé par une caméra et intersecte certains fils. La détermination des points appartenant à des fils électriques s'effectue par soustraction avec le fond de la scène. Les coordonnées des "points candidats " sont alors calculées par triangulation. Ces points candidats ne sont pas tous des extrémités de fils, sauf accident. Plusieurs coupes de la scènes sont effectuées, et les différentes intersections avec le plan de lumière sont repérées et associées d'une coupe à l'autre. La discrétisation intrinsèques aux différentes prises d'image qui sont effectuées ne permet pas d'assurer l'appartenance de chaque extrémité de fil à au moins l'un des plans décrits. Lorsqu'un fil "disparaît" dans une coupe, un retour à l'image digitalisée représentant la scène permet de préciser la position dans l'image de l'extrémité du fil, et une extrapolation de sa position dans l'espace réel est effectuée à partir des coordonnées tridimensionnelle des points des deux coupes précédentes qui appartiennent au même fil.

2.2 Le Système de TSUJI, YACHIDA, GUO & CHOU

Le but poursuivi par le système de [TSUJI-83] est la localisation de l'extrémité de fils électriques en vue de leur saisie par un robot manipulateur. La méthode, astucieuse, ne nécessite qu'une unique caméra noir-et-blanc. L'information de distance d'un point est obtenue par analyse d'une image contenant à la fois les fils électriques et leurs ombres projetées sur une table de travail (cf. figure C.06.a). Les ombres sont



C.06.a: *(d'après [TSUJI-83-])* Le système de Tsuji, Yachida, Guo et Chou



C.06.b: *(d'après [KONIS-84-])* Le système de Konishi, Takagi et Kitsuki

Figure C.06: Différentes approches des objets filiformes (3)

obtenues par utilisation d'une source lumineuse ponctuelle. La reconstruction de la géométrie des extrémités des fils est obtenue par localisation des fils et de leurs ombres et grâce à des connaissances a priori sur la disposition de la caméra, de la table, et de la source d'illumination.

Les fils et leurs ombres sont extraits d'une image initiale (256 x 256 et 16 niveaux de gris) en appliquant des opérateurs de seuillage et de squelétisation.

La correspondance entre chaque fil et son ombre est établie en examinant la coincidence entre d'une part, les droites définies par les extrémités des indices d'image détectés et d'autre part, les rayons émis par la source de lumière.

La localisation tridimensionnelle et l'orientation des extrémités de fils sont alors déduites de la géométrie connue de l'environnement.

Notons que dans le cas d'indétermination due à un trop grand nombre d'alignements accidentels des indices d'image extraits, il est possible d'utiliser une deuxième source de lumière pour résoudre le conflit.

La méthode utilisée se heurte à plusieurs inconvénients. Tout d'abord, l'utilisation des ombres comme indices primaires nécessite que chaque ombre présente dans l'image signifie quelque chose vis-à-vis des fils observés, et réciproquement que chaque fil devant être observé possède une ombre observable dans l'image. D'autre part, tout comme dans le système de Vernon, l'utilisation d'opérateurs de seuillage et de squelétisation conduisent à tous les inconvénients classiques de la vision binaire.

Les autres inconvénients de la méthode concernent les contraintes imposées sur l'environnement. Ainsi la source de lumière ponctuelle, essentielle pour la méthode, voit sa position contrainte par celles du plan de travail et de la caméra. D'autre part, des informations tridimensionnelles ne peuvent être extraites que si cette source ponctuelle se trouve assez près du plan de travail; mais si cette source est trop proche du plan de travail, il s'ensuit une plus grande difficulté d'extraction des indices d'ombre.

Finalement, et encore plus que dans le cas du système de Vernon, la méthode est liée à l'existence de la table de travail.

2.3 Le Système de KONISHI, TAKAGI & KITSUKI

Les système de [KONIS-84] traite de fils électriques qui sont associés dans un même montage. La méthode suivie permet de reconstruire une partie de l'espace réel observé sur la base d'une inférence de formes à partir de la stéréovision couleur (cf. figure C.06.b). La méthode de reconstruction utilisée permet de déterminer une information 3-D à partir de deux image orthogonales de fils électriques qui sont marqués de couleurs spécifiques. L'orientation dans l'image selon laquelle les projections de fils sont observées est contrainte a priori. La méthode de segmentation des images couleur en régions repose sur l'utilisation d'une combinaison linéaire d'images de couleur filtrées. Sachant que toute région extraite dans une des images filtrées correspond forcément à la marque d'un des fils recherchés et inversement, un

algorithme d'extraction de régions par seuillage et squelétisation suffit à extraire la forme des fils dans chacune des deux images de la paire stéréoscopique étudiée. La reconstruction de la forme des fils est déterminée par une mise en correspondance entre indices de régions qui s'effectue sur l'unique indice de couleur.

3 Discussion

Par rapport à chacune des méthodes des méthodes que nous venons de décrire, la méthode que nous avons développée se distingue à chaque fois suivant plusieurs directions : c'est ce qui ressortira implicitement à la fin de cette partie du rapport.

Un point de différence commun de ces méthodes avec notre approche est la non-distinction (sauf peut-être chez [AKITA-82]) entre indices d'image et indices de scènes. Il nous semble que c'est une des raisons pour lesquelles tous les systèmes tridimensionnels que nous venons de décrire ne peuvent plus fonctionner dès que le nombre d'objets filiformes est trop important. Inversement, l'existence de ces méthodes démontre que la distinction entre indices d'image et indices de scène n'est pas indispensable si le nombre des objets filiformes est très faible.

Finalement, très appuyé par l'analogie avec la méthode de [INOUE-84], il existe avec notre propre méthode un point de convergence de toutes les méthodes tridimensionnelles que nous venons d'évoquer : dans tous les cas, l'étude tridimensionnelle des objets filiformes est plus liée à une nécessité de coopération bras-oeil qu'à un réel souci de description de la scène observée.

Chapitre C.III

Identification et localisation d'objets filiformes

Les solutions aux problèmes plus théoriques que nous venons d'aborder dans le premier chapitre de cette partie, ajoutés aux premiers résultats obtenus sur les image noir-et-blanc (cf. partie B) nous ont encouragés à poursuivre sur la voie de l'implantation partielle des niveaux que nous défendons ici. Ce chapitre présente, du niveau IMAGE au niveau OBJET, les outils qui ont été développés pour notre seconde expérimentation : l'analyse de scènes composées d'objets flexibles et gauches (mais filiformes), sur la base de l'inférence de formes à partir de la stéréovision simple couleur.

Néanmoins, les algorithmes d'inférence, de contrôle et d'enrichissement des niveaux les plus bas ne sont pas détaillés ici explicitement : ne faisant pas intervenir la stéréovision, nous nous contentons de donner les adaptations nécessaires aux algorithmes déjà découverts (partie B), pour tenir compte de la donnée supplémentaire que constitue la couleur. La présentation de leur fonctionnement est présentée de façon simplifiée dans le chapitre suivant qui décrit l'application industrielle à laquelle ces travaux ont donné lieu. Les paragraphes 1 à 3 décrivent donc les extensions au niveaux IMAGE, INDICES D'IMAGE et INDICES DE SCENE.

Les processus de plus haut niveau quant à eux sont la traduction directe de notre position vis-à-vis de la stéréovision, que nous avons explicitée au début de cette par-

tie. C'est sur ces niveaux que nous avons travaillé avec le plus d'approfondissement. Les outils nécessaires à l'atteinte du niveau OBJET (et tout spécialement ceux liés au problème de la localisation que nous n'avions pas étudié pour notre première expérimentation) sont explicités au cours du paragraphe 4, et sont détaillés par les paragraphes 5 et 6.

La figure C.07 montre où se situent les différentes procédures inter- et intra-niveaux qui sont entrées en jeu. Une partie des travaux rapportés ici sont exposés dans [DEMAZ-85a] et [DEMAZ-85b].

Là encore, d'un point de vue de l'implantation, seuls le niveau IMAGE, les procédures d'enrichissement de ce niveau et certains outils mathématiques nécessaires à l'interprétation en indices d'image sont écrits en FORTRAN sur LSI 11/23 (Système RT11B), ainsi que tous les algorithmes se rapportant au traitement des attributs de couleur. Tous ces travaux ont été réalisés par Jose-Luis GORDILLO au sein du système CAICOU. Tout comme dans le cas du système CAIMAN, nous ne reviendrons pas ici sur le travail déjà consigné dans [GORDI-86].

Les niveaux INDICES D'IMAGE, INDICES DE SCENES et OBJET, ainsi que toutes les autres procédures ont été écrites en MACLISP sur HB70 (Système Multics).

1 Les Indices d'Image et les Indices de Scène à rechercher

" Percevant les branches enchevêtrées d'un arbre mort, je ne puis d'abord distinguer si la branche a *est en avant ou en arrière de la branche* b *jusqu'au moment où, atteignant leur point d'intersection, je vois passer* a *sur* b *et* b *sous* a *: cette relation "*a *sur* b*" acquiert alors le rôle d'un indice dont l'utilisation me permet de structurer immédiatement l'ensemble des positions relatives des autres segments de* a *et de* b*. Mais cet indice n'est ainsi qu'une partie ou un aspect de l'ensemble total constituant le signifié. "* [PIAGE-61]. Cette petite phrase de Piaget illustre parfaitement les indices d'image et leurs interprétations en indices de scène dans le contexte des objets filiformes :

- *"... atteignant leur point d'intersection ..."* : l'occultation entre objets filiformes (information de niveau INDICES DE SCENE) se traduit par l'existence de pseudo-sommets, dont la localisation au niveau IMAGE se situe aux intersections des lignes de contours définissant les bords apparents des objets filiformes observés. A priori, deux objets filiformes s'occultant définissent ainsi quatre points d'intersection qui sont des noeuds-image tels que nous les avons définis dans notre première expérimentation. Si on joue sur le fait qu'un objet filiforme est parfaitement défini par son axe médian et une fonction de distance reflétant l'"épaisseur apparente" de l'objet filiforme, on peut considérer qu'une occultation entre objets filiformes ne correspond qu'à un seul "noeud-image double", qui est défini comme l'intersection des axes médians des objets concernés. De la même façon, les autres "noeuds-image double"

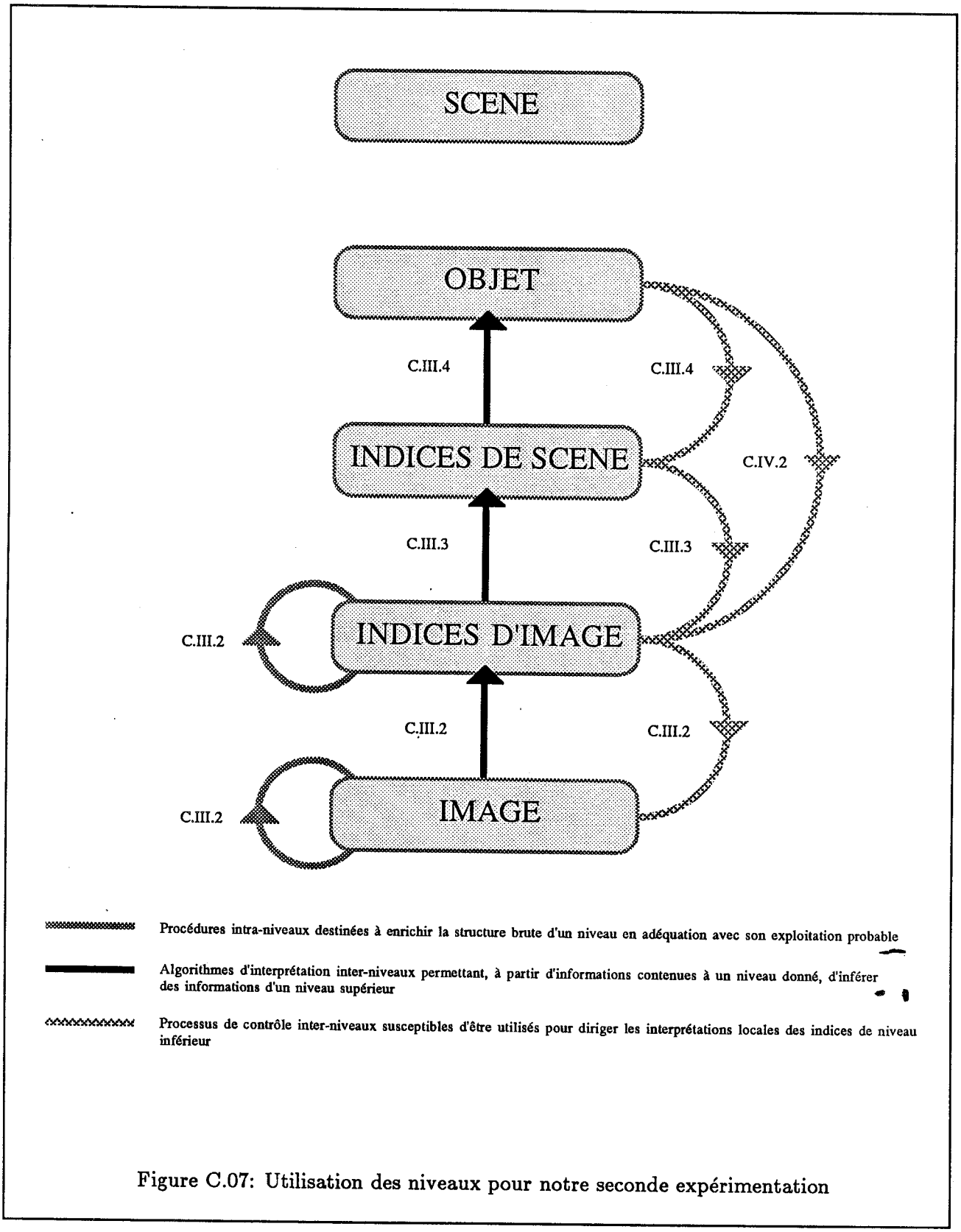


Figure C.07: Utilisation des niveaux pour notre seconde expérimentation

concernant un objet filiforme sont définis comme les extrémités de son axe médian.

Par souci de cohérence vis-à-vis de la définition des noeuds-image double, il est nécessaire que les deux lignes de contraste délimitant la région occuppée dans l'image par un objet filiforme ne puissent engendrer qu'une unique "ligne-image double". Notons que si la ligne-image double est extraite une suite connexe de segments de droites, nous introduisons de ce fait la notion de "droite-image double" en parfaite cohérence avec ce que nous avions défini comme "droite-image" pour notre première expérimentation. Ceci n'est possible qu'à la condition d'accepter dans le même temps le fait que l'intersection de deux droites-image doubles, consécutives et constitutives de la ligne-image double considérée, définissent un noeud-image double. Notons que dans ces conditions, une ligne-image double est alors simplement le reflet d'une double "ligne-brisée-image" au sens où nous l'avons introduit pour les images noir-et-blanc.**Nous intéressant à la seule notion de filiformité, nous avons donc introduit deux nouveaux indices primaires : "les droites-image double" et "les noeuds-image doubles".**

- *"... je vois passer* a *sur* b *..."* : pour voir passer un objet filiforme au-dessus d'un autre, il faut pouvoir vaincre la symétrie apparente du problème, identique à celle de l'interprétation des noeuds-image de type "X" pour le cube de Necker (cf. figure B.03). Nous disposons ici d'un moyen de distinguer les deux lignes-image double : leurs informations colorées. Cela nous amène à introduire toutes les définitions liées à la couleur. Un "point de couleur" (notion de même niveau que le point de contraste) d'une ligne-image double est défini par les coordonnées RVB du point dans l'image de l'axe médian de l'objet filiforme auquel il correspond (ou par tout autre combinaison de valeurs RVB faisant intervenir un voisinage de ce point). Une "droite de couleur" est une succession de points dont les valeurs de couleur restent dans un domaine de valeurs RVB donné, tout comme une "droite de contraste " est une succession de points qui restent dans une direction donnée. Par extension, il est immédiat de définir une "droite de couleur et de contraste", comme une double droite de contraste la plus "longue" possible et dont l'axe médian soit aussi une droite de couleur, ou inversement. L'interprétation au niveau INDICES D'IMAGE d'une droite de couleur et de contraste s'appelle une "droite-image double et de couleur". Ses deux extrémités s'appellent des "noeuds-image double et de couleur". *Notons que les extrémités d'une droite-image double et de couleur correspondent à des ruptures de contraste ou de suivi de couleur.* **Nous intéressant à la notion de filiformité et de couleur, nous avons encore pu introduire deux nouveaux indices primaires : "les droites-image doubles de couleur" et "les noeuds-image doubles de couleur".**

Le tableau suivant montre les dénominations des indices du niveau IMAGE et par extension, ceux du niveau INDICES D'IMAGE :

IMAGE N&B POLYEDRES	IMAGE N&B OBJETS FILIFORMES	IMAGE COULEUR OBJETS FILIFORMES
contraste	contraste	contraste ET couleur
droite-image	droite-image double	droite-image double et de couleur
noeud-image	noeud-image double	noeud-image double et de couleur
ligneB-image	ligne-image double	ligne-image double et de couleur

Pour nous rapprocher de notre première expérimentation et tenir compte de la nouvelle information de couleur, nous considérons essentiellement dans le raisonnement qui suit pour les niveaux INDICES DE SCENE et OBJET les "droites-image double et de couleur", les "noeuds-image double et de couleur", et les "lignes-image doubles et de couleur". Ces indices d'image primaires sont différents des indices d'image que nous avions introduits dans la première expérimentation. Mais pour une meilleure convivialité et surtout parce que les traitements ultérieurs seront identiques à ceux que nous avons exposés dans la seconde partie de ce manuscrit, il nous arrivera à partir de maintenant d'omettre totalement ou partiellement la désinence "double et de couleur" après les termes "droites-image", "noeuds-image" et "lignes-image".

2 Extraction des Indices d'Image

Comme nous l'avons suggéré à plusieurs reprises, la puissance d'un système de vision dépend fortement de la qualité de sa capacité à extraire de "bons" indices d'image. Il est très difficile d'extraire des lignes-image (de couleur ou non) qui ne soient pas géométriquement droites, si on cherche à les exprimer autrement qu'en termes d'une suite de segments de droites. C'est ce que montrent des travaux comme ceux de [DUDA-72] ou encore ceux de [AGIN-81].

Nous nous plaçons délibérément dans le cas où une ligne-image de couleur est une succession de droites-image de couleur. Il suffit d'un point de vue de l'extraction d'utiliser un algorithme d'extraction d'indices qui (si on souhaite éviter l'utilisation de procédures d'enrichissement au niveau INDICES D'IMAGE) considère effectivement une ligne courbe comme une succession de segments de droite [LUX-83a]. C'est grâce à la version couleur du système CAIMAN, c'est-à-dire le système CAICOU [GORDI-86], que nous avons pu extraire les indices d'image des scènes observées. Tout comme le logiciel CAIMAN fournit des indices de contraste dans une image noir-et-blanc, le système CAICOU permet d'extraire incrémentalement des indices de contraste et/ou de couleur dans des images colorées, dans un formalisme parfaitement adapté au notre. Par exemple, l'activation de l'opérateur de "suivi double avec extraction de couleur" correspond effectivement à l'activation de deux opérateurs, dont l'action simultanée permet de découvrir ce qui, au niveau INDICES D'IMAGE est interprétable en une "ligne-image double et de couleur". L'opérateur de suivi double avec extraction de contraste fournit une ou plusieurs lignes double qui sont des droites de contraste, et telles que l'axe médian de chaque ligne double possède une valeur de couleur RVB.

Actuellement, la segmentation est possible sur la base des informations en rouge, vert ou bleu, mais est prévue pour fonctionner à partir des données obtenues par application de la transformée de Karhunen-Loeve sur l'espace RVB.

La qualification des indices de couleur s'effectue dans l'espace RVB, et la rupture de continuité suivant l'une de ces trois dimensions ou suivant H, L, ou S, entraîne la création d'un noeud-image.

Les procédures d'inférence et de contrôle utilisées pour interpréter les données au niveau INDICES D'IMAGE sont exactement les mêmes que dans notre première expérimentation. Certaines de ces procédures seront détaillées dans l'exposé de notre expérimentation ayant trait à l'identification et la localisation de fils électriques. La figure C.08 montre les différentes causes d'arrêt d'un suivi double avec extraction de couleur dans le cas particulier où la couleur de chaque objet est uniforme.

Les procédures d'enrichissement sont tout aussi identiques. La création de "lignes-image double et de couleur" s'effectue en fait implicitement par application itérative de l'opérateur de suivi, tant qu'aucun croisement réel n'est rencontré. C'est exactement ce qu'illustre la phrase de Piaget que nous avons citée au début de ce chapitre.

Le type de chaque noeud-image peut lui aussi être calculé comme nous l'avions fait dans la première expérimentation, en ne considérant le calcul que vis-à-vis des axes médians des droites-image doubles extraites, et il est inutile avec cette définition d'introduire de nouveaux types de noeuds-image ...

3 Des Indices d'Image aux Indices de Scène

... Généralement (c'est-à-dire sauf jonction accidentelle), les noeuds-image d'une ligne-image représentant un objet filiforme sont 1/ de type "Y" ou "X" dans le cas des croisements d'objets filiformes, 2/ de type "PENDANT" dans le cas de leurs extrémités, et 3/ de type "L" dans le cas où le noeud-image correspond à une discontinuité dans la qualification de la couleur de l'indice. On en déduit immédiatement la nature des opérateurs d'interprétation des indices d'image en indices de scène : elle est identique à celle que nous avons déjà introduit, à ceci près que chaque noeud-image, et chaque droite-image est associée à des attributs colorés exprimés dans un repère RVB ou HLS, tandis que les attributs des indices de scène correspondant sont exprimés dans le repère HLS.

L'inférence des indices de scène à proprement parler se fait toujours, elle aussi, de la même façon. Mais il est clair que la forte contrainte imposée par la définition même du domaine d'expérimentation permet d'adjoindre aux ensemble d'opérateurs d'interprétation des règles contextuelles restreignant fortement la combinatoire d'interprétation. L'application des règles contextuelles sur les les ensembles I_{TYPE} des opérateurs d'interprétation possible d'un noeud-image N de type $TYPE$ engendrent des "ensembles d'interprétation réduits" IR_{TYPE} dont nous donnons des exemples maintenant, en supposant que toute extrémité de ligne-image est un noeud-image de type PENDANT :

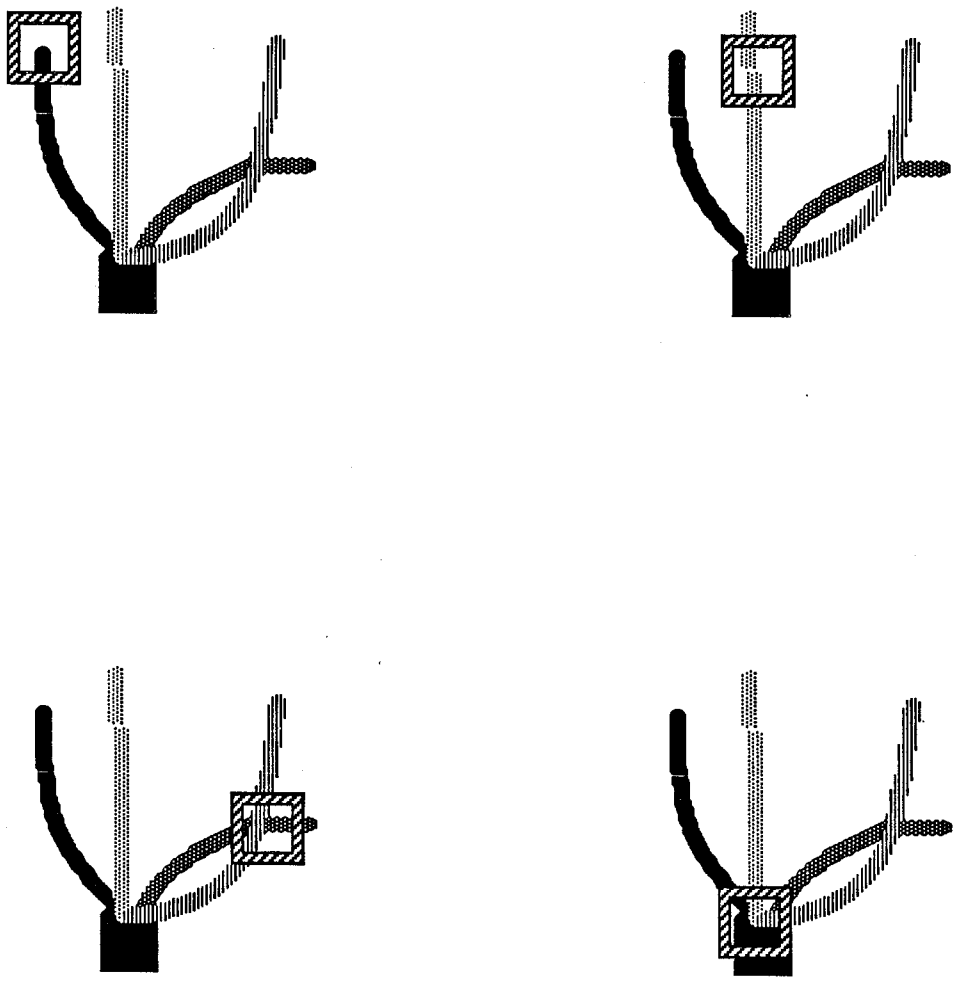


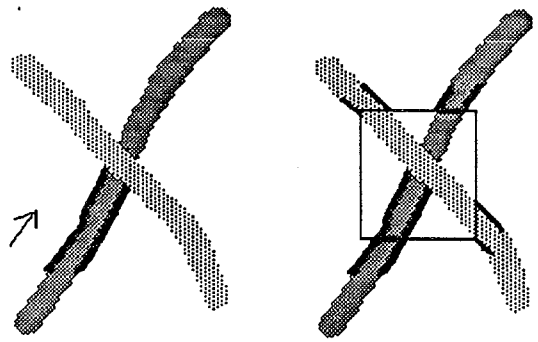
Figure C.08: Extraction d'indices d'image: causes d'arrêt d'un suivi double

- $N.IR_{PENDANT} = \{\{\{A_1\}\}\}$
- $N.IR_L = \{\{\{A_{12}\}\}\}$
- $N.IR_Y = \{\{\{A^1_{12}\}\}\{\{A^2_{23}\}\}Pi\}$, les Pi étant obtenus par permutations circulaires sur la liste des indices (1 2 3)
- $N.IR_{XX} = \{\{\{A_{13}\}\{A_{24}\}\}\{\{A_{13}\}\{A_{24}\}\}Pi\}$ les Pi étant obtenus par permutations circulaires sur la liste des indices (1 2 3 4)

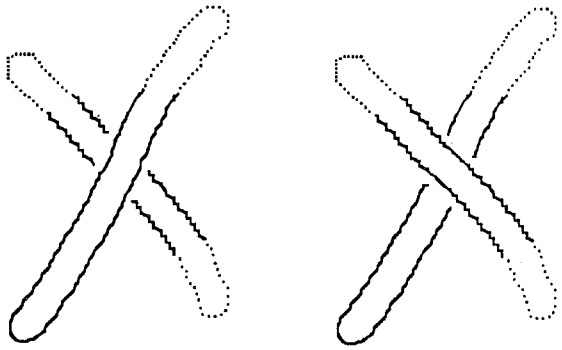
De façon générale, l'interprétation d'un noeud-image ne peut engendrer la création d'un sommet qu'à partir du moment où ce noeud-image est l'extrémité réelle de la projection d'un des objets filiformes présent dans la scène observée. Inversement, chaque ligne-image double et de couleur est interprété comme une surface de révolution de couleur uniforme, dont le profil est fourni par une des deux lignes de contraste qui lui ont donné naissance.

Les procédures de contrôle sont elles aussi identiques à ce que nous avons vu dans le cas des "indices d'image simples" qui sont utilisées pour étudier les objets polyédriques de notre première expérimentation. Il s'agit d'assurer la cohérence et de guider l'interprétation des indices d'image non encore interprétés, par propagation des contraintes imposées par les interprétations possibles déjà déduites.
La figure C.09b est l'occasion de renouveler l'exposé des différences entre ces procédures de contrôle et un algorithme de contrôle global tel qu'il est illustré : après interruption de la poursuite d'une ligne-image (niveau INDICES D'IMAGE) et analyse dans une fenêtre de recherche autour du point d'arrêt (figure C.09.a), l'interprétation au niveau INDICES DE SCENE des départs potentiels de droites-image (cf. figure C.09.b) permet de choisir (cf. figure C.09.c) quelle droite-image doit être poursuivie en priorité. Ce type de contrôle global est du même ordre que celui que [SOUVI-83] utilise pour sa phase de reconnaissance d'objets pouvant s'occulter partiellement (même si le niveau auquel le contrôle s'effectue à un niveau qui n'est pas explicité comme étant un niveau INDICES DE SCENE)

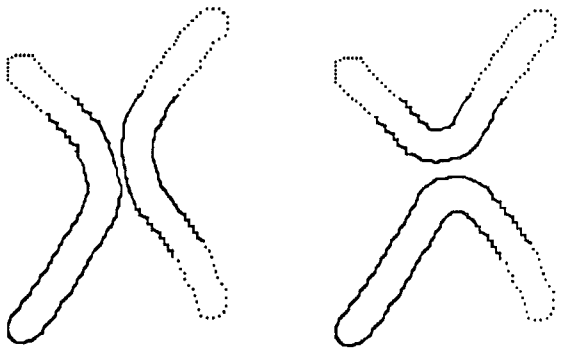
Il est temps à ce moment de l'exposé, de ne pas oublier que **tous les traitements dont nous avons parlé du niveau IMAGE jusqu'à ce point de la présentation ont trait au travail indépendant de toute stéréovision qui est effectué dans chaque image en parallèle.** Mais l'inférence des indices de scène est l'occasion de montrer toute la puissance des procédures de contrôle local dans le cas de la stéréovision. En effet, nous avons dit au cours de la deuxième partie qu'à chaque occultation d'objet correspond une relation symbolique entre indices de scène. Connaissant parfaitement la géométrie (tant interne qu'externe) du montage stéréoscopique, il est donc possible d'exploiter toute relation symbolique inférée dans une image de la paire stéréoscopique pour aider à l'interprétation des indices d'image de l'autre image de cette paire. Nous n'avons pas encore réellement exploité cette contrainte, dont l'utilité est suggérée par exemple chez [OHTA-83] [THORP-83a] [HERMA-84] ou encore [YUILL-84]. La puissance d'une telle contrainte nous paraît dépasser de beaucoup celle de la simple contrainte sur les lignes



C.09.a
Analyse locale d'un noeud-image
à l'interruption d'un suivi double



C.09.b
les quatre interprétations
contextuelles valides



C.09.c
Le choix de la ligne-image
à extraire en priorité

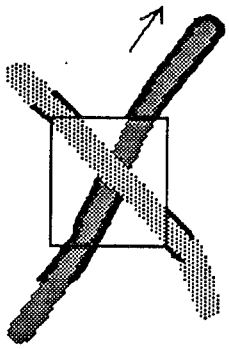


Figure C.09: Interprétation des indices d'image en indices de scène

épipolaires. En effet, cette dernière ne fait que refléter les relations symboliques d'adjacence entre indices de niveau quelconque, en ignorant l'information contenue dans les relations de profondeur. L'utilisation de cette "contrainte sur les profondeurs" est tout aussi fortement suggérée par [PIAGE-61] que nous avons évoqué au début de ce chapitre et dont nous poursuivons maintenant la citation :

" Mais cet indice n'est ainsi qu'une partie ou un aspect de l'ensemble total constituant le signifié. Et il n'en consiste qu'une partie interchangeable, car j'aurais pu d'abord percevoir que a *est plus proche de moi que* b*, ce qui m'aurait conduit à anticiper, en cherchant le point d'intersection, le passage de* a *sur* b *: en ce cas, l'évaluation globale des distances m'eût servi d'indice ou de signifiant perceptif et la position de* a *sur* b *en leur intersection eût été par exemple éclairée par cet indice, si elle avait été peu visible (devenant ainsi "signifiée"). En bref, l'indice perceptif est déjà un signifiant, mais ne constituant qu'un aspect partiel et interchangeable du signifié, tandis qu'un symbole et a fortiori un signe sont de plus en plus différenciés de leurs signifiés. "*

4 Interprétation des Indices de Scène en Objets

Nous sommes ici arrivés à un niveau de représentation tel que l'abstraction est presque totale. Il reste à travailler sur l'objet pour découvrir sa forme et sa nature réelles. Il reste aussi à se décentrer totalement par rapport à l'observateur, c'est-à-dire à passer dans un espace tridimensionnel où les objets peuvent être décrits dans leur propre repère. Si on considère l'interprétation en objets comme composé de deux phases, l'une de mise en correspondance géométrique et topologique (i.e. une "identification"), l'autre comme une localisation, alors c'est la première phase qui assure l'abstraction, tandis que pour la décentration est assurée pour sa plus grande partie par le seconde phase. Ce sont ces deux phases que nous expliquons dans les paragraphes suivants, en insistant sur la décentration.

Notons simplement que la phase d'identification permet d'obtenir des cylindres généralisés facilement modélisables du fait que l'axe médian observé *est* la génératrice de l'objet observé. L'attribut de couleur est exprimé dans un repère de type Munsell (cf. figure C.01) qui en particulier permet d'exprimer la couleur observée par une association de termes intelligibles par tout le monde et facilitant tout dialogue au niveau OBJET.

Tout aussi simplement, un calibrage inverse simple permet de retrouver les coordonnées de cet objet dans un repère tridimensionnel quelconque qui n'est plus lié à l'observateur.

5 Identification - Modélisation des objets filiformes

5.1 Critères de Mise en Correspondance

A un niveau d'abstraction aussi élevé que le niveau INDICES DE SCENE, la mise en correspondance peut s'effectuer entre indices riches d'informations significatives du

niveau atteint. Par ailleurs, dans le cas d'une stéréovision large telle que nous l'avons adoptée, il n'est plus possible d'effectuer une correspondance sur la base simple d'un calcul de disparité. Même si nous ne l'avons pas encore étudié, nous pensons que le problème de mise en correspondance est la recherche d'une structure extrémale dans un graphe traduisant des relations géométriques mais aussi topologiques entre indices de scène. En cela, nous rejoignons les travaux de [MOHR-84].

Pour l'instant nous ne nous sommes contentés que de mesures de cohérence entre informations de couleur, et de la compatibilité par rapport à la contrainte de la ligne épipolaire. C'est ainsi que, sans même établir de graphe, nous établissons une mise en correspondance entre éléments gauche / droite dès qu'ils ont :

- soit des valeurs de luminance comparables
- soit des valeurs de teinte comparables
- soit des valeurs de saturation comparables
- soit des ordonnées comparables dans le repère image, ce qui est la traduction exacte du fait que nous nous contentons d'une approximation d'horizontalité des lignes épipolaires.

Toutes ces relations n'ont pas le même poids dans la décision de mise en correspondance. Ainsi, nous avons observé statistiquement que du point de vue des informations colorées, les coordonnées T L S sont d'importance décroissante dans cet ordre. Cela s'explique en particulier par le fait que la teinte est une information intrinsèque d'une surface et qu'elle ne varie ni selon l'intensité de la source lumineuse ni selon le point de vue de l'observateur (dans ce sens, passer de l'information RVB à une information TSL correspond aussi à une décentration). Ces résultats, cohérents avec ceux de [KENDE-76] sur le choix des critères de correspondance, traduisent aussi le fait que la saturation est un indice très peu fiable. La comparaison de la contrainte de la ligne épipolaire par rapport aux trois coordonnées colorées ne peut par contre relever que de l'empirisme. Tout dépend de la confiance qui est accordée au dessin du montage stéréoscopique. Dans tous les cas, il nous paraît plus important de sacrifier la qualité des couples de correspondance à leur quantité, ce qui correspond à une avoir une "vision floue" de la scène, quitte à ce que des mesures ou des déplacements ultérieurs viennent a posteriori préciser la localisation de ces objets.

Finalement, à tous les niveaux, les éléments de représentation possèdent des attributs supplémentaires par rapport à leurs homologues de la première expérimentation. Ils contiennent tous des informations colorées, de la façon et suivant les modèles qui ont été présentés lors de notre étude du problème de la couleur (cf. figure C.01). C'est une des raisons pour lesquelles l'identification d'un objet filiforme peut on seulement s'effectuer au niveau OBJET comme nous l'avons réalisé, mais aussi (et ceci en ne considérant que la couleur) à tous les niveaux intermédiaires.

5.2 Modélisation des objets filiformes

Aux niveaux INDICES D'IMAGE et INDICES DE SCENE, les modèles sont des représentations "par l'axe médian" auxquelles on associe une mesure d'"épaisseur ". L'axe médian est lui-même représenté par une ligne-brisée constituée de droites : ceci est parfaitement cohérent avec la notion, au niveau indices d'image, de "droite de couleur".

Plus généralement, la démarche que nous avons suivie pour le choix de nos modèles de représentation est identique à celle de notre première expérimentation : les modèles de représentation sont hérités de la représentation au niveau OBJET qui semble la mieux adaptée. La seule vraie différence avec la première expérimentation est que toutes les représentations ont ici rapport non plus à une modélisation par "sommet et arêtes", mais à une modélisation par "cylindres généralisés", notion que nous avons décrit précédemment (cf. figure C.10).

C'est le type de représentation le plus populaire aux niveaux équivalents à notre niveau OBJET. De très nombreux objets complexes peuvent être modélisés par composition de cylindres généralisés. Dans notre cas, le choix d'une telle représentation ne fait que traduire un des critères que nous avons utilisé pour définir la notion d'objet filiforme. Il est clair qu'une telle représentation convient particulièrement aux objets composés de parties allongées. Les formes non allongées peuvent aussi être représentés par ce moyen, mais le choix de l'axe est moins évident : par exemple, un cube possède trois axes possibles tous aussi convenables.

6 Modèle et Calibrage du Capteur - Localisation

6.1 Calibrage Stéréoscopique

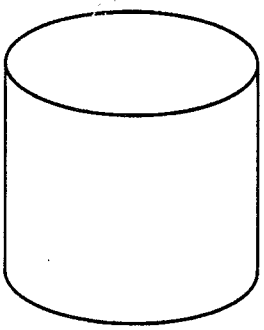
Dans un premier temps, et pour introduire les simplifications de calibrage que nous avons pu effectuer, nous définissons de façon analytique le montage stéréoscopique que nous avons dessiné. Les repères dans lesquels les points sont exprimés sont identiques à ceux de [YUILL-84].

6.1.1 Notations et Définitions (cf. figure C.11.a)

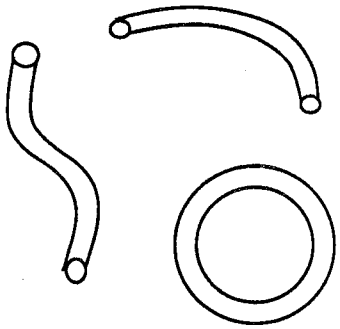
◇ Nous définissons pour chacune des deux caméras composant le capteur stéréoscopique, un repère orthonormé direct, soit $R_g(O_g, \iota_g, \xi_g, \kappa_g)$ pour l'écran E_g et $R_d(O_d, \iota_d, \xi_d, \kappa_d)$ pour l'écran E_d.

◇ De par la restriction du modèle de Longuet-Higgins, et en accord avec le dessin de notre capteur, 1/ la dimension verticale est la même pour chacune des deux caméras, donc $\xi_g = \xi_d$ et 2/ les longueurs focales des deux caméras sont identiques et égales à une même valeur "m" (soit F_g et F_d les points focaux des deux caméras).

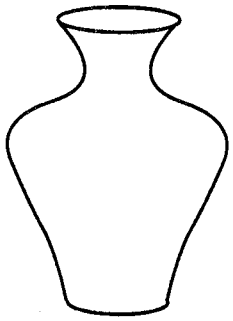
◇ Soit R le repère (O, xx, yy, zz), repère orthonormé direct de l'espace réel. R peut être choisi de sorte que 1/ O soit le milieu de F_g et F_d, 2/ $F_gF_d // xx$, 3/ $\xi_g = \xi_d = zz$.



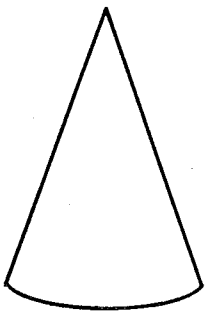
cylindre ordinaire

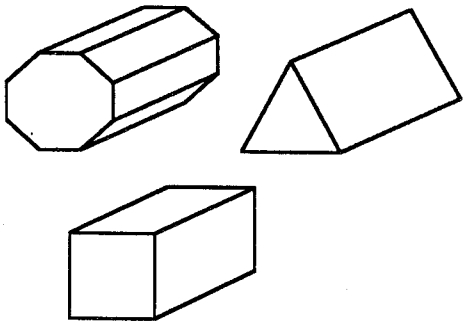


extension à une génératrice gauche

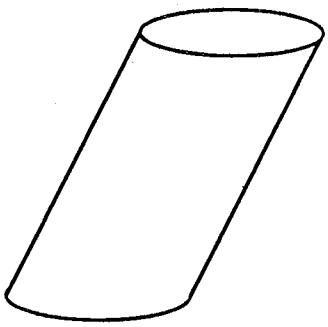


extension à une section
de balayage circulaire
à rayon variable



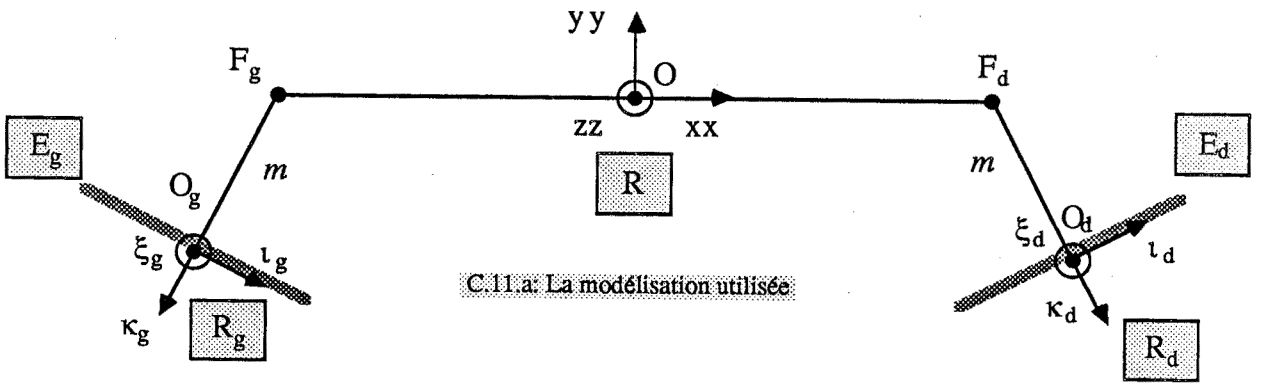


extension à une section
de balayage non circulaire

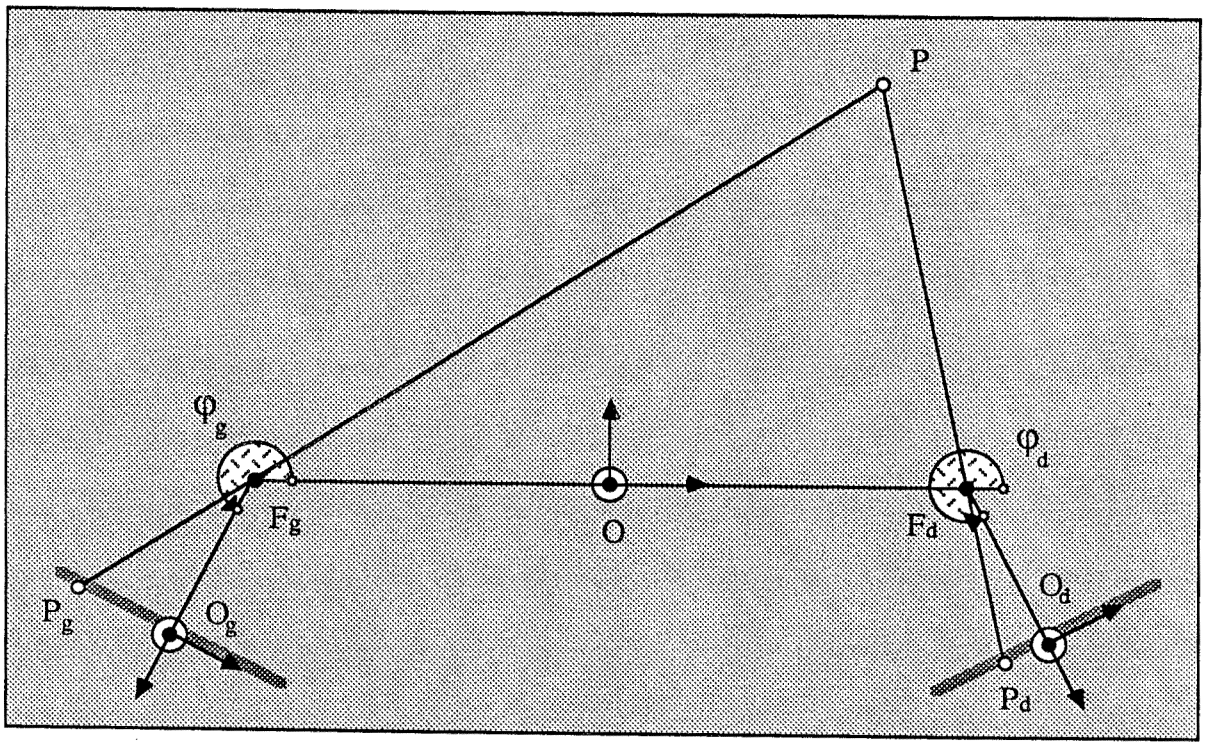


extension à une section de
balayage non perpendiculaire
à la génératrice

Figure C.10: Modélisation des objets filiformes par cylindres généralisés



C.11.a: La modélisation utilisée



C.11.b: Calibrage

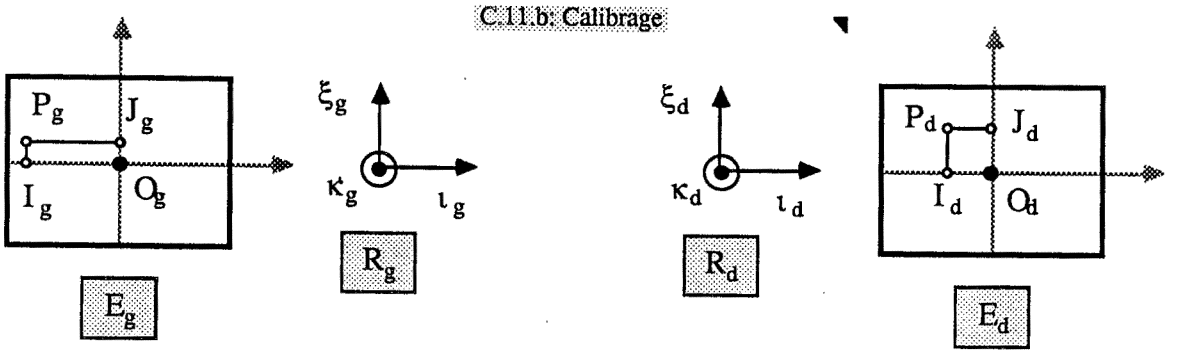


Figure C.11: Modélisation et calibrage du capteur stéréoscopique

6.1.2 Détermination des équations de calibrage (cf. figure C.11.b)

◇ Un point $P(X,Y,Z)_R$ de l'espace réel se projette sur les écrans E_g et E_d respectivement en $P_g(I_g,J_g,0)_{Rg}$ et en $P_d(I_d,J_d,0)_{Rd}$. Le calibrage direct consiste à fournir les formules définissant P_g et P_d en fonction de P.

◇ Les positions des centres des caméras, soit de O_g et de O_d sont fournies par les équations (1) :

$$\begin{aligned} OO_g &= OF_g + m\kappa_g \\ OO_d &= OF_d + m\kappa_d \end{aligned} \tag{1}$$

Les équations des écrans E_g et E_d sont quant à elles définies par les formules (2) :

$$\begin{aligned} OM.\kappa_g &= OO_g.\kappa_g = OF_g.\kappa_g + m \\ OM.\kappa_d &= OO_d.\kappa_d = OF_d.\kappa_d + m \end{aligned} \tag{2}$$

◇ Soit P_g la projection de P sur E_g, alors P_g appartient à la ligne de vue définie par l'équation paramétrique (3).

$$OM(\lambda) = OP + \lambda PF_g = OP + \lambda(OF_g - OP) \tag{3}$$

◇ Sachant que P_g appartient de plus à E_g, à partir de $(2a)$ et (3), il est possible de déduire (4), ce qui fournit la valeur du paramètre λ(5).

$$OP_g.\kappa_g = OP.\kappa_g + \lambda(OF_g - OP).\kappa_g = OF_g.\kappa_g + m \tag{4}$$

$$\lambda - 1 = \frac{m}{(OF_g - OP).\kappa_g} \tag{5}$$

◇ Le point P_g est donc parfaitement défini par l'équation (6) issue de (5) et de (3).

$$OP_g = OF_g + \frac{m(OF_g - OP)}{(OF_g - OP).\kappa_g} \tag{6}$$

◇ Relativement au centre de l'écran E_g, soit O_g, on en déduit $(7a)$, d'après $(1a)$ et (6). Un calcul similaire pour l'écran droit fournit un résultat de même nature $(7b)$.

$$O_gP_g = OP_g - OO_g = -m\frac{((OF_g - OP).\kappa_g)\kappa_g - (OF_g - OP)}{(OF_g - OP).\kappa_g}$$

(7)

$$O_dP_d = OP_d - OO_d = -m\frac{((OF_d - OP).\kappa_d)\kappa_d - (OF_d - OP)}{(OF_d - OP).\kappa_d}$$

◇ Les formules (7) sont générales et ne tiennent pas compte de toutes les restrictions précédemment citées. Rappelons par exemple, que par rapport au repère R choisi, nous avons :

$$OF_d = -OF_g = (f, 0, 0)_R \tag{8}$$

◇ D'autre part, soit ϕ_g et ϕ_d les angles issus de κ_g et de κ_d, définis par :

$$\kappa_g \ : (\cos\phi_g, \sin\phi_g, 0)_R$$

$$\kappa_d \ : (\cos\phi_d, \sin\phi_d, 0)_R$$

Sachant que ξ_g et ξ_d sont tels que :

$$\xi_g \ : (0, 0, 1)_R$$

$$\xi_d \ : (0, 0, 1)_R$$

de la définition des repères R_g et R_d, on en déduit (9), coordonnées des vecteurs ι_g et ι_d, en utilisant les égalités $\iota_g = \xi_g \wedge \kappa_g$ et $\iota_d = \xi_d \wedge \kappa_d$.

$$\iota_g \ : (-\sin\phi_g, \cos\phi_g, 0)_R$$

(9)

$$\iota_d \ : (-\sin\phi_d, \cos\phi_d, 0)_R$$

◇ Nous avons posé initialement le fait que dans le repère R_g (resp. R_d), les coordonnées de P_g (resp. P_d) sont $(I_g, J_g, 0)_{Rg}$ (resp. $(I_d, J_d, 0)_{Rd}$). Mais ces coordonnées sont également fournies par les équations (10a) (resp. (10b)) :

$$O_gP_g = (O_gP_g.\iota_g)\iota_g + (O_gP_g.\xi_g)\xi_g + (O_gP_g.\kappa_g)\kappa_g$$

(10)

$$O_dP_d = (O_dP_d.\iota_d)\iota_d + (O_dP_d.\xi_d)\xi_d + (O_dP_d.\kappa_d)\kappa_d$$

Notons de plus que $(O_gP_g.\kappa_g) = (O_dP_d.\kappa_d) = 0$. Ces égalités résultent du fait que les points P_g et P_d appartiennent respectivement à E_g et E_d.

◇ A partir des formules (7), (8), (9) et (10), on peut alors, par substitution, déduire les relations (11), fournissant les coordonnées image (I_g, J_g) et (I_d, J_d) dans les deux écrans, des points projetés d'un point P de coordonnées (X, Y, Z) de l'espace réel.

$$I_g = O_gP_g.\iota_g = m\frac{-(\sin\phi_g)X + (\cos\phi_g)Y - (\sin\phi_g)f}{(\cos\phi_g)X + (\sin\phi_g)Y + (\cos\phi_g)f}$$

$$J_g = O_gP_g.\xi_g = m\frac{Z}{(\cos\phi_g)X + (\sin\phi_g)Y + (\cos\phi_g)f} \tag{11}$$

$$I_d = O_dP_d.\iota_d = m\frac{-(\sin\phi_d)X + (\cos\phi_d)Y + (\sin\phi_d)f}{(\cos\phi_d)X + (\sin\phi_d)Y - (\cos\phi_d)f}$$

$$J_d = O_dP_d.\xi_d = m\frac{Z}{(\cos\phi_d)X + (\sin\phi_d)Y - (\cos\phi_d)f}$$

6.2 Restriction apportée à notre modèle

◇ Le dessin de notre capteur, utilisant une caméra unique, nous a conduit à utiliser des angles ϕ_g et ϕ_d tels que :

$$\phi_d = \pi - \phi_g = \pi - \phi \begin{cases} cos\phi_d = -\cos\phi_g = -\cos\phi \\ sin\phi_d = \sin\phi_g = \sin\phi \end{cases} \tag{12}$$

Dans ce cas précis, où les miroirs sont symétriques par rapport à l'axe focal de la caméra réelle, nous obtenons les équations de calibrage suivantes :

$$I_g = m\frac{-(\sin\phi)X + (\cos\phi)Y - (\sin\phi)f}{(\cos\phi)X + (\sin\phi)Y + (\cos\phi)f}$$

$$J_g = m\frac{Z}{(\cos\phi)X + (\sin\phi)Y + (\cos\phi)f} \tag{13}$$

$$I_d = m\frac{-(\sin\phi)X - (\cos\phi)Y + (\sin\phi)f}{-(\cos\phi)X + (\sin\phi)Y + (\cos\phi)f}$$

$$J_d = m\frac{Z}{-(\cos\phi)X + (\sin\phi)Y + (\cos\phi)f}$$

◇ Ce modèle de calibrage nous a alors tout naturellement conduit à étudier dans quelle mesure il était possible de rendre égales les valeurs J_g et J_d, c'est-à-dire, *par abus de langage* de réaliser la "propriété d'horizontalité des lignes épipolaires". De fait, cette contrainte est fréquemment utilisée en stéréoscopie simple et surtout dans le cas où le cas des axes optiques des deux caméras sont parallèles. Cependant cette dernière configuration matérielle n'est pas la seule à fournir la propriété recherchée. En effet, d'après (13*b*) et (13*d*), la propriété est satisfaite dès que $J_g = J_d$, c'est-à-dire dès qu'il est possible d'annuler le terme "$(\cos\phi)X$". Les paragraphes qui suivent consignent quelques configurations intéressantes satisfaisant à cette condition.

6.2.1 $\phi = 3\pi/2$ **: Axes optiques parallèles (cf. figure C.12.a)**

$$\begin{aligned} I_g &= m\frac{X+f}{-Y} \\ J_g &= m\frac{Z}{-Y} = J_d \\ I_d &= m\frac{X-f}{-Y} \end{aligned} \tag{14}$$

Ce cas correspond donc à la vision de loin chez l'homme, et c'est celle-là qui est exploitée d'un point de vue de la disparité par [GRIMS-81]. Pour tous les points de la scène, les lignes épipolaires sont horizontales.

6.2.2 $X = 0$ **: Axe bissecteur des axes optiques (cf. figure C.12.b)**

$$\begin{aligned} I_g &= m\frac{(\cos\phi)Y - (\sin\phi)f}{(\sin\phi)Y + (\cos\phi)f} = -I_d \\ J_g &= m\frac{Z}{(\sin\phi)Y + (\cos\phi)f} = J_d \end{aligned} \tag{15}$$

Quel que soit l'angle formé par les deux axes optiques des caméras, tout point de l'axe bissecteur est associé à des lignes épipolaires horizontales.

6.2.3 $(\cos\phi)X = o[(\sin\phi)Y + (\cos\phi)f]$ **(cf. figure C.13)**

$$\begin{aligned} I_g &= m\frac{-(\sin\phi)X + (\cos\phi)Y - (\sin\phi)f}{(\sin\phi)Y + (\cos\phi)f} \\ J_g &= m\frac{Z}{(\sin\phi)Y + (\cos\phi)f} = J_d \end{aligned} \tag{16}$$

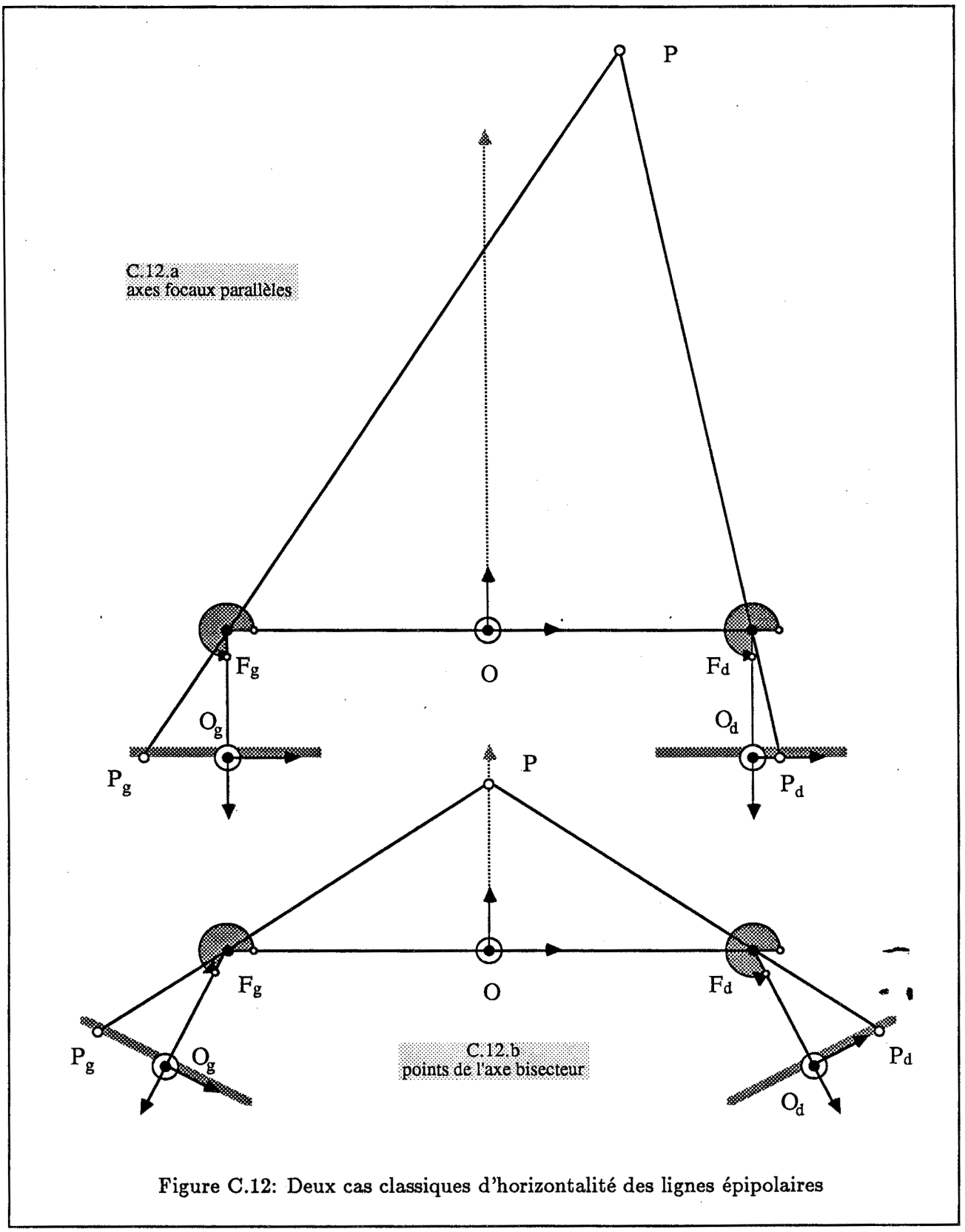


Figure C.12: Deux cas classiques d'horizontalité des lignes épipolaires

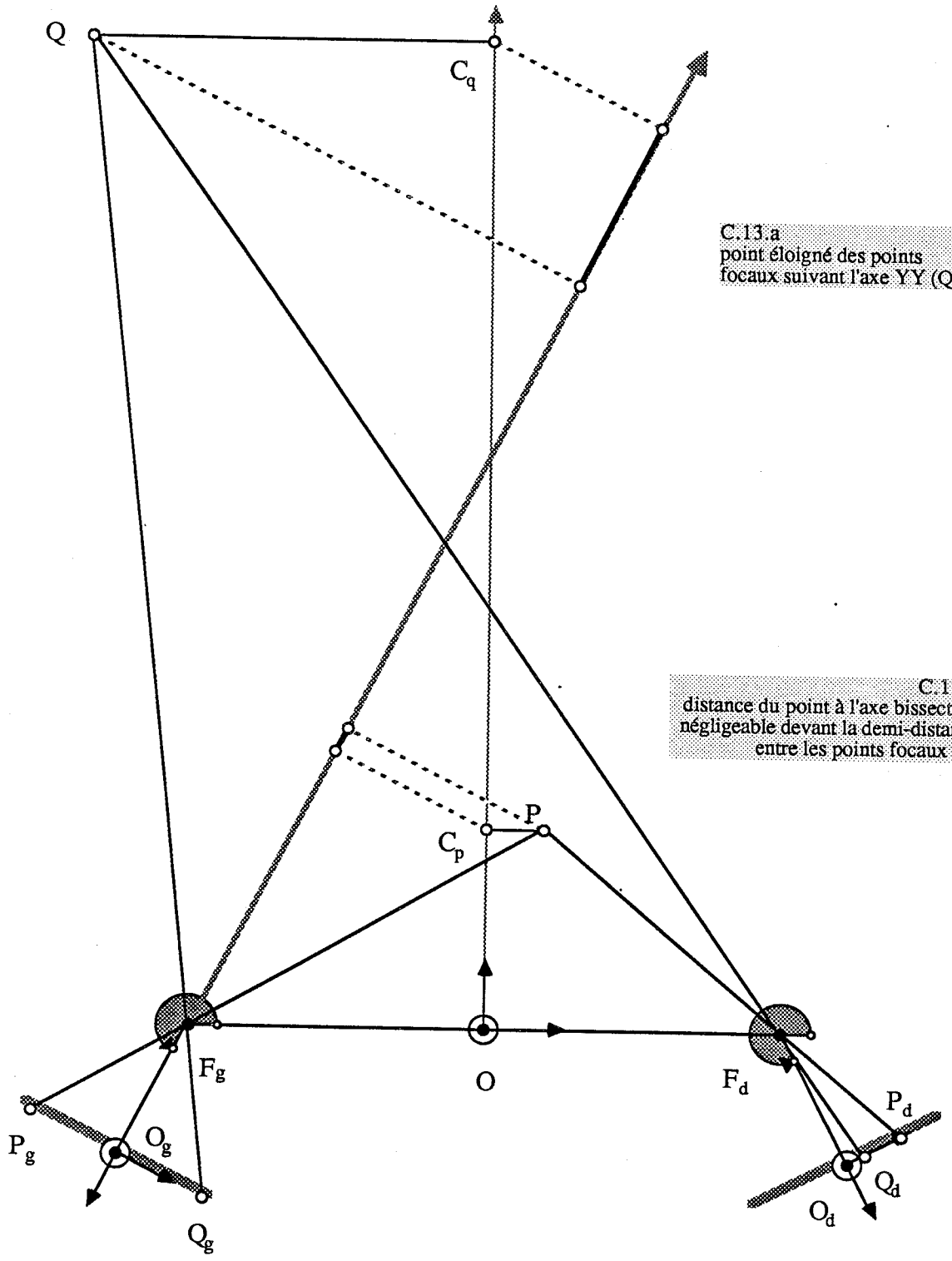


Figure C.13: Approximation des conditions d'horizontalité de l'épipolarité

$$I_d = m\frac{-(\sin\phi)X - (\cos\phi)Y + (\sin\phi)f}{(\sin\phi)Y + (\cos\phi)f}$$

Ce cas est un cas d'approximation général. Si ces valeurs approchées sont retenues pour modéliser le calibrage, il faut alors délimiter les conditions d'utilisation du montage stéréoscopique. En effet, ces équations ne sont valables que

- dans le cas où les points observés sont éloignés des points focaux suivant l'axe des Y,
- dans le cas où la distance des points observés à l'axe bissecteur est négligeable devant la moitié de la distance entre les points focaux des deux caméras.

On retrouve donc ici sous forme approchée, les deux autres cas généraux où les lignes épipolaires sont horizontales.

6.3 Dynamique du Calibrage Simplifié

Par rapport à ce que nous avons déjà dit au début de cette partie, nous nous plaçons donc dans le cas d'une horizontalité des lignes épipolaires. Quelles que soient les raisons de cette horizontalité, on dispose donc d'un système d'équations de forme générale :

$$X = \frac{a_{01}Ig + a_{02}J + a_{03}Id + a_{04}}{a_{13}Ig + a_{14}J + a_{15}Id + 1}$$

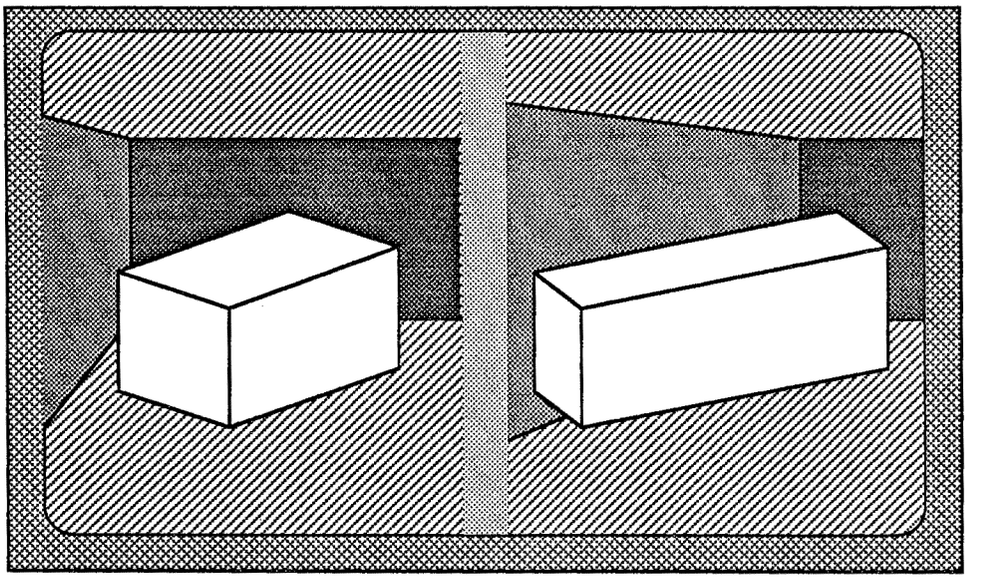
$$Y = \frac{a_{05}Ig + a_{06}J + a_{07}Id + a_{08}}{a_{13}Ig + a_{14}J + a_{15}Id + 1}$$

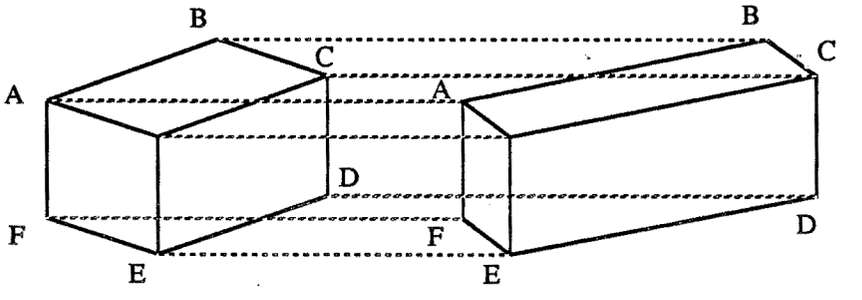
$$Z = \frac{a_{09}Ig + a_{10}J + a_{11}Id + a_{12}}{a_{13}Ig + a_{14}J + a_{15}Id + 1}$$

Quels que soient les repères choisis, la fonction F ($(Ig, J, Id) = F(X, Y, Z)$) de calibrage perspectif direct, et la fonction G de calibrage perspectif inverse ($(X, Y, Z) = G(Ig, J, Id)$) peuvent être calculées à partir d'un modèle homographique en utilisant un minimum de cinq points de contrôle. Cette propriété est tout à fait générale et ne dépend pas de la qualité ni de la spécificité des repères image et scène utilisés.

La figure C.14 montre un montage expérimental permettant d'obtenir un tel quintuplet de points. Tout comme dans le cas du calibrage d'une caméra unique, les paramètres du modèle de caméra sont obtenus en appliquant une méthode de résolution de systèmes linéaires, comme la méthode de Gauss ou celle des moindres carrés, selon le nombre de points de contrôle observés.

Il faut insister sur un des avantages à avoir un modèle de calibrage direct approché. Alors que le modèle du calibrage inverse est souvent difficile à calculer,





$$X = \frac{a01\ Ig + a02\ J + a03\ Id + a04}{a13\ Ig + a14\ J + a15\ Id + 1}$$

$$Y = \frac{a05\ Ig + a06\ J + a07\ Id + a08}{a13\ Ig + a14\ J + a15\ Id + 1}$$

$$Z = \frac{a09\ Ig + a10\ J + a11\ Id + a12}{a13\ Ig + a14\ J + a15\ Id + 1}$$

Figure C.14: Acquisition de points de contrôle pour le calibrage simplifié

notre modèle de calibrage est bijectif (et donc forcément involutif). On montre facilement que le calibrage inverse peut s'exprimer sous la même forme que le calibrage direct, et que les paramètres inconnus du calibrage inverse s'expriment linéairement en fonction des paramètres du calibrage direct. Cette dernière remarque est très importante : étant donnée la connaissance de l'un des deux modèles de calibrage, le second peut être 1/ soit calculé avec les mêmes points de contrôle, 2/ soit analytiquement déduits des valeurs des paramètres connus. Une dernière conséquence est que, vu le caractère involutif des modèles, il est donc possible de déterminer dynamiquement une erreur de chacun des modèles.

6.4 Localisation des Objets

Il n'y a plus beaucoup de choses à dire sur la localisation des objets. La mise en correspondance s'effectuant au niveau INDICES DE SCENE en assurant une décentration totale, elle assure bien une "localisation des objets". Plus généralement, si on dispose (stéréovision simple) à tout moment des deux modèles de calibrage direct et inverse, il est possible, à chaque instant, de déterminer le lieu de l'espace où se trouve une correspondance entre indices de niveau de représentation quelconque. Inversement, la connaissance de ces modèles permet toujours de diriger la reconnaissance d'un objet dont on connaît la localisation approximative.

Chapitre C.IV

Application à la Manipulation des Fils Electriques

Ce chapitre présente l'application industrielle sur laquelle nous avons expérimenté certaines des recherches théoriques présentées précédemment, et tout particulièrement au sujet de la couleur et de stéréoscopie simple.

Le problème qui nous a été soumis peut être résumé par la question : "comment identifier et localiser des brins d'un même câble électrique présentant une éventuelle différence de couleur en vue de leur préhension sélective?".

Outre le fait que la réalisation de cette application industrielle nous ait une nouvelle fois permis d'implanter avec succès plusieurs niveaux dans le processus de compréhension de la scène observée, elle a en outre assis notre thèse selon laquelle il est nécessaire de distinguer entre indices d'image et indices de scène, même dans les cas a priori les plus simples; ceci est particulièrement probant dans le cas de fils qui apparaissent se croiser dans l'image.

La présentation de cette application industrielle est l'objet spécifique du second paragraphe. Cependant, dans le souci de restituer l'application industrielle dans son contexte plus général (auquel sont consacrés les premier et dernier paragraphe), nous présentons au cours des paragraphes 3 et 4 les traitements ultérieurs dont les fils font l'objet, après avoir été localisés par le système de vision.

1 Le problème

La fabrication de nombreux sous-assemblages électriques nécessite l'emploi de fils électriques isolés. En collaboration avec la Compagnie des Machines Bull, le Laboratoire LIFIA s'est tout particulièrement intéressé à l'étude du montage de plusieurs fils électriques sur un ou plusieurs connecteurs. Chaque fil peut être libre (unifilaire), torsadé avec un ou deux autres fils (multifilaire), ou faire partie d'un ensemble (câble). Dans tous les cas, chaque fil doit être dénudé puis, soit soudé, soit serti et inséré dans un connecteur. A l'heure actuelle, certaines de ces opérations peuvent être effectuées automatiquement par des machines spécialisées qui par exemple dénudent et sertissent, mais toutes les autres manipulations, et spécialement celle qui consiste en l'alimentation des machines en fils électriques (opération de "repérage - saisie - amenée") et celle de l'insertion des fils électriques dans les connecteurs (opération d"'insertion") restent assurées par un opérateur humain. L'opérateur saisit par exemple un "brin" d'un câble, et l'amène à la machine à dénuder; ou encore, il saisit un brin déjà serti et l'insère avec précaution dans l'alvéole adéquate du connecteur. Toutes ces opérations sont itérées plusieurs milliers de fois par jour. De plus, outre le "contrôle de puissance" qui doit être effectué sur tous les ensembles câbles connecteurs fabriqués, l'insertion manuelle conduit à effectuer une opération supplémentaire de "contrôle de bonne connexion" pour vérifier qu'un fil serti est correctement positionné dans les deux connecteurs auxquels il est lié. Dans ce contexte, le laboratoire LIFIA a tout particulièrement étudié les points suivants :

La Localisation et L'Identification de Fils
(phase de repérage-saisie)

La Saisie et le Dénudage-Sertissage des Fils
(phase d'amenée et de dénudage-sertissage)

L'Insertion des Fils dans un Connecteur
(phase d'enfichage)

Ces opérations ont été étudiées en vue de leur intégration au sein d'une cellule robotisée automatisant la production des différents types d'ensembles câbles connecteurs.

2 La Localisation et l'Identification

Le poste de "repérage - saisie" soulève un problème essentiel, celui de pouvoir saisir un brin parmi plusieurs, et dont on ne connaît pas la position exacte. L'exemple type traité est celui d'un câble constitué de plusieurs brins, qui ne sont libres que sur une longueur réduite (3, 4 ou 6 cms). Tous les brins doivent être saisis un à un, avant qu'ils ne puissent être l'objet de traitements ultérieurs. Il faut donc disposer d'un système quelconque (système de proximité ou système de vision par exemple) permettant d'identifier un brin parmi plusieurs, et de le localiser dans l'espace avec suffisamment de précision pour qu'un robot puisse venir le saisir.

2.1 Principes de la Méthode

Nous disposons d'un câble composé de plusieurs fils et nous voulons à la fois pouvoir distinguer entre ces fils et localiser chacun d'entre eux dans l'espace réel. Le but principal du système que nous avons développé [DEMAZ-85a] est la reconstruction du modèle tridimensionnel de la scène. Utilisant la stéréovision (la localisation des fils sous-entend une information tridimensionnelle) et la vision couleur (l'identification des fils nécessite une information de couleur), la mise en correspondance entre les interprétations (indices de scène) des indices d'image (obtenus par segmentation) fournit une telle description. Un résultat auxiliaire du système est le recouvrement d'informations de prise pour le robot, qui peuvent aisément être extraites à partir de la description tridimensionnelle de la scène.

Les fils électriques sont des objets flexibles : ils forment souvent un bouquet (dont le modèle géométrique ne peut être parfaitement défini a priori) à l'extrémité du câble. C'est pourquoi les systèmes de vision classiques pour la reconnaissance d'objets rigides ne conviennent pas pour le problème des fils. De plus, des contraintes de fabrication telles que "fil rouge pour alimentation et fil noir pour la masse" nécessitent l'emploi d'un système de vision

2.1.1 La Stéréovision

La vision stéréoscopique (ou stéréovision) nous permet d'obtenir des informations tridimensionnelles. Nous avons utilisé le capteur de vision stéréoscopique développé au laboratoire LIFIA [DEMAZ-85b], décrit au début de cette partie (cf. figure C.02). Le calibrage homographique perspectif qui lui est associé a été décrit avec précision dans le chapitre précédent. Aussi, nous ne reviendrons donc pas sur ces deux points.

Notons simplement que l'expérimentation a montré qu'avec l'utilisation d'un tel capteur, l'erreur relative sur un point calculé de l'image (resp. de l'espace) est inférieure à 0.5% (resp. 0.2%). De telles performances sont compatibles avec la plupart des applications industrielles, et en particulier pour le problème de la localisation des fils.

2.1.2 La Couleur

La vision couleur est nécessaire pour résoudre les problèmes d'identification et accélère le processus de mise en correspondance entre les images d'une paire stéréoscopique.

Dans un premier temps, les images de couleur ont été obtenues à l'aide de filtres de couleurs (série trichrome KODAK WRATTEN), puis par illumination directe avec les couleurs primaires (R,V,B). Si le but était d'attacher un nom de couleur à chaque fil électrique observé, ces techniques poseraient un problème délicat, mais dans le cas des fils, les deux images sont saisies simultanément avec une seule caméra, et les couleurs sont donc dégradées de la même façon dans les deux images; cette propriété permet d'exploiter l'information de couleur dans le processus de mise en correspondance.

2.2 Extraction des Indices d'Image

La méthode utilisée est fondée sur l'utilisation de base du système CAICOU. Rappelons-nous que ce système permet de définir incrémentalement des indices de couleur tout autant que des lignes de contraste. L'activation de l'opérateur "droite double de couleur" correspond à l'activation simultanée des opérateurs "droite de contraste " et "ligne de couleur", et fournit une ou plusieurs lignes qui sont des lignes droites, et qui possèdent chacune une valeur de couleur. Typiquement, à l'occasion d'un suivi double d'un fil bicolore jaune - vert, le résultat de l'extraction est une suite de segments de droite, alternativement jaune puis vert.

2.2.1 Initialisation

Chaque câble est supposé se présenter dans le champ de vision suivant une orientation qui est toujours à peu près la même. Cela nous permet de guider la recherche initiale en affirmant que pour une ordonnée J donnée, une ligne horizontale intersecte chaque fil présent dans les deux images. C'est ainsi que nous utilisons un opérateur de CAICOU qui détecte les points de contraste à partir d'un pixel donné et suivant une direction donnée. Le résultat de cette initialisation est un nombre pair de points de contraste (cf. figure C.15.a).

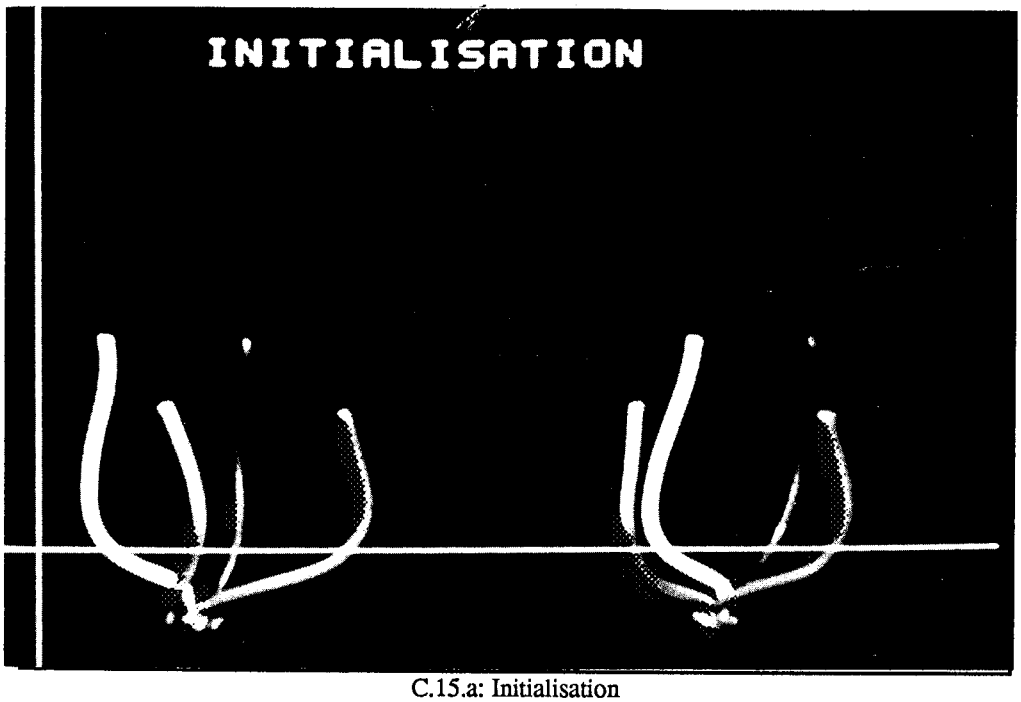
2.2.2 Suivi Double avec Extraction d'une Droite de Couleur

Cet opérateur suit deux lignes de contraste et calcule incrémentalement les valeurs de couleur de la région délimitée. Le suivi s'arrête lorsqu'il n'y a plus de contraste, ou lorsque les contraintes de distance entre les deux lignes de contraste sont dépassées. En appliquant cet opérateur à partir des paires de points de contraste détectés lors de l'initialisation, nous obtenons un contour qui est censé au niveau INDICE D'IMAGE, représenter un fil (cf. figures C.15.b et C.16.a).

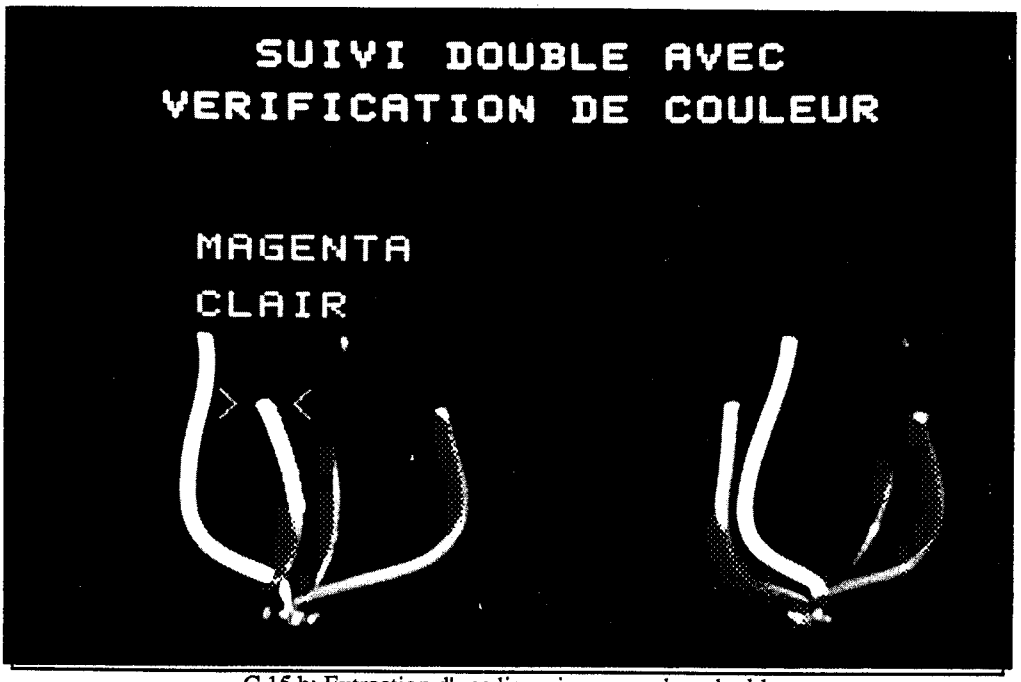
2.2.3 Analyse de l'Extrémité d'un Fil

Cet opérateur détecte les points de contraste le long d'un carré centré autour d'un pixel donné P. Lorsqu'un suivi double s'est interrompu, cette analyse locale autour des points d'arrêt fournit un nombre de points de contraste correspondant aux solutions suivantes :

1. si deux points seulement sont détectés, les points d'arrêt représentent une extrémité libre d'un fil.
2. si plus de deux points sont détectés, cela signifie que :
   - soit le fil croise d'autres fils. Dans ce cas, une interprétation locale de l'arrangement des droites-image détectées doit permettre de découvrir où le fil se poursuit.
   - soit le fil rejoint le câble; dans ce cas, la détection de l'extrémité libre du fil peut être obtenue en inversant la direction du suivi double.

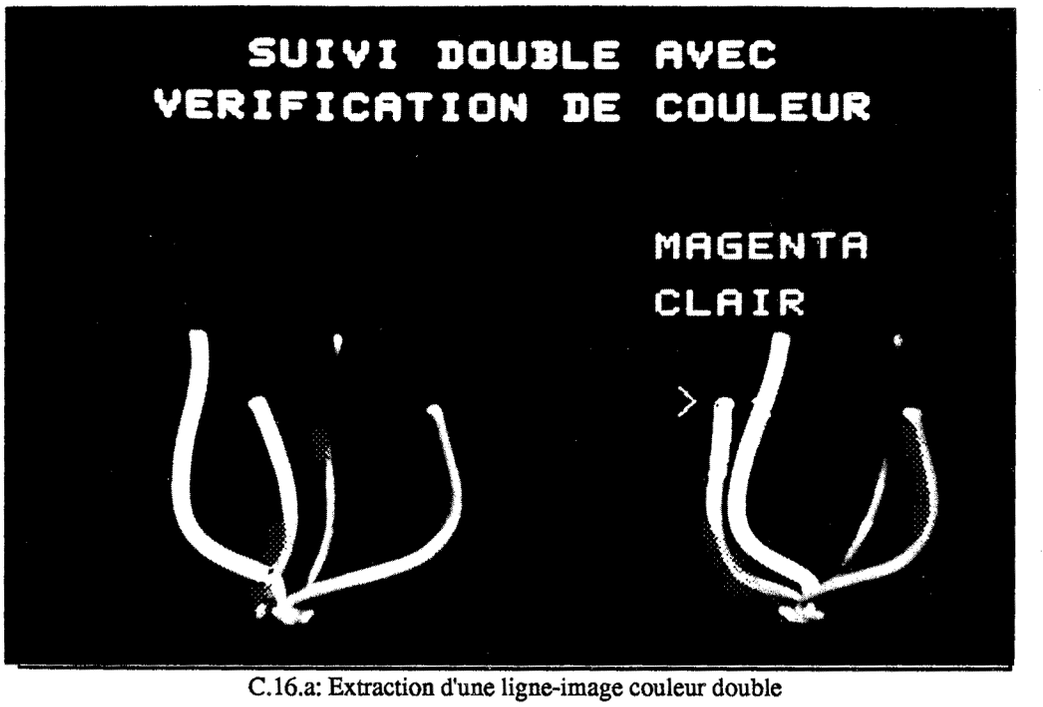


C.15.a: Initialisation

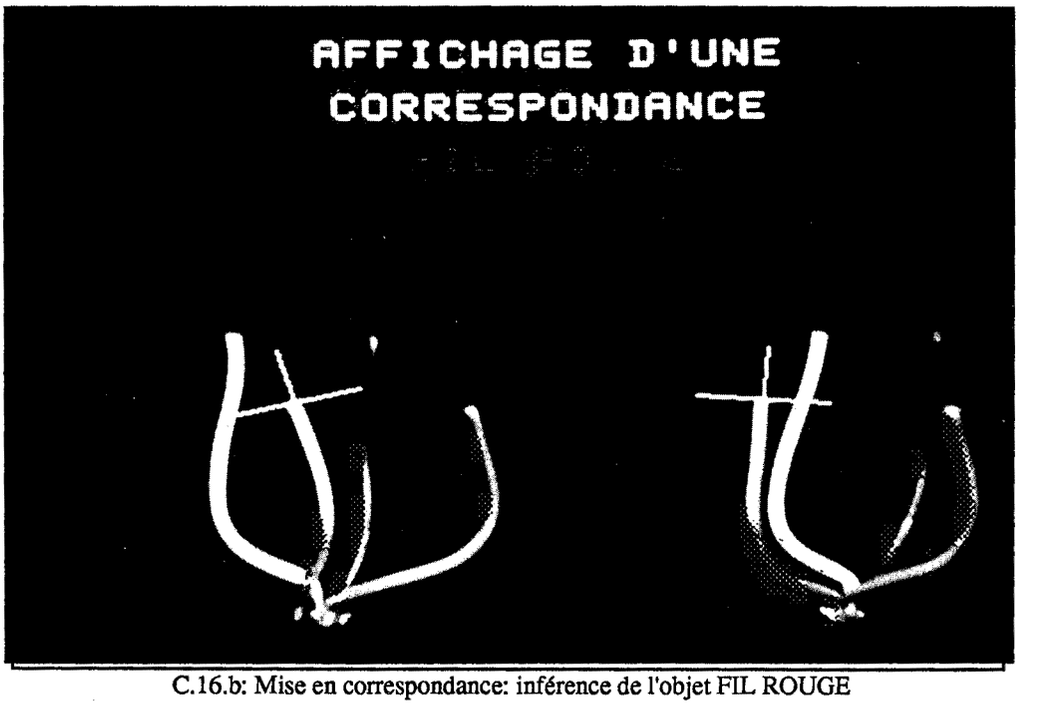


C.15.b: Extraction d'une ligne-image couleur double

Figure C.15: Localisation et Identification de Fils: résultats expérimentaux (1)



C.16.a: Extraction d'une ligne-image couleur double



C.16.b: Mise en correspondance: inférence de l'objet FIL ROUGE

Figure C.16: Localisation et Identification de Fils: résultats expérimentaux (2)

- soit quatre points ont été détectés, ce qui signifie que le suivi double s'est interrompu à cause d'un manque de contraste; la détection de l'extrémité libre du fil peut être obtenue en initialisant un nouveau suivi double à partir des deux nouveaux points détectés.

2.3 Interprétation des Indices d'Image en Indices de Scène

Dans le meilleur des cas, c'est-à-dire lorsqu'aucun fil n'en croise un autre, chaque suite de segments de droite avec ses attributs de couleur représente un fil, et il est alors tentant de ne pas tenir compte de la distinction entre indices d'image et indices de scène. Pourtant, lorsque deux fils se croisent, et cela arrive fréquemment dans des images quelconques, il est nécessaire de décider quel fil est devant l'autre, afin d'inférer des structures d'indices de scène qui traduisent les relations d'adjacence convenables vis-à-vis de la réalité de la scène.

2.4 Mise en Correspondance et Reconstruction de Formes

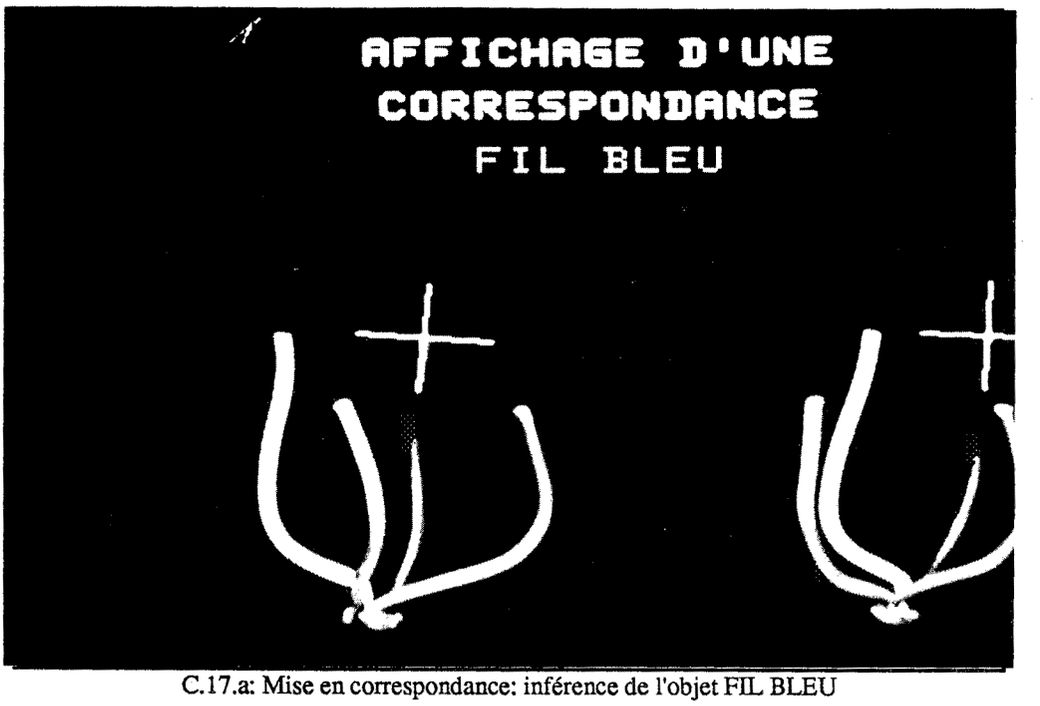
La troisième étape de notre analyse, qui concerne la reconstruction de la forme tridimensionnelle des fils, est assurée par une mise en correspondance entre les indices de scène des deux images, à partir du moment où chaque projection de fil présente dans l'une des images est associée à une projection de fil de l'autre image. Ceci est réalisé grâce à des contraintes de couleur, c'est-à-dire des comparaisons entre les valeurs de couleur en termes de teinte, saturation et luminance et grâce à la contrainte géométrique (fournie par le capteur stéréoscopique) sur la coordonnée J de l'extrémité du fil (l'extrémité du fil est observée à la même ordonnée dans les deux images).

Une fois la mise en correspondance effectuée, chaque fil de l'espace réel est associé avec deux structures (une par image). Il est alors facile de déduire la forme tridimensionnelle des fils en utilisant la fonction de calibrage perspectif inverse (cf. figures C.16.b, C.17.a et C.17.b).

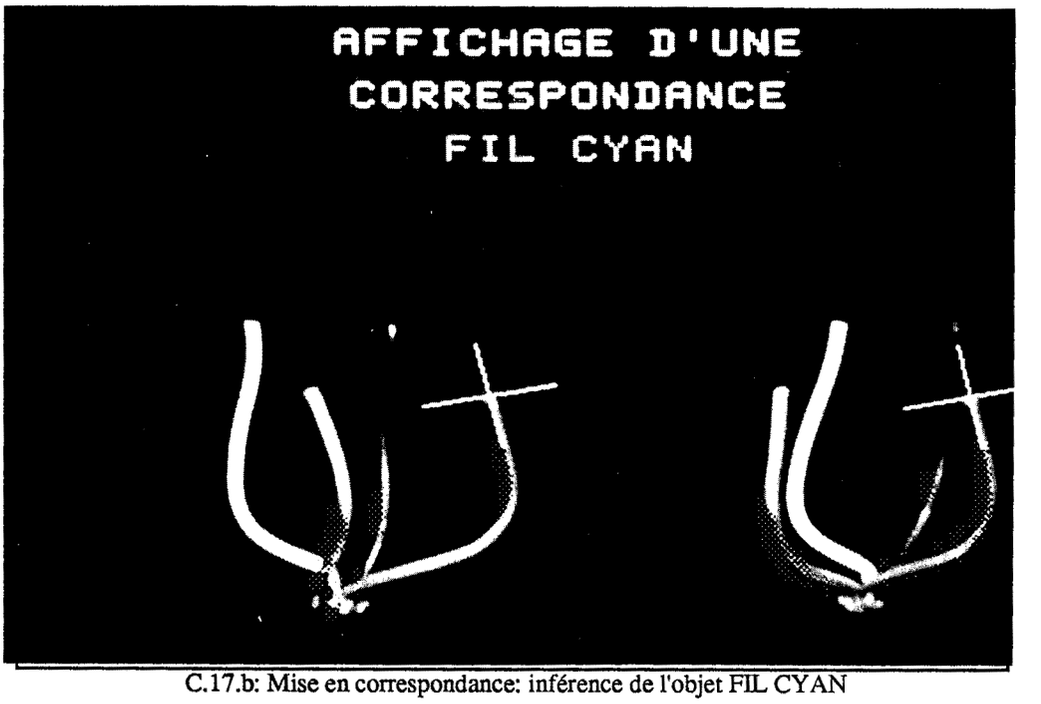
2.5 Détection d'un point et d'une direction de prise

La dernière étape consiste à récupérer des informations tridimensionnelles pour le robot : point de prise, direction de prise, point extrémité et direction extrémité. Bien que le modèle de la forme tridimensionnelle du fil contient implicitement toutes ces informations, il est clair que cette phase est tout à fait spécifique à l'application envisagée puisque les opérations qui doivent être effectuées ultérieurement par le robot concernent l'extrémité des fils, de sorte que le robot ne peut pas directement saisir ces fils à leur extrémité. Bien au contraire, il doit les prendre en des points situés à une distance donnée de leur extrémité. Dans le cas général, les quatre informations sont requises par le robot quelle que soit l'opération postérieure. Si cela paraît évident pour les informations de position, la nécessité de l'information d'orientation mérite quelques explications.

- Au point de prise, le fil doit être saisi entre deux mors qui soient parallèles; dans



C.17.a: Mise en correspondance: inférence de l'objet FIL BLEU



C.17.b: Mise en correspondance: inférence de l'objet FIL CYAN

Figure C.17: Localisation et Identification de Fils: résultats expérimentaux (3)

tous les autres cas, la saisie du fil provoque une déformation du fil en rotation et détruit la connaissance acquise sur le point extrémité. Notons toutefois qu'à la rencontre d'un tel problème, il est toujours possible d'utiliser un "feed-back" visuel (permis par le calibrage bijectif du système de vision) pour de nouveau préciser la position et l'orientation de ce point extrémité (contrôle du niveau OBJET sur le niveau INDICES D'IMAGE.
o A l'extrémité du fil, la connaissance de l'orientation du fil permet d'insérer le fil dans la machine à dénuder-sertir suivant une direction tangente au fil en son extrémité.

Une autre méthode (et c'est celle que nous avons utilisée pour l'expérimentation car la machine à dénuder-sertir, une PP3, nécessitait un accès frontal) consiste à rechercher un segment de droite à partir de l'extrémité du fil (par un "suivi double") qui soit assez long par rapport à la longueur minimale requise pour les manipulations ultérieures effectuées par le robot. En outre, la position exacte de l'extrémité du fil peut être précisée à partir d'une analyse locale ("suivi simple"), dans chaque image : elle permet d'extraire le point de contraste qui décrit le fil et qui est le plus éloigné dans la direction du segment de droite découvert.

2.6 Algorithmique de Contrôle

Tel que nous l'avons présenté, un algorithme de contrôle global du processus de repérage peut être présenté simplement comme suit :

1. initialisation
2. extraction de la totalité des indices d'images par enchaînement des primitives de "suivi double avec extraction d'une droite de couleur" et d"'analyse de l'extrémité d'un fil"
3. interprétation des indices d'image en indices de scènes
4. mise en correspondance (contraintes de couleur et de géométrie)
5. *pour un ou plusieurs fils sélectionnés*
   - reconstruction de la forme tridimensionnelle du fil
   - détection d'un point et d'une direction de prise

Cet algorithme peut bien sûr être amélioré en fonction des contraintes de production; c'est d'ailleurs dans ce sens que d'autres algorithmes optimisés et mieux adaptés ont été développés [DEMAZ-85a].

2.7 Configuration Logicielle et Materielle

Le système CAICOU est écrit en FORTRAN sur un LSI 11/23.
Les autres procédures sont écrites en MACLISP sur un HB 70.
Le matériel de traitement d'images utilisé est un GRINNELL couleur GMR 274.

3 La Saisie et le Dénudage-Sertissage

A partir des coordonnées du point de prise et de celles de l'extrémité d'un fil, le programme calcule une trajectoire d'approche pour la prise du fil, de façon à dégager la longueur nécessaire à l'insertion du fil dans la machine à dénuder-sertir. Une première *phase de calibrage* sert à décrire les positions relatives de la machine et du robot chargé de la manipulation des fils, ainsi qu'à calibrer le robot et le système de vision stéréoscopique. Il est en effet nécessaire de calibrer le système de vision relativement au robot car c'est lui qui fournit effectivement les informations tridimensionnelles concernant les fils à manipuler. La seconde phase est la *phase de manipulation des fils* proprement dite.

3.1 Modélisation et Calibrage

La phase de calibrage est assurée par des déplacements manuels de la pince du robot à des endroits précis utilisés comme référentiels de base. La construction des modèles de calibrage est alors réalisée sur la base d'un modèle géométrique partiel de l'univers contenant et la pince, et la machine à dénuder-sertir, et le système de vision.

Le modèle de la pince est décrit par un repère lié au robot.

Deux repères sont associés au modèle de la machine à dénuder-sertir : l'un décrit la position du palpeur qui déclenche le mécanisme de sertissage de la machine PP3, l'autre correspond à la position où la pince n'a besoin que d'un déplacement rectiligne donné pour amener l'extrémité du fil au palpeur.

Quant au modèle du système de vision, il est associé à un repère unique, déterminé par un objet étalon que le robot tient entre les mors de sa pince et qu'il présente au système de vision. L'objet utilisé est un parallélépipède dont les caractéristiques sont connues du système de vision. Il sert d'étalon pour créer un repère de référence commun au robot et à la vision. Le système de vision se calibre par rapport à cet objet comme cela a été décrit dans le chapitre précédent. Le robot, quant à lui, tient l'objet de calibrage entre les mors de sa pince, et est donc capable d'en déduire le repère associé au système de vision.

Les modèles de calibrage sont constitués d'une suite de mouvements (translations et rotations), associés aux positions courantes de la main du robot lorsqu'elle décrit les repères importants des objets.

Une fois ces modèles déterminés, il est établi des zones d'interdiction (correspondant à des butées logicielles) pour la sécurité de la maquette et dans lesquelles le robot ne pourra pas se déplacer.

3.2 Manipulation des Fils

Les opérations de prise, d'amenée et d'insertion des brins représentent un intérêt particulier pour la Robotique, du fait de la manipulation "d'objets mous" comme

les fils.

La méthode utilisée pour effectuer ces différentes opérations est décrite dans l'algorithme qui suit. Certaines actions sont directement issues des résultats de l'expérimentation. Le cycle, effectué pour chaque fil à traiter, comprend l'exécution séquentielle des opérations suivantes :

Attente de l'achèvement de la tâche d'identification et de localisation des brins par le système de vision.

Réception des coordonnées fournies par le système de vision, c'est-à-dire des informations sur le point de prise et sur l'extrémité du fil à dénuder et sertir. A la limite, seules les positions de ces points suffisent au robot : effectivement, l'accès frontal à la machine PP3 impose un fil relativement droit à son extrémité et cette propriété peut être vérifiée par le système de vision, comme cela a été décrit au paragraphe C.IV.2.. Une trajectoire d'approche peut éventuellement être requise.

Prise du fil. Si la trajectoire a été fournie, le programme l'utilise en vérifiant uniquement qu'elle ne traverse pas la zone d'interdiction créée dans la phase de calibrage. Dans le cas où la trajectoire d'approche n'est pas fournie, le programme doit la calculer et la vérifier. Dans les deux cas, la pince se referme en emprisonnant le fil de sorte que le robot puisse présenter l'extrémité libre de ce fil à la machine à dénuder-sertir (cf. figure C.18.a).

Amenée du fil à la machine à dénuder-sertir. Pour ce qui est de la programmation des mouvements du robot, il apparaît que le transport d'objets est une tâche délicate, en particulier lorsqu'ils sont flexibles. D'autre part, la prise des fils et l'accès frontal de la machine demandent des configurations complexes pour les articulations du robot. Ces différentes remarques expérimentales nous ont conduit à décomposer les mouvements à effectuer en mouvements élémentaires sans contrainte.

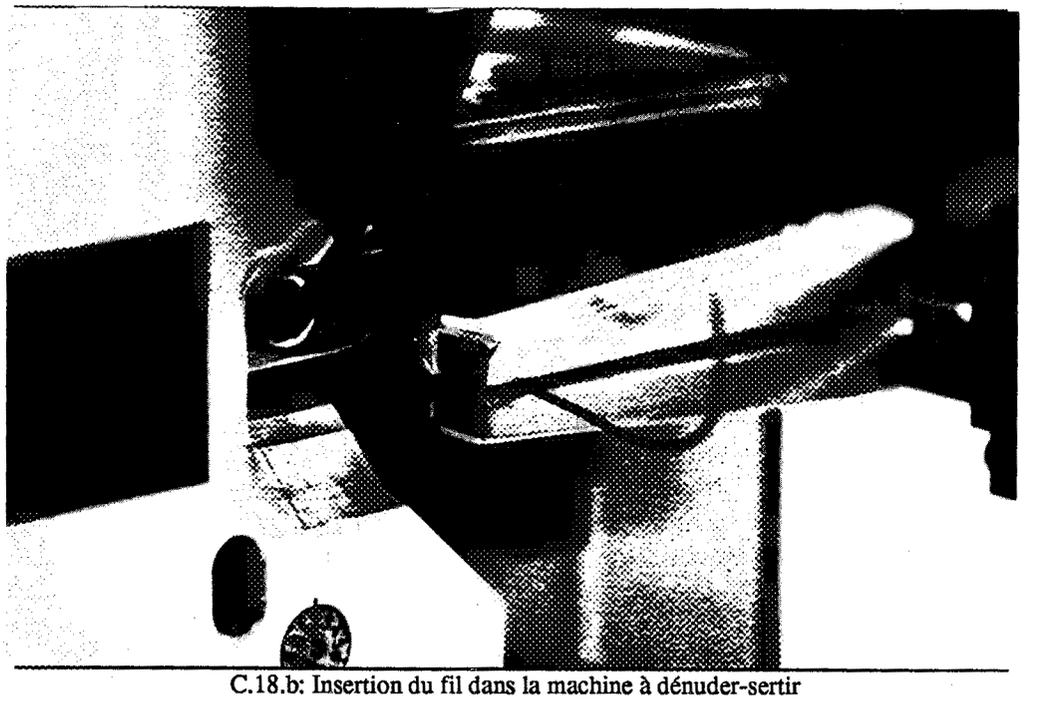
Insertion du fil dans l'ouverture de la machine à dénuder-sertir. La précision du calibrage de la machine assure une insertion parfaite, à condition que le fil soit correctement pris et positionné. Cette opération est un mouvement réflexe : le contact du fil sur le palpeur déclenche le mécanisme de dénudage et sertissage de la machine. Le fil serti est alors éjecté latéralement. La pince dégage le fil en suivant le mouvement latéral de l'éjection, en évitant ainsi un éventuel nouveau contact du fil serti avec le palpeur (cf. figure C.18.b).

Libération du fil. Cette opération consiste sinon à déposer le fil dans un vrac temporaire (unifilaire), du moins à lâcher le fil.

3.3 Configuration Matérielle et Logicielle

Le robot utilisé est un SCEMI 6 axes, programmé en langage LM [MAZER-85] et piloté par un LSI 11-23.

C.18.a: Prise d'un fil dont les coordonnées ont été fournies par le système de vision



C.18.b: Insertion du fil dans la machine à dénuder-sertir

Figure C.18: Saisie et Dénudage-Sertissage de fils: résultats expérimentaux

4 L'Insertion dans un Connecteur

Le poste d"'enfichage" permet d'insérer les brins sertis dans les différentes alvéoles d'un connecteur. Le principal problème qui se pose pour cette phase consiste en l'insertion même du contact serti. Elle demande, pour certains types de contacts, un très bon positionnement du contact par rapport au connecteur, position qui n'est pas toujours donnée initialement. En particulier, la flexibilité des brins peut introduire une incertitude sur cette orientation. La solution adoptée par le Laboratoire LIFIA consiste en l'utilisation de capteurs de force à jauges de contraintes pour orienter le contact avant ou pendant son insertion. Un autre problème majeur lié à cette phase concerne l'accessibilité des alvéoles. Il est résolu par utilisation d'un mors spécialisé et grâce à un protocole d'ordonnancement de l'insertion des différents fils sertis.

Dans un premier temps, nous décrivons le matériel utilisé pour la réalisation de cette phase. En effet, la méthode suivie pour l'insertion des fils dans les alvéoles a été fortement conditionnée par les choix matériels que nous avons été amenés à effectuer. Puis nous détaillons les deux phases de la méthode utilisée.

Le second paragraphe concerne le calibrage entre robot et support de brins ainsi que le choix des couples brin-alvéole. En effet, à l'initialisation du processus, il s'agit de calibrer l'ensemble peigne-connecteur placé sur la table de travail. Suivant le choix interactif fourni par l'utilisateur des contacts qui doivent être enfichées et des alvéoles correspondantes du connecteur, le système détermine alors l'ordre dans lequel les contacts seront saisis sur le peigne et enfichés dans le connecteur.

Le dernier paragraphe a trait au traitement de chaque brin serti devant être enfiché dans le connecteur : chaque brin serti est saisi sur le peigne, orienté dans la bonne direction, puis enfiché dans l'alvéole qui lui a été assignée, sous le contrôle permanent du capteur de forces.

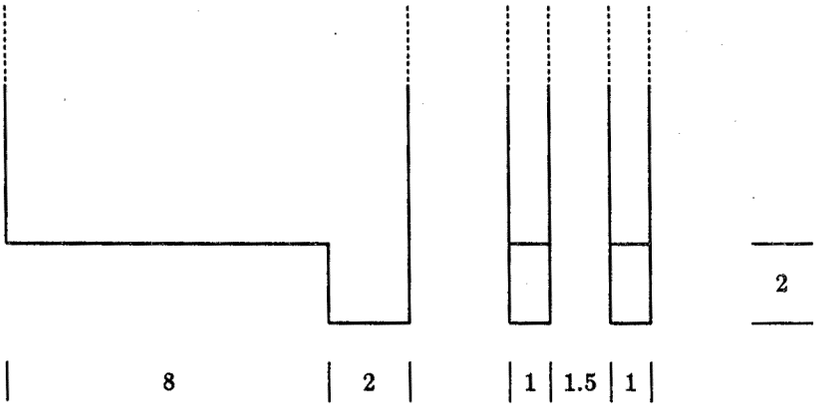
4.1 Description du Matériel

Robot et Calculateur Pour réaliser ce projet nous avons utilisé un robot SCEMI 6 axes muni de capteurs de forces, programmé en langage LM et piloté par un LSI 11-23.

Capteur de forces Le bras manipulateur est doté d'un capteur de forces RTL situé dans le poignet. Le capteur délivre huit tensions analogiques qui sont converties par une carte A/N en un vecteur V8 à huit composantes. A l'aide d'un moyennage (assurée par un calculateur LSI 11-02) sur plusieurs lectures des valeurs et par multiplication avec une matrice (8×6) caractéristique du capteur de forces, un vecteur à six composantes est obtenu, soit V6. Ce vecteur représente les forces Fx, Fy, Fz et les moments Mx, My et Mz délivrés par le capteur de forces. Ces six valeurs sont alors acheminées au LSI 11-23 par une ligne parallèle. Au niveau de l'expérimentation, il est à noter que le capteur possède une plus grande sensisibilité aux moments Mx et My qu'aux forces Fx, Fy et Fz, et fournissent de bien meilleurs résultats.

Pince La contrainte imposée par les faibles distances entre les alvéoles du connecteur (2.74mm entre les centres de deux alvéoles d'une même ligne, 2.84mm entre les centres des alvéoles de deux lignes successives) ainsi que la géométrie des contacts nous ont amené à dessiner une pince spécialement destinée à l'opération d'enfichage. L'extrémité de la pince que nous avons conçue (réalisant la préhension des brins) possède les caractéristiques suivantes :

- dimension de chacun des mors : $2mm \times 2mm \times 1mm$
- écartement maximum entre les mors : $1.5mm$



Peigne et Support Nous avons réalisé un peigne semi-circulaire dont le rayon extérieur présente des encoches destinées à recevoir des brins. Le peigne est solidaire d'un montage qui permet de recevoir un connecteur. Le support est fixé sur le peigne de telle sorte que les alvéoles du connecteur soient dans un plan vertical, et à environ 8mm derrière le centre origine du peigne.

4.2 Calibrage et Choix des Couples Brin-Alvéole

Pour réaliser l'enfichage des contacts dans les alvéoles d'un connecteur, il est nécessaire de disposer d'une précision de l'ordre de 0.1mm dans la modélisation des contacts, de la pince du robot et des alvéoles du connecteur.

Le calibrage du robot avec l'ensemble peigne-connecteur est nécessaire pour procéder à l'enfichage des contacts dans les alvéoles d'un connecteur. Il s'agit de déterminer, par rapport au référentiel du robot, un repère lié à l'ensemble peigne-connecteur. Nous avons donc neuf valeurs à déterminer avec une précision de 0.1mm. Pour pouvoir obtenir une telle précision, nous avons utilisé le capteur de forces du robot (première utilisation du capteur de forces). Compte tenu de la géométrie de la pince et du support, il nous a paru préférable lors du calibrage, de n'effectuer que des micro-déplacements dans les directions X et Z du repère de la pince. Le problème revient alors à déterminer la position de cinq points sur le support, avec comme intention de ne connaître pour chaque point que deux coordonnées sur trois. Pour chaque position à déterminer, la

pince est amenée à proximité du point, puis effectue des micro-déplacements de $0.1mm$ en X et Z jusqu'à ce qu'elle touche le support ($Fz > 2N$ pour un déplacement en Z, $My > 2Ncm$ pour un déplacement en X), ce qui lui permet d'acquérir les deux coordonnées recherchées.

Le choix des couples brin-alvéole du connecteur. Les brins sertis n'étant pas nécessairement placés sur le peigne dans l'ordre dans lequel ils seront enfichés, l'algorithme est réalisé de sorte que l'ordre des couples brin-alvéole donnés par l'utilisateur soit arbitraire. Une fois que tous les couples sont choisis, le programme les ordonne d'une façon telle que l'enfichage se fasse de gauche à droite en commençant par la ligne inférieure.

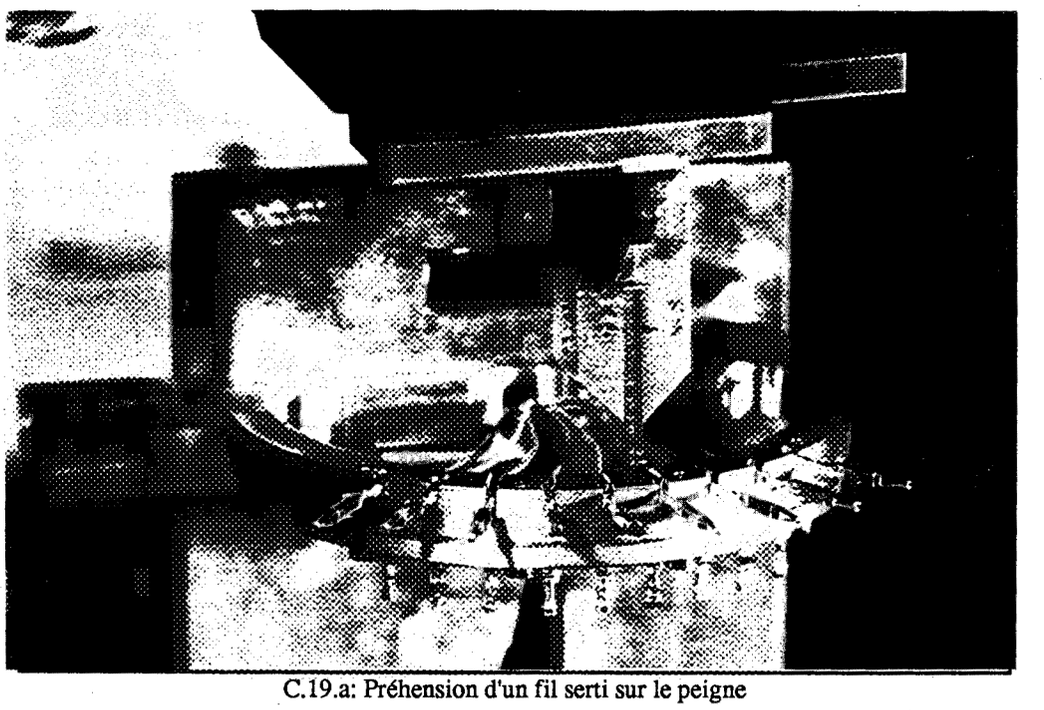
4.3 Description de la Manipulation des Brins Sertis

Préhension d'un brin sur le peigne. Au moment du sertissage d'un contact sur un brin, une section aplatie d'environ 3mm de longueur se forme sur le contact du côté du brin. Lorsque les brins sont placés dans les encoches du peigne, cette section est approximativement dans un plan vertical et son centre se trouve à environ 3.5mm du périmètre extérieur du peigne. C'est à cet endroit précis que le contact est saisi par la pince. Le brin serti est alors dégagé et transporté à une position intermédiaire (cf. figure C.19.a).

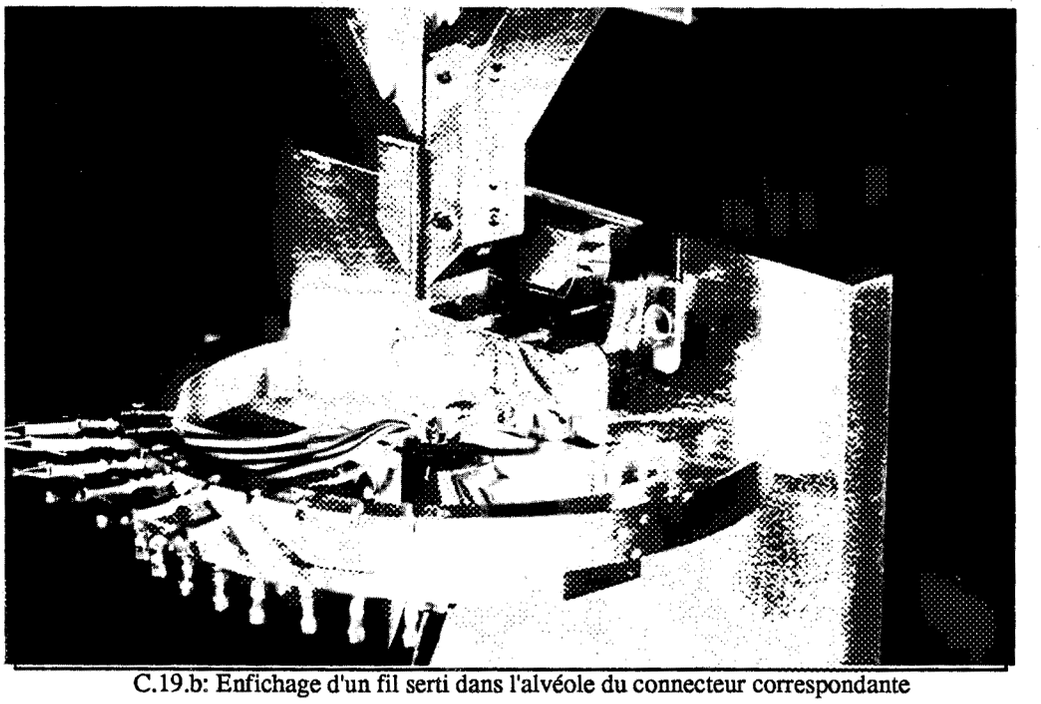
Détermination de l'orientation du contact serti à l'extrémité du brin. Dans un premier temps, nous avions choisi de tenter "directement" l'enfichage. En cas d'échec dû à une mauvaise orientation du contact dans la pince, nous procédions alors à la réorientation du contact puis revenions faire l'enfichage. Cette approche s'est avérée inefficace, car dans plus de 90% des cas le contact devait être réorienté. Nous avons donc décidé de procéder à la réorientation systématique de tous les contacts (seconde utilité du capteur de forces, après celle du calibrage). L'extrémité du contact est amenée à proximité du support du connecteur, et un mouvement de rotation autour de Z couplé à un mouvement de translation en X (jusqu'à ce que la pointe du contact touche le support), permet de déterminer la correction angulaire horizontale. De même, un mouvement de translation en Z couplé à un mouvement de translation en X à proximité de l'arête du support nous permet d'évaluer la correction verticale.

Notons toutefois que dans une phase industrielle, le caractère systématique de cette réorientation peut être évité par utilisation d'un préhenseur dont les mors seraient moulés en adéquation avec la zone des brins sertis que le robot vient saisir.

Enfichage du contact dans l'alvéole du connecteur. Une fois l'orientation déterminée, le contact est amené devant l'alvéole. Le robot procède alors à l'insertion en s'assurant d'abord que l'extrémité du contact pénètre bien dans l'alvéole, puis que le renflement du contact, à environ 5mm de l'extrémité, ne bute pas sur l'arête de l'alvéole (troisième et principale utilité du capteur de



C.19.a: Préhension d'un fil serti sur le peigne



C.19.b: Enfichage d'un fil serti dans l'alvéole du connecteur correspondante

Figure C.19: Insertion de brins dans un connecteur: résultats expérimentaux

forces). Lorsque les deux tiers du contact sont dans l'alvéole, la pince libère le contact et vient saisir le brin à environ 2mm derrière le contact pour terminer l'enfichage. Cette façon de procéder est nécessaire pour assurer une insertion complète et pour éviter une collision de la pince avec le support du connecteur. L'enfichage de ce contact étant achevée, la pince retourne chercher un autre contact à enficher et procède de la même façon jusqu'à montage complet du connecteur (cf. figure C.19.b).

5 Le Projet d'Intégration des Méthodes au sein d'un Prototype

Une étude a été menée en vue de l'intégration de ces différents postes de traitement des fils électriques au sein d'une cellule robotisée automatisant la production des différents types d'ensembles câbles connecteurs [DEMAZ-84b]. Les solutions proposées sont influencées par l'organisation actuelle de la fabrication et sont adaptées aux technologies des ensembles câbles-connecteurs aujourd'hui en vigueur et à leur évolution probable.

5.1 La méthode choisie

La méthode choisie, illustrée par la figure C.20, met en valeur les points suivants :

- les opérations d'AMENEE sont facilitées par l'utilisation de goulottes d'alimentation en unifilaires, multifilaires et câbles.
- les opérations de base sont le DENUDAGE, le SERTISSAGE (parfois simultané avec le DENUDAGE comme dans le cas de notre exemple courant) et l'ENFICHAGE, opérations qui s'effectuent toutes brin par brin (soit N le nombre de brins intervenant dans la fabrication de l'ensemble câble-connecteur).
- pendant la phase de REPERAGE-SAISIE, les différents brins constituant le futur câble-connecteur sont rangés dans un peigne, permettant de conserver la position de chacun, et éventuellement l'orientation du contact dans le cas d'un brin serti.
- mise à part l'opération de REPERAGE-SAISIE, l'enchaînement des opérations qui figure sur le schéma correspond à l'état actuel des étapes manuelles nécessaires à la fabrication des câbles-connecteurs. C'est-à-dire que tous les brins d'un même câble passent par le même poste de base, avant que le câble ne soit l'objet d'une autre opération.
- les opérations d'ENTREE-SORTIE sont facilitées par l'utilisation d'une chaîne de transfert circulaire, extérieure aux différents postes de travail. Cette chaîne est destinée à supporter les peignes (sur lesquels les brins sont rangés) et permet la présentation des brins l'un après l'autre, devant tous les postes de travail de base.

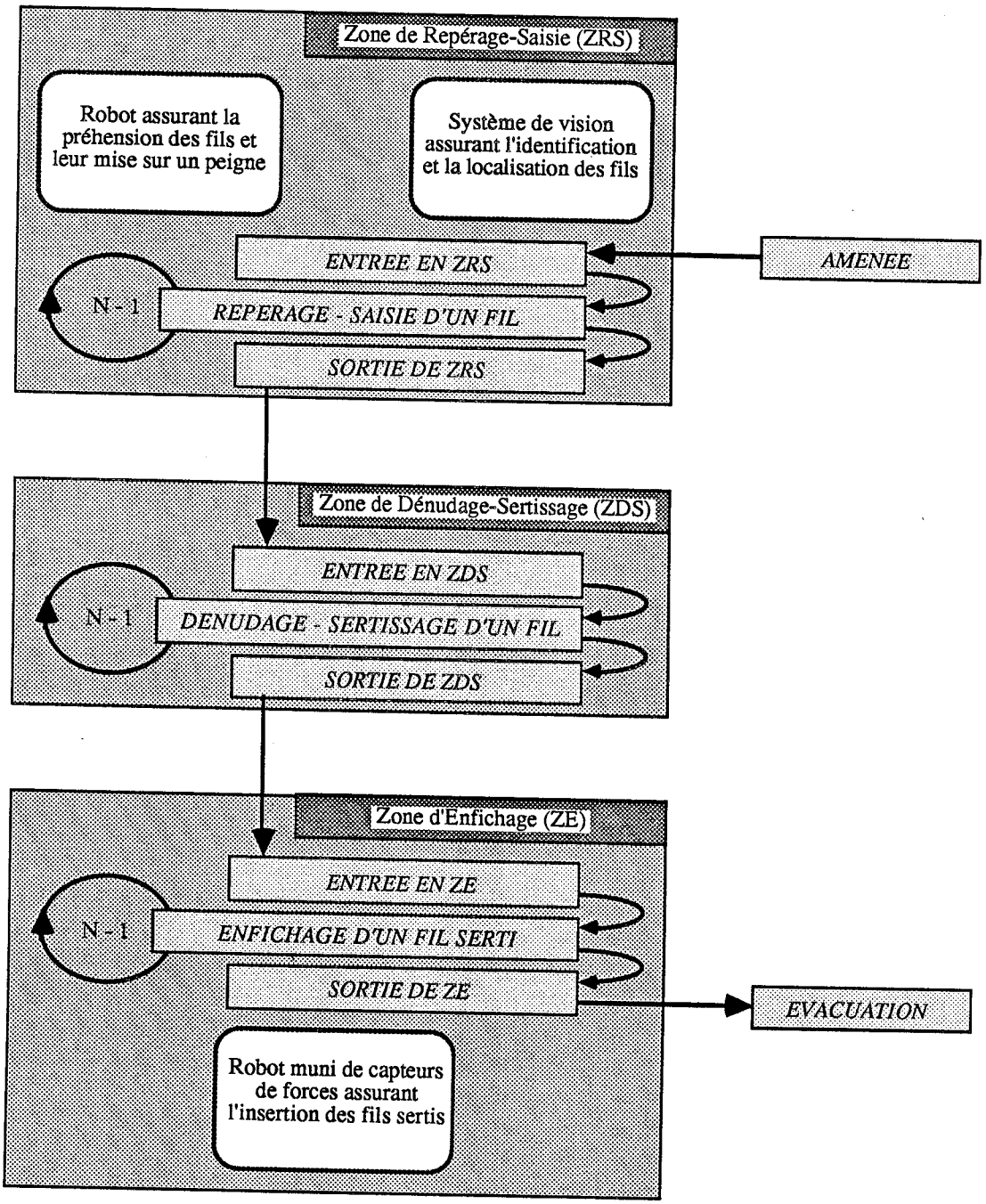


Figure C.20: Fabrication Assistée par Ordinateur d'ensembles câbles-connecteurs

5.2 Fonctionnement de la Cellule

1. Les différents brins, paires et tierces sont débités, et transportés automatiquement dans les goulottes.

2. Les brins (unifilaires) sont saisis et transportés au poste de repérage-saisie. Dans le cas de l'unifilaire, la position de la goulotte est connue, elle correspond à un brin qui a une position bien définie sur le peigne. Ainsi, il n'est pas nécessaire d'utiliser la vision; les brins sont donc saisis par le robot et insérés sur les encoches du peigne.

3. Les paires, les tierces et les câbles sont saisis et amenés au poste de repérage-saisie. Le système de vision identifie chaque brin, que le robot saisit après localisation pour installation sur le peigne.

4. Le peigne est alors entièrement constitué et est transporté au poste de dénudage-sertissage pour traitement des brins.

5. Le peigne est ensuite transporté au poste d'enfichage. Dans le même temps, le poste d'enfichage est alimenté avec un connecteur du type correspondant au câble-connecteur désiré. Les contacts sertis sont alors enfichés à leur place dans le connecteur.

6. Lorsque l'enfichage est terminé, l'ensemble est finalement évacué dans un vrac.

6 Conclusion

Nous avons présenté les trois phases qui ont été particulièrement étudiées au laboratoire LIFIA en vue de l'automatisation de la production des ensembles câbles-connecteurs. La réalisation de la phase de repérage-saisie est originale dans le sens où elle fait appel à un procédé de stéréovision couleur, principe totalement ignoré actuellement dans l'industrie. L'originalité des méthodes développées réside aussi dans l'usage multiple qui est fait d'un capteur de forces dans le cadre de l'insertion des brins sertis dans les alvéoles d'un connecteur.

Nous nous sommes efforcés de développer des techniques et des méthodes adaptables aux différents types de câbles-connecteurs fabriqués par la CMB, ainsi qu'à l 'évolution possible de la câblerie en général. Les solutions proposées aux problèmes soulevés sont complètes. Elles ne se cantonnent pas simplement à l'environnement immédiat du système de vision par exemple, mais ont trait à l'ensemble de la cellule.

Les résultats déjà obtenus nous confortent dans l'idée que la Robotique, et plus précisément l'utilisation de robots dotés de capteurs (capteur de vision pour l'opération de repérage-saisie, capteur de forces pour l'opération d'enfichage) peut permettre d'améliorer considérablement sinon le rendement, du moins la qualité de la fabrication des ensembles câbles connecteurs.

Conclusion

Si on pense que le but des études dans le domaine du traitement de l'information est de formuler et de comprendre des problèmes particuliers du traitement de l'information, c'est alors la structure de ces problèmes qui est le problème central et non les mécanismes à travers lesquels ils sont implantés.

David Marr (1977)

Conclusion

" Quels sont selon vous, parmi les différents points évoqués dans ce rapport, ceux qui constituent l'originalité du travail effectué? "

Nous avons voulu montrer que les systèmes de Vision par Ordinateur, quelles que soient les techniques utilisées, doivent distinguer plusieurs niveaux de représentation entre celui de l'image initialement acquise et celui de la description de scène. De notre point de vue, cinq niveaux de représentation sont nécessaires : les niveaux IMAGE, INDICES D'IMAGE, INDICES DE SCENE, OBJET et SCENE. En outre ils sont suffisants jusqu'au niveau OBJET, par rapport à l'état de l'art des méthodes développées en Vision par Ordinateur. Nous avons justifié la nécessité de l'existence de ces niveaux en regard de résultats obtenus dans les trois domaines de l'informatique, de la psychologie et de la neurophysiologie, et par des critères scientifiques et informatiques non antropomorphes. Nous croyons effectivement que les études effectuées dans chacun des domaines intéressés par les images et la vision se doivent de ne pas ignorer les résultats obtenus dans les autres disciplines. L'essai de formalisation que nous avons introduit repose sur la prise en compte des critères d'abstraction et de décentration. Mais si les noms de ces deux critères sont hérités de la psychologie, nous les avons redéfinis : c'est ainsi qu'une évolution en abstraction (resp. en décentration) implique un changement de type de représentation (resp. de type de référentiel).

Les cinq niveaux présentés n'ont pas tous été étudiés avec le même soin. Nous avons étudié plus particulièrement les niveaux intermédiaires INDICES D'IMAGE et INDICES DE SCENE. Cette étude a été effectuée à travers deux expérimentations fondées sur des méthodes d'inférence de formes analysant des caractéristiques différentes dans des domaines d'étude eux aussi différents. Les deux expérimentations illustrent et cautionnent la thèse défendue. Chacune d'entre elles s'est prolongée par

la résolution de problèmes industriels qui montrent l'intérêt pratique des résultats obtenus.

Notre étude de l'inférence de formes à partir des contours dans le domaine du monde des blocs (application à un problème industriel de localisation de paquets sur une palette) nous a permis de montrer que l'inférence d'indices de scène ne se réduisait pas à un simple étiquetage d'indices d'image. Cette expérimentation nous a en outre permis d'exhiber différents algorithmes de construction du niveau INDICES D'IMAGE.

Notre étude de l'inférence de formes à partir de la stéréovision simple couleur dans le domaine des objets filiformes (application industrielle d'identification et de localisation de fils électriques) nous a permis de prolonger la distinction entre INDICES D'IMAGE et INDICES DE SCENE dans un domaine où les surfaces du monde réel ne sont pas planes. Cette seconde expérimentation nous a de plus amené à exposer d'autres résultats connexes : la notion de "stéréovision large" et la "modélisation d'un capteur stéréoscopique adapté à une coopération bras-oeil efficace", ainsi que la possibilité d'une mise en correspondance stéréoscopique entre INDICES DE SCENE contribuant à atteindre le niveau OBJET.

" Quelles réserves émettez-vous à la validité des idées défendues? "

De nombreuses remarques viennent ternir le tableau dressé ci-dessus.

D'un point de vue théorique, la spécification et la formalisation des niveaux, des notions d'abstraction et de décentration, et surtout des procédures inter- et intra-niveaux restent faibles. Il est tout aussi évident que l'analogie au système humain est pour l'instant très limitée, et que les problèmes de communication avec d'autres systèmes intelligents, si ils ont été évoqués, restent inexplorés.

D'un point de vue plus expérimental, il est dommage que seuls les niveaux les plus bas aient été étudiés, au détriment des niveaux OBJET et SCENE qui, à terme, sont les plus importants. Même aux niveaux les mieux étudiés, les indices exhibés et les procédures envisagées pour les extraire sont loin d'être exhaustifs : ainsi les INDICES D'IMAGE considérés sont en faible nombre et, par exemple, ni les régions, ni les groupements perceptuels n'ont été pris en compte. Les INDICES DE SCENE en subissent les conséquences. Cet état des choses est en particulier dû au fait que l'expérimentation réalisée est en elle-même limitée : ne s'attacher qu'à deux caractéristiques seulement, et a fortiori dans des domaines d'étude restreints ne représente qu'une goutte d'eau dans la mer des combinaisons possibles entre méthodes d'inférence existantes et opérant dans un univers réel.

" Dans la première partie du rapport, vous exposez des idées théoriques et des tentatives de formalisation en ne faisant aucunement allusion à un contexte quelconque. A l'opposé, les expérimentations que vous avez menées, et a fortiori les applications à des problèmes industriels que vous avez réalisées, semblent fortement conditionnées par leur environnement. De quelle façon un système de Vision par Ordinateur doit-il être dépendant de connaissances spécifiques sur l'environnement dans lequel

il évolue? "

Cette question traduit bien pour nous toute l'importance du CONTEXTE au sein d'un système de vision par ordinateur. S'il est vrai que nous reconnaissons pleinement cette présence au niveau des expérimentations, il est aussi vrai que cette notion n'a malheureusement pas (encore) été bien appréhendée au niveau de la formalisation des niveaux. Par rapport à la "généralité" de nos idées "théoriques" relatives aux cinq niveaux de représentation que nous avons présentés, il apparaît clairement que nos deux expérimentations (et plus encore les applications industrielles correspondantes) présentent des restrictions, et tout spécialement quant à la limitation des domaines dans lesquels les systèmes sont utilisables.

- La première expérimentation, traitant l'inférence de formes à partir de contours, a été menée dans le domaine restreint des objets polyédriques. L'application industrielle s'y rapportant est limitée à l'inférence de formes parallélépipédiques.

- La seconde expérimentation, traitant de l'inférence de formes à partir de la stéréovision simple, permet d'extraire des objets filiformes, et dans le même temps, l'application industrielle se limite à traiter des fils électriques.

Au-delà des raisons temporelles, de la réduction de la combinatoire ou même de la commodité d'implantation, il est important d'expliciter les trois principaux avantages que nous trouvons à limiter le domaine dans lequel un système de compréhension de scène peut opérer : simplification de représentation, spécialisation des inférences et aide au contrôle. Des exemples de ces trois avantages sont globalement représentés (sans distinction de leur type) par la figure CONT et nous en expliquons ici certains dans le cas de notre première expérimentation :

SIMPLIFICATION DE REPRESENTATION La simplification des modèles de représentation se produit évidemment à tous les niveaux de représentation. Se limiter à des objets polyédriques permet, au niveau INDICE D'IMAGE, de ne représenter que des droites-image. De même, au niveau INDICES DE SCENE, ne figurent alors que les surfaces planes, tandis qu'au niveau OBJET, il suffit d'un modèle de représentation par sommets et arêtes.

SPECIALISATION DES INFERENCES La restriction au contexte d'objets polyédriques permet de spécialiser les règles d'interprétation des noeuds-image en indices de scène aux seules interprétations cohérentes avec une rétroprojection fournissant des surfaces planes du monde réel. C'est-à-dire que certaines inférences deviennent inutiles.

AIDE AU CONTROLE La restriction du contexte guide le contrôle du processus de compréhension de scènes. Sachant par exemple que l'on traite de scènes composées d'objets polyédriques, le système aura pour priorité absolue, dans son fonctionnement prédicatif, d'extraire des lignes-image droites plutôt que des lignes-image

DOMAINE	NIVEAU INDICES D'IMAGE	NIVEAU INDICES DE SCENE	NIVEAU OBJET
MONDE DES BLOCS	Les indices d'image linéaires à extraire sont des droites-image simples Les surfaces sont homogènes: pas de variation de contraste a priori	Restriction aux seuls opérateurs qui ne font intervenir que des indices d'image rectilignes Les surfaces extraites sont directement interprétées comme des faces du monde réel	Modélisation par sommets et arêtes
+ PENTAEDRES		+ Pour chaque noeud-image, restriction de l'ensemble des opérateurs d'interprétation du fait de la valuation maximale d'un sommet réel au plus égale à 3	+ Chaque objet est composé de cinq volumes identiques d'où des conditions d'orthogonalité et de parallélisme à respecter
+ PAQUETS	+ Limitation sur le type des noeuds recherchés (L,T et X)	+ Pour chaque noeud-image, limitation du nombre des opérateurs d'interprétation du fait de la coplanarité des arêtes constituant un sommet	+ Tous les objets à inférer sont identiques à une même forme parallélépipédique
OBJETS FILIFORMES	Les indices d'image linéaires à extraire sont des lignes-image doubles	La filiformité des objets limite les interpétations des indices linéaires à des contours virtuels d'objets gauches Les régions sont interprétees comme des surfaces cylindriques	Modélisation de chaque objet extrait par un unique cylindre généralisé Les objets sont directement inférés à partir des surfaces cylindriques extraites
+ FILS ELECTRIQUES	+ Limitation sur le type des noeuds recherchés (Tc, Xc)	+ Pour chaque noeud-image, limitation du nombre des opérateurs d'interprétation du fait de l'existence d'un unique sommet multivalué (correspondant à l'extrémité de l'indice décrivant le câble multifilaire)	+ Modélisation de chaque objet extrait par un cylindre dont la seule généralisation est la possibilité d'une génératrice gauche + Organisation connue des objets entre eux

Figure Cont: Avantages liés à une spécification du contexte

courbes. Et ce qui est vrai au niveau des indices d'image l'est aussi aux niveaux de représentation d'abstraction plus élevés, et est tout particulièrement sensible dans le cadre de la reconnaissance d'objets [BALLA-78].

Déduits de nos expérimentations, ces avantages ne sont cependant pas réellement dépendants des méthodes d'inférence de formes utilisées. Nous pensons en outre que les avantages à spécifier le contexte environnant, réduisent bien sûr un peu la complexité du problème posé par la compréhension de scènes, mais surtout sa combinatoire.

" Dans quelle mesure un système de Vision par Ordinateur doit-il adapter sa démarche et ses méthodes d'analyse et de compréhension quand son but premier devient la recherche d'un objet particulier (ou d'une configuration) dans son environnement? "

Notre interprétation de la question nous a conduit à situer la position de la RECONNAISSANCE, voire de la DESIGNATION par rapport à la conception globale d'un système avancé de compréhension de scènes. Avant de nous préoccupper du problème même de la reconnaissance, et pour finir d'insister sur l'importance du contexte, il est intéressant de noter combien, au niveau humain, ces deux notions sont liées.

CONTEXTE ET RECONNAISSANCE Cette indissociabilité est tout particulièrement sensible lorsqu'on s'intéresse aux performances du système visuel humain quant à la reconnaissance de visages. Rien ne justifie au premier abord que l'image d'un visage humain doive être traitée différemment des autres images. Et pourtant les psychologues lui accordent une attention particulière car, d'une grande complexité, le visage humain peut apparaître sous des formes avec des caractéristiques extrêmement variées. Rien n'est indiscutablement démontré, mais certains résultats d'expériences en psychologie montrent par exemple que la suppression du "contexte d'encodage" au moment d'une reconnaissance se traduit par une diminution de la performance mnésique. Ainsi, n'ayant aperçu une personne donnée qu'une seule fois dans sa vie dans un hall de gare, cette personne aura beaucoup plus de chances d'être "reconnue" que le "contexte d'étude" dans laquelle elle sera vue une seconde fois sera aussi un hall de gare : la reconnaissance est d'autant plus facile que le degré d'association entre le contexte de reconnaissance et le contexte d'encodage est élevé. Inversement, on peut supposer que ce ne sont pas les changements de contexte en tant que tels qui perturbent la reconnaissance, mais l'absence de relation entre le nouveau contexte de reconnaissance et l'ancien contexte d'encodage [TIBER-83]. La restriction à un contexte particulier permet donc de réduire une combinatoire à un point tel qu'il est alors permis de reconnaître des objets qui ne l'auraient pas été sinon. Cette performance est à relier directement à l'avantage d'"aide au contrôle" explicité précédemment. Plus encore, ces expériences en psychologie nous montrent que les "objets" vus par l'être humain ne sont pas seulement mémorisés isolément, mais gardent quelque part souvenir de leur contexte d'encodage, ce qui cautionne la thèse de représentations visuelles relationnelles.

RECONNAISSANCE Dans la plupart des systèmes de vision par ordinateur (le lecteur pourra par exemple se référer aux exemples de systèmes décrits par [BALLA-78] [BINFO-82] [BROOK-81] [SHIRA-75] [SHIRA-78] ou [SOUVI-83]), les informations extraites d'une image servent à des fins de reconnaissance sous la forme de l'identification et de la localisation d'objets connus, et les systèmes développés sont naturellement dirigés vers le but ("goal oriented" ou "model-based"). Disposant d'une image digitalisée d'une part, et d'une "banque de modèles" d'autre part, ces systèmes s'efforcent de construire une structure de l'image interprétée qui puisse être inférée au mieux avec les modèles, et de reconnaître les objets en scène par mise en correspondance géométrique entre formes extraites et modèles connus du système. Ceci dit, il est naturel (et ceci est caractéristique des systèmes de vision industriels) d'intégrer au plus tôt le plus de connaissances spécifiques que l'on peut détenir sur la scène à analyser. Cette prise en compte d'informations supplémentaires, généralement utilisée de manière prédictive, accélère notablement le processus de vision, et présente, là encore, tous les avantages liés à une spécification du contexte.

Il n'est pas toujours évident d'admettre les différences entre compréhension et reconnaissance. Car en effet, pourquoi comprendre des scènes contenues dans des images si ce n'est pour réutiliser les connaissances acquises lors d'analyses ultérieures, et qui donc feraient inévitablement appel à des activations de processus de reconnaissance?

o Selon nous, la "reconnaissance d'objets" est un processus "agent", ponctuellement activé de façon consciente par les formes supérieures de l'intelligence, et de ce fait tout naturellement dirigé vers la satisfaction d'un but précis, donc plutôt descendant. La délimitation des raisonnements entrant en jeu tient simplement au choix de telle ou telle variété d'actions ou d'opérations en vue du but à atteindre.

o Au contraire, la "compréhension de scènes" est un processus essentiellement "patient" (au sens où il est inéluctable), permanent, et plutôt ascendant, dont le but est, à tout moment, de construire une description symbolique : - géométrique, relationnelle, fonctionnelle et temporelle - de l'environnement dans lequel il évolue. Il n'intervient aucune décision d'activation des processus d'inférence, tous les facteurs en présence agissant concurremment puisque cette description évoluée de formes n'est alors pas librement composée mais découverte ou reconstituée en fonction des données objectives.

Nous considérons donc que le système de vision par ordinateur possède une double fonctionnalité : c'est 1/ un ensemble de processus d'acquisition qui fonctionnent en permanence et 2/ c'est un ensemble de processus dont la mise en activité intermittente correspond à la satisfaction d'un but précis.

DESIGNATION Comment se place la désignation d'objets pour un système de vision dont l'objectif premier est de décrire une scène en termes de volumes et de relations entre ces volumes? Comme nous l'avons dit, et d'un point de vue strictement visuel (et donc en dehors des raisonnements qu'elle induit), la désignation ne

constitue à notre avis de vue qu'un "plus", et n'intervient directement en première approche qu'au niveau objet, voire au niveau scène, même s'il est clair que le résultat d'une désignation permet d'engendrer des "rétroactions" sur les niveaux inférieurs. Même dans le cas simple où la désignation se réduit à associer un nom à une forme tridimensionnelle extraite, le processus, activé consciemment, fait appel aux fonctions les plus raisonnées d'un système de vision : utilisant une "base de modèles" associée à un vocabulaire, il nécessite un processus de mise en correspondance efficace [GRANG-85]. Dans ce sens où il est conscient, le problème de la désignation est à rapprocher de celui de la reconnaissance. Comme dans le cas de la reconnaissance, les modèles connus du système ne sont d'ailleurs pas forcément des modèles purement géométriques, munis d'attributs qui les caractérisent; l'acte de désignation fait appel à des connaissances qui ne sont pas seulement celles qui sont contenues dans une simple image, de nature a priori géométrique; des exemples de telles connaissances concernent autant par exemple la fonctionnalité des objets que le langage employé pour les décrire.

Une dernière observation se situe par rapport au niveau auquel la désignation est effectuée : à première vue, elle n'intervient qu'une fois la scène décrite à "un niveau d'abstraction suffisamment élevé". Cette dernière affirmation est pourtant à moduler, dans le sens où le but premier de la désignation est selon nous d'autoriser un dialogue évolué avec le monde extérieur. Pourtant il est aussi vrai que du point de vue simplement vision, le raisonnement sur les interprétations de l'image (et de la scène) peut se faire à tous les niveaux, tout dépend bien entendu de la nature des connaissances utilisées. Ainsi Garvey raisonne-t-il au niveau IMAGE [GARVE-76]. De même, Gordillo raisonne-t-il au niveau INDICES D'IMAGE [GORDI-85] [LUX-85b], comme la plupart des systèmes actuels.

" Quelles suites est-il envisageable de donner à vos travaux? "

Notre volonté est de poursuivre les travaux dans la même direction, en suivant la méthodologie actuelle des études en Vision par Ordinateur : 1/ choisir plusieurs méthodes, où pour chacune une source particulière d'information est isolée et étudiée à fond, 2/ coordonner les méthodes étudiées suivant l'idée que seule la conjonction de plusieurs de ces méthodes permet d'inférer une description complète de la scène, et de vérifier en permanence la cohérence des connaissances inférées. C'est dans cette optique que nous comptons proposer dans un prochain rapport les grandes lignes du système de Vision par Ordinateur SATURNE (abréviation pour Système d'Analyse Tridimensionnel Utilisant une Représentation par Niveaux Explicites) [DEMAZ-87]. Incluant les différents niveaux de représentation défendus dans cette thèse, cette proposition constitue le prolongement immédiat de nos idées actuelles par rapport aux problèmes du contexte, de la reconnaissance et de la désignation. L'expérience que nous avons acquise par le travail rapporté ici et les nouvelles expérimentations que nous pensons mener devraient permettre de mieux formaliser tant les différents niveaux de représentation que les procédures d'enrichissement, de contrôle et d'inférence. Mais le dessein de la proposition est aussi de constituer une base pour des études ultérieures destinées à approcher les niveaux

de représentation OBJET et SCENE pour mieux les étudier.

" Les cinq niveaux et leurs procédures associées mises à part, quelles semblent être les particularités de la proposition SATURNE? "

Nous formulons ici des objectifs, illustrés par la figure SATU. Nous sommes d'ailleurs conscients et quelque peu impressionnés par l'ampleur de la tâche.

La première des particularités de cette proposition de système en gestation sera la possibilité, pour une même structure, d'accueillir les quatre différents types de méthodes en Vision par Ordinateur, types que nous avons explicités au cours de la partie A.

Au cours des expérimentations que nous avons présentées, nous avons vu comment exploiter certaines propriétés analytiques, géométriques et physiques du monde réel. Le système devra en sus intégrer des connaissances d'ordre fonctionnel, temporel, relationnel, et linguistique (au sens des langages de communication avec d'autres êtres intelligents).

La possibilité de communication entre êtres intelligents aux différents niveaux de représentation (en particulier dans une optique de coopération bras-oeil) sera une autre caractéristique du système [TROCC-86].

" Comment situer SATURNE par rapport aux différentes tentatives de construction de systèmes de compréhension de scène qui ont été développés dans le passé et/ou qui sont en cours de développement? "

La comparaison par rapport aux premiers systèmes avancés peut difficilement être effectuée, l'approche du problème de Vision par Ordinateur n'étant pas, à l'époque, celle qui caractérise la plupart des systèmes en cours de développement (étude poussée de plusieurs caractéristiques puis coopération entre les méthodes développées). Dans notre réponse à cette question, la comparaison ne pourra donc être qu'implicite. L'analogie avec les systèmes en cours de développement est sans doute un peu plus aisée. Nous séparerons donc notre réponse en deux parties. Dans tous les cas, le nombre des systèmes évoqués est bien loin de celui de la liste exhaustive dont ils font partie.

PAR RAPPORT AUX PREMIERS SYSTEMES AVANCES La plupart des premiers systèmes de Vision (presque exhaustivement décrits dans [LUX-83b]) ont été réalisés dans un but précis, mais aucun n'a cherché à être interprété comme modèle de la vision. Cet état de choses est tout particulièrement perceptible dans les tous premiers systèmes de reconnaissance. Il utilisent un nombre très réduit de niveaux de représentation, dont l'existence est jugée indispensable [BINFO-82]. Par exemple, de nombreux systèmes construisent une structure de l'image digitalisée en termes de régions et/ou de lignes qui sont directement interprétées par mise en correspondance avec des modèles d'objets. La présence d'autres niveaux dans ces systèmes, ne correspond qu'au moyen imaginé pour résoudre le problème de la synthèse de sources d'information multiples.

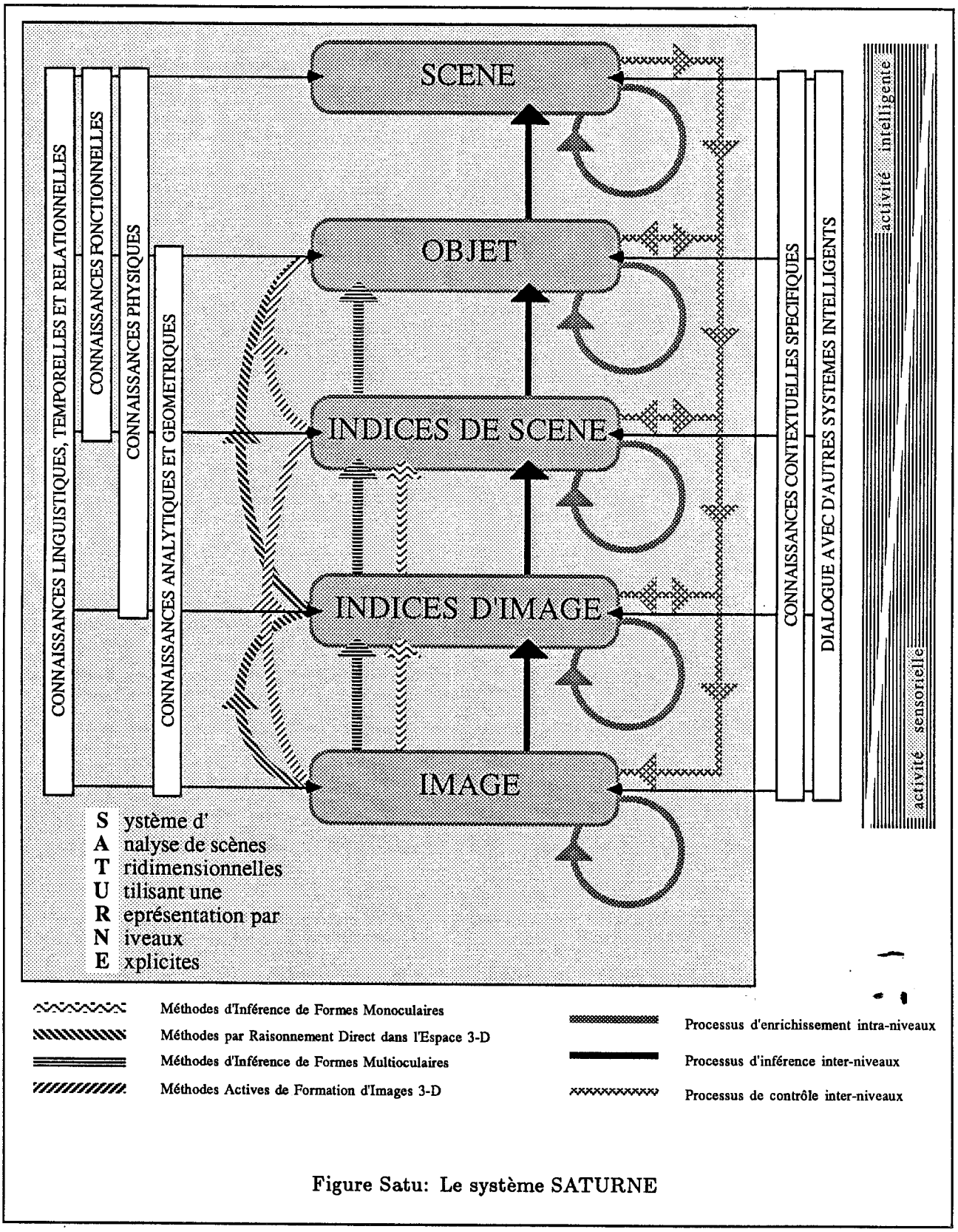


Figure Satu: Le système SATURNE

C'est typiquement le cas des travaux effectués par Hanson et Riseman qui sont l'illustration même de la volonté à maîtriser plusieurs sources d'informations différentes [HANSO-78]. En effet, le système VISIONS tient compte de la potentielle multiplicité des sources d'information (informations de contraste, de couleur, de distance) et les niveaux présents reflètent essentiellement le besoin de posséder des points de contrôle où plusieurs sources d'information interviennent simultanément, mais la raison d'être de ces niveaux ne se traduit bien souvent que par le besoin de développer de nouvelles entités géométriques nécessaires aux inférences ultérieures.

Le système très spécifique de Ballard, Brown et Feldmann est entièrement dirigé par le but, et si les niveaux y sont implicitement présents, l'interprétation se fait plutôt au coup par coup, un niveau étant créé dynamiquement dès que le modèle choisi fait apparaître ce dernier [BALLA-78]. Les niveaux considérés sont donc ceux du modèle par rapport auquel s'effectue la reconnaissance.

Le système développé par Barrow et Tenenbaum ([BARRO-75] [BARRO-78]), un peu moins ambitieux que les précédents, s'intéresse essentiellement aux indices qui peuvent être extraits directement de l'image, indépendamment de tout contexte (images intrinsèques). Sa force réside dans le fait que ces indices sont obtenus en totale indépendance vis-à-vis de la base de modèles des objets et du contexte que le système connaît.

PAR RAPPORT AUX SYSTEMES EN COURS DE DEVELOPPEMENT La plus grande ambition des systèmes en cours de développement est liée à plusieurs phénomènes dont en particulier l'introduction de la notion de cylindres généralisés par [BROOK-81], mais aussi par les résultats obtenus par l'équipe de Marr [ULLMA-80] [MARR-82].

Les niveaux présents dans le système de Marr reflètent une étude minutieuse de l'image, uniquement basée sur les propriétés physiques du monde réel. L'implémentation réalisée n'est que partielle : les parties développées concernent surtout le "croquis élémentaire" [MARR-80], (assimilable à notre niveau INDICES D'IMAGE) la stéréoscopie [MARR-79] [GRIMS-81], le mouvement [ULLMA-79], et quelques uns des autres modules participant à la construction du croquis 2,5-D (assimilable à notre niveau INDICES DE SCENE).

Une hiérarchie de niveaux figure aussi dans les travaux de Crowley [CROWL-84] [CROWL-86]. Ils se concentrent sur un niveau de représentation, le Modèle Composite de Surfaces, qui est à positionner entre notre niveau INDICES DE SCENE et notre niveau OBJET. Ce niveau est actuellement obtenu par un enrichissement incrémental, sur la base d'études menées en stéréovision généralisée et sur l'expérimentation de méthodes actives de formation d'images 3-D.

Une idée similaire guide les travaux de Herman [HERMA-82] [HERMA-84] et de Kanade [KANAD-83]. Ils s'attachent eux aussi à la construction d'un niveau similaire à celui de Crowley, mais sur la base d'une étude en parallèle de l'inférence de formes à partir des contours et de l'inférence de forme à partir de la stéréovision simple, méthodes dont les résultats alimentant séparément un modèle incrémental du niveau recherché, très proche de notre niveau OBJET.

Des points de vue de la représentation de l'information visuelle à différents niveaux de représentation et de la conjonction de plusieurs méthodes [GLICK-83], nos travaux sont donc à rapprocher de ces trois systèmes en cours de développement, bien que nous pensions qu'une synthèse de résultats obtenus par deux méthodes différentes puisse être intégrés à n'importe lequel des cinq niveaux que nous défendons, et non à un seul. Des points de vue de la spécification et de la formalisation des niveaux et des procédures inter- et intra-niveaux, SATURNE est plus à rapprocher des travaux de Marr et Poggio, mais aussi de la réflexion menée par Havens et Mackworth [HAVEN-83].

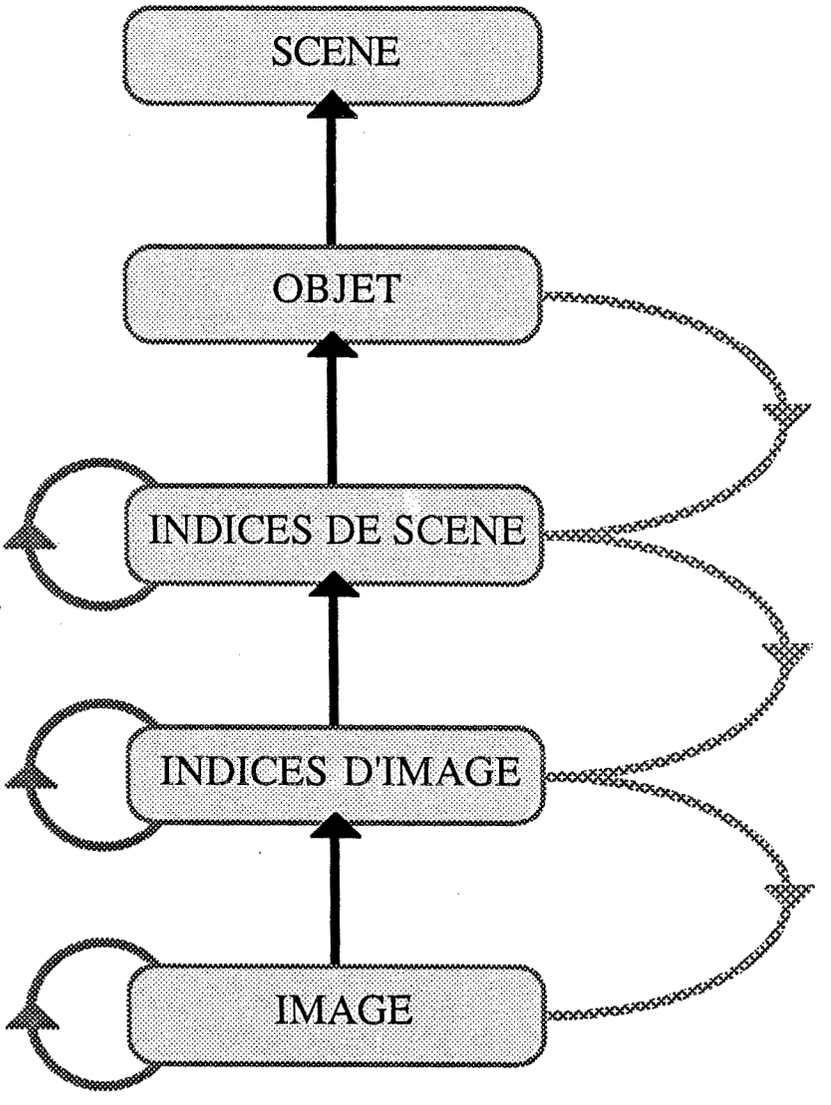
Notons finalement qu'il existe une différence essentielle entre SATURNE et tous les systèmes qui ont été évoqués dans cette réponse à la question posée : SATURNE n'est qu'une proposition qui n'a pas encore fait ses preuves et n'a fourni de résultats que par le truchement des travaux ici rapportés, alors que tous ces systèmes ont déjà donné lieu à un travail technique considérable!

" Justement, là est le problème. Quels axes de recherche sous-entend la réalisation de ces futurs développement et comment comptez-vous les aborder? "

A priori, les objectifs majeurs de la recherche poursuivie seront donc aussi ceux qui ont constitué nos justifications à l'existence des niveaux : abstraction et décentration, multiplicité des sources d'information, communication et ouverture à un dialogue avec d'autres agents intelligents. Vis-à-vis du problème de la synthèse d'informations multiples, une toute première expérimentation pourrait être par exemple l'analyse de scènes composées d'objets de révolution qui peuvent être translucides, au moyen des trois méthodes d'inférence respectivement fondées sur les contours, la stéréovision simple et la réflectance. Nous espérons pouvoir montrer deux choses : 1/ alors que chacune de ces méthodes ne peut assurer à elle seule une compréhension de telles scènes, la conjonction de ces trois méthodes, elle, y parvient (cf. figure EXPE), et 2/ que la formalisation qui sera faite pour SATURNE permet d'adapter directement les cinq niveaux au traitement d'un nouveau problème.

Pour progresser vers ces objectifs, il est donc nécessaire que les axes déjà abordés aujourd'hui soient poursuivis et accentués. Plus précisément, deux sujets de recherche ponctuels devront être :

- la découverte de nouveaux types indices d'image, de nouveaux indices de scène, avant même d'envisager de formaliser la notion d'objet. De ce point de vue, la vision par ordinateur est un domaine où il n'est pas simple de trouver un expert qui puisse expliquer non seulement ce qu'il voit, mais aussi pourquoi il le voit et comment. Un "système expert de vision" est difficilement imaginable, et il est tout aussi difficile de découvrir de nouveaux indices encore inconnus! Plus que jamais, nous pensons qu'il est possible de découvrir de nouveaux indices visuels par le biais de la psychologie et celui de la neurophysiologie, mais aussi en s'inspirant de domaines informatiques comme par exemple celui de la programmation des jeux : quels sont les indices qu'utilise un joueur de go pour représenter et raisonner sur une situation



Procédures intra-niveaux destinées à enrichir la structure brute d'un niveau en adéquation avec son exploitation probable

Algorithmes d'interprétation inter-niveaux permettant, à partir d'informations contenues à un niveau donné, d'inférer des informations d'un niveau supérieur

Processus de contrôle inter-niveaux susceptibles d'être utilisés pour diriger les interprétations locales des indices de niveau inférieur

Figure Expe: Utilisation des Niveaux pour une troisième expérimentation

intermédiaire particulière du go-ban, qui n'est ni plus ni moins qu'une image à trois niveaux de gris?

- la maîtrise de la combinatoire de l'interprétation des indices d'image en indices de scène, car les méthodes développées restent encore fortement combinatoires. A ce point, nous pensons par exemple pouvoir améliorer et diminuer la combinatoire de l'interprétation des indices d'image en indices de scènes par l'utilisation de règles d'inférence laissant apparaître des symboles de variable, et en se rapprochant donc d'un modèle de programmation logique. Par ailleurs, nous pensons aussi que dans le cas de la stéréoscopie, une meilleure exploitation des contraintes géométriques entre indices d'image et aussi entre indices de scènes devrait permettre une réduction notoire de la combinatoire de ces interprétations : et si autant ce dernier objectif est précis, autant le travail qu'il sous-entend est important.

Mais il est tout aussi clair que, très rapidement, au moins deux autres axes de recherche seront inévitablement abordés.

La gestion temporelle des modèles de représentation des connaissances Il ne s'agit pas ici d'étudier le temps lorsqu'il est considéré comme un facteur de contrainte (par rapport à un problème de planification en temps réel par exemple), mais comme un facteur inévitable durant l'écoulement duquel de nouveaux phénomènes se produisent et donc au cours duquel de nouveaux indices apparaissent. La question est de savoir comment ces nouveaux indices s'intègrent aux - mais aussi éventuellement modifient l'accessibilité des - modèles de représentation des connaissances en cours de construction, à quelque niveau d'abstraction et de décentration que ce soit. A ce propos, nous pensons que l'on retrouve à tous les niveaux de représentation ce même processus fondamental suivant lequel l'acquisition d'une connaissance, et surtout d'un ensemble de connaissances, ne s'effectue pas par voie exclusivement additive, mais comporte de continuelles réorganisations à partir d'éléments ou de relations initialement privilégiées. Nos réponses à ce problème sont selon nous à rapprocher du caractère incrémental de la mise à jour du "modèle composite" développé par Crowley [CROWL-86], mais sont aussi connexes au caractère anticipatif du modèle de Garvey [GARVE-84], même si le système de ce dernier présuppose une connaissance a priori d'un modèle bien défini du monde réel. Là encore, nous pensons que des travaux a priori déconnectés de la Vision par Ordinateur peuvent alimenter une source d'inspiration importante : une telle ouverture existe par exemple du coté de la dynamique de modélisation des connaissances avalancheux (et météorologiques de façon générale), technique utilisée pour être capable en permanence de répondre à une quelconque demande de prévision de risques d'avalanches [GRANI-84].

L'apprentissage au sein d'un système de vision Dans l'instant présent, nous pensons qu'il peut exister en vision trois types d'apprentissage, que nous définissons par rapport aux procédures d'inférence.

L'"apprentissage d'INDICES" (ou "apprentissage d'objets") consiste, pour des procédures d'inférence données, à effectuer des optimisations tant quantitatives

(dans un souci de réduction de la combinatoire) que qualitatives (pour une meilleure réalité de l'inférence considérée). Cela peut se traduire, dans le cas de l'inférence des indices de scène telle que nous l'avons abordée, par la pondération (qui peut être initialement empirique) des opérateurs d'interprétation d'un noeud-image de type donné règles avec des coefficients qu'il s'agit ensuite de "mettre au point" quitte à ce qu'il provoque une disparition totale de l'opérateur d'interprétation. Pour l'extraction des indices d'image, cela correspond par exemple à la mise au point d'un algorithme d'extraction de droite-image par exemple.

L'"apprentissage de METHODES" (ou "apprentissage de concepts".) consiste à découvrir et à créer de nouvelles procédures d'inférence, dans une optique à rapprocher de celle de [SAINT-86]. Là encore, pour l'interprétation en indices de scène, cela peut correspondre à découvrir, pour un nouveau type de noeud-image considéré, quels sont les opérateurs d'interprétation qui lui sont applicables, sur la base des autres inférences connues par le système et tout en vérifiant la cohérence de l'ensemble des interprétations possibles. Pour l'extraction des indices d'image, cela se traduit par exemple par la détection de nouveaux indices d'image, problème que nous avons déjà évoqué précédemment.

L'"apprentissage de COMPORTEMENT" (ou "apprentissage de CONTROLE ") consiste, pour un ensemble de procédures d'inférence donné, et relativement à un but fixé (de compréhension d'une classe d'image ou de reconnaissance d'une classe d'objet), à acquérir une méthodologie d'étude du but fixé, et donc de déterminer un ensemble de procédures de contrôle pour organiser l'activation des procédures de manière à atteindre ce but.

Les deux problèmes que nous venons d'évoquer traduisent le fait que le but du système de Vision par Ordinateur n'est pas d'aboutir à une construction définitive dans laquelle les processus les plus "intelligents" n'auraient qu'à puiser des connaissances pour les généraliser ou même éventuellement pour les prolonger, mais que la connaissance se construit ou se reconstruit à chaque palier de développement, à chaque observation d'un nouvel indice, à quelque niveau que ce soit. Cette reconstruction (ou "réorganisation"), qui doit parfois s'accompagner de modifications structurelles, peut bien sûr s'appuyer sur les constructions antérieures mais pour les transporter puis pour les dépasser.

" Ces deux derniers thèmes, ainsi que d'autres évoqués précédemment, ne relèvent pas seulement de la Vision par Ordinateur, mais pour beaucoup du domaine de l'Intelligence Artificielle. Comment situer votre proposition par rapport aux systèmes d'Intelligence Artificielle? "

La proposition que nous émettons est destinée à servir de base pour construire un système de compréhension de scène. Si les références auxquelles il faut le comparer sont le système expert de reconnaissance de galaxies développé par [THONN-85] ou le modèle du K + I + C ("Knowledge - Inference - Control") proposé par [LUX-85a], fortement empreints de techniques et de représentations typiques du domaine de l'Intelligence Artificielle, alors SATURNE est peut-être un système intelligent, à

condition d'accepter la distinction fortement marquée entre plusieurs niveaux de représentation. Par rapport à notre interprétation du K + I + C, nous pouvons d'ailleurs dire que les travaux que nous avons effectués se rapportent surtout au K d'un point de vue de vue théorique, au K et au I pour les expérimentations dans le domaine de la couleur et de l'inférence de formes à partir de contours et à partir de la stéréovision simple, et que le C n'a presque pas été étudié.

Si le schéma du "système intelligent" est maintenant quelque chose de la forme "perception - raisonnement - action", le système SATURNE peut ne pas en être un, puisqu'il se limite à une perception visuelle pour un raisonnement "gratuit", et qu'on peut donc le considérer comme restreint à son seul module de perception.

Dans tous les cas, nous considérons l'intelligence comme un système d'ensemble de fonctions cognitives. Le problème de savoir si un système de vision par ordinateur est un système d'intelligence artificielle est un faux problème, tant les domaines sont mal définis chacun séparément, et tant la frontière entre les deux domaines est mince. Par contre, il est une remarque importante, par rapport à la structure même du système que nous proposons ou de tout autre système reconnaissant l'existence de différents niveaux de représentation, et ceci sans préjuger de la nature même de ces niveaux : admettre, comme nous l'avons fait, que les points de dialogue avec les autres systèmes intelligents *sont* les niveaux de représentation qui sont explicitement présents au sein du système de Vision par Ordinateur, conduit à supposer que les autres systèmes comprennent les communications qui leur sont adressées et sont eux-mêmes capables de formuler leurs demandes à ces niveaux de représentation : nous sommes donc parfois tentés de penser que les systèmes intelligents possèdent une structure unique! La structure même du système SATURNE permet effectivement la communication entre plusieurs systèmes de vision par ordinateur, puisqu'elle souhaite intégrer *tous* les types de méthodes utilisées en Vision par Ordinateur, mais qu'en est-il de la communication avec les systèmes intelligents d'un autre domaine? Nos connaissances ne nous permettent pas de répondre. Pourtant, toujours dans l'hypothèse où les niveaux de communication *seraient* les niveaux de représentation, peut-être faudrait-il envisager l'étude d'une modélisation commune adaptée aux différents domaines où existe une collaboration entre systèmes intelligents? C'est tout au moins ce que nous croyons nécessaire dans un environnement comprenant des systèmes d'Intelligence Artificielle, de Robotique, et de Vision par Ordinateur, et c'est un des besoins inévitable de la coopération bras-oeil.

" Pour en revenir une dernière fois au système de Vision par Ordinateur SATURNE, une telle synthèse est-elle d'ores et déjà nécessaire? "

Une telle démarche se justifie d'un premier point de vue par le simple fait que les travaux que nous avons effectués à ce jour concernent essentiellement les niveaux IMAGE, INDICES D'IMAGE et INDICES DE SCENE. Dans l'état actuel, les niveaux OBJET et SCENE n'ont été que partiellement étudiés, alors qu'ils semblent au moins tout aussi importants sinon plus que les autres : étant les plus décentrés et les plus abstraits, ils contiennent en effet de nombreux indices et notions qui peuvent facilement être partagées par les trois domaines de la Vision par Ordinateur,

de la Robotique et de l'Intelligence Artificielle. Cette absence est en particulier due à ce que, ni la dimension temporelle, ni l'apport de connaissances non visuelles (géométriques, fonctionnelles et/ou physiques) n'ont été jusqu'alors ouvertement considérées. Et c'est pourquoi, autant il nous semble acquis que les deux niveaux OBJET et SCENE doivent figurer dans tout système de compréhension de scènes, autant il nous paraît indispensable de construire un système comme SATURNE (même si les nombreux objectifs visés ne doivent être que partiellement atteints) pour réellement permettre de mieux appréhender chacun et l'ensemble des cinq niveaux de représentation IMAGE, INDICES D'IMAGE, INDICES DE SCENE, OBJET et SCENE.

Par ailleurs, et comme nous l'avons dit, il est clair que les méthodes et les résultats issus de la seule étude de deux caractéristiques (contours et stéréovision simple) ne peuvent constituer une démonstration indiscutable de la proposition SATURNE, et ne font qu'appuyer la thèse défendue sans la vérifier irréfutablement. Seules les études de nouvelles caractéristiques, si elles parviennent à s'intégrer directement au modèle SATURNE, permettront de valider la proposition et d'en corriger les erreurs, tout comme les expériences en physique permettent de valider des modèles de phénomènes qui sont tout aussi réels.

Bibliographie

[AGIN-76] Gerald J. AGIN and Thomas O. BINFORD. Computer description of curved objects. *I.E.E.E. Transactions on Computer*, C-25(4):439–449, April 1976.

[AGIN-81] Gerald J. AGIN. *Fitting Ellipses and General Second-Order Curves.* Technical Report CMU RI-TR-81-5, Carnegie-Mellon University - The Robotics Institute, July 1981.

[AKITA-82] Koichiro AKITA and Hideki KUGA. A computer method of understanding ocular fundus images. *Pattern Recognition*, 15(6):431–443, November 1982.

[AYACH-85] Nicolas AYACHE and Bernard FAVERJON. Fast stereo matching of edge segments using prediction and verification of hypotheses. In *Conference on Computer Vision and Pattern Recognition*, pages 662–664, I.E.E.E., San Francisco, June 1985.

[BALLA-78] D. H. BALLARD C. M. BROWN and J. A. FELDMANN. An approach to knowledge-directed image analysis. In Allen R. Hanson and Edward M. Riseman, editors, *Computer Vision Systems*, Academic Press, New York, 1978.

[BALLA-82] Dana H. BALLARD and Christopher M. BROWN. *Computer Vision.* Prentice-Hall, Englewood Cliffs, 1982.

[BALLA-84] Dana H. BALLARD. Parameter nets. *Artificial Intelligence*, 22:235–267, April 1984.

[BARLO-85] H. B. BARLOW. Perception: what quantitative laws govern the acquisition of knowledge from the senses. In Clive Warwick Coen,

editor, *Functions of the Brain*, Clarendon Press, Oxford, 1985.

[BARNA-83] Stephen T. BARNARD. Interpreting perspective images. *Artificial Intelligence*, 21:435–462, November 1983.

[BARNA-84] Stephen T. BARNARD. *Choosing a basis for perceptual space.* Technical Note 315, SRI International - Artificial Intelligence Center, January 1984.

[BARRO-75] Harry G. BARROW and Jay M. TENENBAUM. *Representation and use of knowledge in vision.* Technical Note 108, SRI International - Artificial Intelligence Center, 1975.

[BARRO-78] Harry G. BARROW and Jay M. TENENBAUM. Recovering intrinsic characteristics from images. In Allen R. Hanson and Edward M. Riseman, editors, *Computer Vision Systems*, Academic Press, New York, 1978.

[BARRO-81] Harry G. BARROW and Jay M. TENENBAUM. Interpreting line drawings as three-dimensional surfaces. *Artificial Intelligence*, 17:75–116, August 1981.

[BESSI-83] Pierre BESSIERE. Etude de la génération de plans en univers multi-agents. Thèse de Docteur-Ingénieur, Laboratoire d'Informatique Fondamentale et d'Intelligence Artificielle, Institut National Polytechnique de Grenoble, Décembre 1983.

[BIEDE-81] Irving BIEDERMAN. On the semantics of a glance at a scene. In M. Kubovy and J. R. Pomerantz, editors, *Perceptual Organization*, Erlbaum, Hillsdale, 1981.

[BINFO-81] Thomas O. BINFORD. Inferring surfaces from images. *Artificial Intelligence*, 17:205–244, August 1981.

[BINFO-82] Thomas O. BINFORD. Survey of model-based image analysis systems. *The International Journal of Robotics Research*, 1(1):18–64, Spring 1982.

[BLAKE-85] Andrew BLAKE. Specular stereo. In *9th International Joint Conference on Artificial Intelligence*, pages 973–976, I.J.C.A.I., Los Angeles, August 1985.

[BOLLE-83] R. C. BOLLES P. HORAUD and M. J. HANNAH. 3DPO: a three-dimensional part orientation system. In *8th International Joint Conference on Artificial Intelligence*, pages 1116–1120, I.J.C.A.I., Karlsruhe, August 1983.

[BOLLE-85] Robert C. BOLLES and H. Harlyn BAKER. Epipolar-plane image analysis: a technique for analyzing motion sequences. In *3rd Workshop on Computer Vision: Representation and Control*, pages 168–178, I.E.E.E., Bellaire, October 1985.

[BORIA-84] Paul-Louis BORIANNE. Contributions à la vision par ordinateur tridimensionnelle. Thèse 3ème Cycle, Laboratoire d'Informatique Fondamentale et d'Intelligence Artificielle, Institut National Polytechnique de Grenoble, Avril 1984.

[BRADY-81] Michael BRADY. The changing shape of computer vision. *Artificial Intelligence*, 17:1–15, August 1981.

[BRADY-82] Michael BRADY. Computer vision (correspondent's report). *Artificial Intelligence*, 19:7–16, 1982.

[BRADY-83] Michael BRADY and Alan L. YUILLE. An extremum principle for shape from contour. In *8th International Joint Conference on Artificial Intelligence*, pages 969–972, I.J.C.A.I., Karlsruhe, August 1983.

[BROOK-81] Rodney A. BROOKS. Symbolic reasoning among 3D models and 2D images. *Artificial Intelligence*, 17:285–348, August 1981.

[BROWN-82] Christopher M. BROWN. *Color Vision and Computer Vision.* TR-108, University of Rochester - Computer Science Department, June 1982.

[CASTA-84] S. CASTAN and J. SHEN. Une méthode photométrique pour déterminer l'orientation des surfaces possédant certaines propriétés de réflectance. In *4ème Congrès Reconnaissance des Formes et Intelligence Artificielle*, pages 81–92, R.F.I.A., Paris, Janvier 1984.

[CHAKR-82] Indranil CHAKRAVARTY. *The Use of Characteristic Views as a Basis for Recognition of Three-Dimensional Objects.* PhD thesis, Rensselaer Polytechnic Institute - Image Processing Laboratory, October 1982.

[CLOWE-71] M. B. CLOWES. On seing things. *Artificial Intelligence*, 2:79–116, 1971.

[COHEN-82] Paul R. COHEN and Edward A. FEIGENBAUM. *The Handbook of Artificial Intelligence.* Volume 3, William Kaufmann Inc., Los Altos, 1982.

[CROWL-84] James L. CROWLEY. *A Computational Paradigm for Three Dimensional Scene Analysis.* Technical Report CMU RI-TR-84-11, Carnegie-Mellon University - The Robotics Institute, April 1984.

[CROWL-85] James L. CROWLEY. Navigation for an intelligent mobile robot. *I.E.E.E. Journal of Robotics and Automation*, 1(1):31–41, March 1985.

[CROWL-86] James L. CROWLEY. Representation and maintenance of a composite surface model. In *3rd I.E.E.E. Conference on Robotics and Automation*, pages 1455–1462, I.E.E.E., San Francisco, April 1986.

[DEMAZ-82] Yves DEMAZEAU. Analyse de scènes visuelles tridimensionnelles. D.E.A. Informatique, Laboratoire d'Informatique Fondamentale et d'Intelligence Artificielle, Octobre 1982.

[DEMAZ-84a] Yves DEMAZEAU. Niveaux de représentation pour la compréhension d'images. In *4ème Congrès Reconnaissance des Formes et Intelligence Artificielle*, pages 199–209, R.F.I.A., Paris, Janvier 1984.

[DEMAZ-84b] Yves DEMAZEAU and Pascal DI GIACOMO. Etude de la réalisation d'un prototype d'une cellule robotisée pour la manipulation et le traitement des câbles. Rapport de contrat entre le Laboratoire d'Informatique Fondamentale et d'Intelligence Artificielle et la Compagnie des Machines Bull de Belfort, Octobre 1984.

[DEMAZ-85a] Yves DEMAZEAU and Jose Luis GORDILLO. A color stereo vision system applied to wire identification and localization. In *4th Canadian CAD/CAM and Robotics Conference*, pages 12.1–12.8, C.C.R.C., Toronto, June 1985.

[DEMAZ-85b] Yves DEMAZEAU. A stereoscopic vision sensor for robotics: use, design and calibration. In *15th International Symposium on Industrial Robots*, pages 357–365, I.S.I.R., Tokyo, September 1985.

[DEMAZ-87] Yves DEMAZEAU. *The SATURNE project: a multi-level representation for the cooperative inference of scenes from images - Work in progress.* Rapport de Recherche, Laboratoire d'Informatique Fondamentale et d'Intelligence Artificielle, 1987.

[DRAPE-81] Stephen W. DRAPER. The use of gradient and dual space in line-drawing interpretation. *Artificial Intelligence*, 17:461–508, August 1981.

[DUDA-72] Richard O. DUDA and Peter E. HART. *Use of the Hough transformation to detect lines and curves in pictures.* Technical Note 36, SRI International - Artificial Intelligence Center, 1972.

[DUDA-73] Richard O. DUDA and Peter E. HART. *Pattern Classification and Scene Analysis.* John Wiley and Sons, New York, 1973.

[FALK-72] Gilbert FALK. Interpretation of imperfect line data as a three-dimensional scene. *Artificial Intelligence*, 3:101–144, Summer 1972.

[FAUGE-84] Olivier FAUGERAS and al. Object representation, identification, and positioning from range data. In Michael Brady and Richard Paul, editors, *Robotics Research: The 1st International Symposium*, pages 425–446, M.I.T. Press, Cambridge, 1984.

[FRISB-79] John P. FRISBY. *Seeing.* Oxford University Press, Oxford, 1979.

[GARVE-76] Thomas D. GARVEY. *Perceptual strategies for purposive vision.* Technical Note 117, SRI International - Artificial Intelligence Center, September 1976.

[GARVE-84] Thomas D. GARVEY. *An AI approach to the integration of information.* Technical Note 318, SRI International - Artificial Intelligence Center, April 1984.

[GENNE-79] Donald B. GENNERY. Stereo-camera calibration. In *'79 Image Understanding Workshop*, pages 101–107, D.A.R.P.A., Los Angeles, November 1979.

[GESCH-79] Clifford GESCHKE. *A robot task using visual tracking.* Master's thesis, University of Illinois, 1979.

[GIRAL-84] G. GIRALT R. CHATILA and M. VAISSET. An integrated navigation and motion control system for autonomous multisensory mobile robots. In Michael Brady and Richard Paul, editors, *Robotics Research: The 1st International Symposium*, pages 192–214, M.I.T. Press, Cambridge, 1984.

[GLICK-83] J. GLICKSMAN. Using multiple information sources in a computational vision system. In *8th International Joint Conference on Artificial Intelligence*, pages 1078–1080, I.J.C.A.I., Karlsruhe, August 1983.

[GORDI-83] Jose Luis GORDILLO. Analyse d'images couleur. D.E.A. Informatique, Laboratoire d'Informatique Fondamentale et d'Intelligence Artificielle, Septembre 1983.

[GORDI-85] Jose Luis GORDILLO. Colour representation for vision machine. In *2nd International Conference on Machine Intelligence*, M.I., Londres, November 1985.

[GORDI-86] Jose Luis GORDILLO. *CAICOU: un Système de Développement Interactif pour l'Analyse d'Images en Couleur.* Rapport de Recherche LIFIA 39, Laboratoire d'Informatique Fondamentale et d'Intelligence Artificielle, Janvier 1986.

[GRANG-85] Catherine GRANGER. *Reconnaissance d'objets par mise en correspondance en vision par ordinateur.* PhD thesis, Institut National de Recherche en Informatique et en Automatique - Centre de Sophia-Antipolis - Université de Nice, Juin 1985.

[GRANI-84] Thierry GRANIER. Etude de l'aspect évolutif dans les informations et de son utilisation dans le raisonnement. application à la prévison des risques d'avalanches en montagne. D.E.A. Informatique, Laboratoire d'Informatique Fondamentale et d'Intelligence Artificielle, Septembre 1984.

[GRIMS-81] William Eric L. GRIMSON. *From Image to Surfaces: a Computational Study of the Human Early Visual System.* M.I.T. Press, Cambridge, 1981.

[GUZMA-68] Adolfo GUZMAN-ARENAS. Decomposition of a visual scene into three-dimensional bodies. In *68' A.F.I.P.S. Fall Joint Conference*, pages 291–304, A.F.I.P.S., San Francisco, December 1968.

[HANSO-78] Allen R. HANSON and Edward M. RISEMAN. Visions: a computer vision system for analysing scenes. In Allen R. Hanson and Edward M. Riseman, editors, *Computer Vision Systems*, Academic Press, New York, 1978.

[HAVEN-83] William S. HAVENS and Alan K. MACKWORTH. Representing knowledge of the visual world. *I.E.E.E. Computer*, 16(10):90–96, October 1983.

[HEALE-85] Glenn HEALEY and Thomas O. BINFORD. Predicting specular features. In *'85 Image Understanding Workshop*, pages 479–488, D.A.R.P.A., December 1985.

[HERMA-82] M. HERMAN T. KANADE and S. KUROE. *Incremental Acquisition of a Three-dimensional Scene Model from Images.* Research Report CMU CS-82-139, Carnegie-Mellon University - Department of Computer Science, October 1982.

[HERMA-84] Martin HERMAN and Takeo KANADE. *The 3D MOSAIC Scene Understanding System: Incremental Reconstruction of 3D Scenes from Complex Images.* Research Report CMU CS-84-102, Carnegie-Mellon University - Department of Computer Science, February 1984.

[HILDR-82] Ellen C. HILDRETH and Shimon ULLMAN. *The measurement of visual motion.* Research Report A.I. MEMO 699, Massachusetts Institute of Technology - Artificial Intelligence Laboratory, December 1982.

[HORAU-85] Radu HORAUD. Spatial object perception from an image. In *9th International Joint Conference on Artificial Intelligence*, pages 1116–1119, I.J.C.A.I., Los Angeles, August 1985.

[HORN-77] Berthold K. P. HORN. Understanding image intensities. *Artificial Intelligence*, 8:201–231, April 1977.

[HUBEL-77] David H. HUBEL and Torsten N. WIESEL. Ferrier lecture: functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London*, B-198:1–59, July 1977.

[HUFFM-71] D. A. HUFFMAN. Impossible objects as nonsense sentences. In R. Meltzer et D. Michie, editor, *Machine Intelligence 6*, Edinburgh University Press, Edinburgh, 1971.

[IKEUC-80] Katsuki IKEUCHI. *Shape from Regular Patterns (an example of constraint propagation in vision).* Research Report A.I. MEMO 567, Massachusetts Institute of Technology - Artificial Intelligence Laboratory, March 1980.

[IMBER-83] M. IMBERT. La neurobiologie de l'image. *La Recherche*, 14(144):600–613, Mai 1983.

[INOUE-84] Hirochika INOUE and Masayuki INABA. Hand-eye coordination in rope handling. In Michael Brady and Richard Paul, editors, *Robotics Research: The 1st International Symposium*, pages 165–174, M:I.T. Press, Cambridge, 1984.

[JUDD-75] Deane B. JUDD and Gunter WYSZECKI. *Color in Business, Science and Industry.* John Wiley and Sons, New York, 1975.

[KANAD-77] Takeo KANADE. Model representations and control structures in image understanding. In *5th International Joint Conference on Artificial Intelligence*, pages 1074–1082, I.J.C.A.I., Cambridge, August 1977.

[KANAD-80a] Takeo KANADE. A theory of Origami world. *Artificial Intelligence*, 13:279–311, May 1980.

[KANAD-80b] Takeo KANADE and John R. KENDER. *Mapping Image Properties into Shape Constraints: Skewed Symetry, Affine-Transformable Patterns, and the Shape from Texture Paradigm.* Research Report CMU CS-80-133, Carnegie-Mellon University - Department of Computer Scienc, July 1980.

[KANAD-81] Takeo KANADE. Recovery of the three-dimensional shape of an object from a single view. *Artificial Intelligence*, 17:409–460, August 1981.

[KANAD-83] Takeo KANADE. Geometrical aspects of interpreting images as a three-dimensional scene. In *I.E.E.E. Conference*, pages 789–802, I.E.E.E., July 1983.

[KENDE-76] John R. KENDER. *Saturation, Hue, and Normalized Color: Calculation, Digitization Effects, and Use.* Research Report CMU CS (INV. 02826), Carnegie-Mellon University - Department of Computer Science, November 1976.

[KENDE-80] John R. KENDER. *Shape from Texture.* PhD thesis, Carnegie-Mellon University - Department of Computer Science, November 1980.

[KENDE-86] John R. KENDER and Earl M. SMITH. Shape from darkness: deriving surface information from dynamic shadows. In *'86 AAAI Conference*, pages 664–669, A.A.A.I., Philadelphie, August 1986.

[KONIS-84] T. KONISHI M. TAKAGI and J. KITSUKI. Shape reconstruction of wires using colour images for automatic soldering. In Michael Brady and Richard Paul, editors, *Robotics Research: The 1st International Symposium*, pages 389–399, M.I.T. Press, Cambridge, 1984.

[LELAN-84] S. LELANDAIS. Numérisation de formes tridimensionnelles: acquisition, traitement. In *4ème Congrès Reconnaissance des Formes et Intelligence Artificielle*, pages 339–356, R.F.I.A., Paris, Janvier 1984.

[LONGU-80] Christopher LONGUET-HIGGINS and K. PRAZDNY. The interpretation of moving retinal images. *Proceedings of the Royal Society of London*, B-208:385–387, 1980.

[LONGU-82] Christopher LONGUET-HIGGINS. The role of the vertical dimension in stereoscopic vision. *Perception*, 11:377–386, November 1982.

[LOWE-85] David G. LOWE. Visual recognition from spatial correspondence and perceptual organization. In *9th International Joint Conference on Artificial Intelligence*, pages 953–959, I.J.C.A.I., Los Angeles, August 1985.

[LUX-83a] Augustin LUX. Caiman: un système interactif pour l'analyse d'images. Note Interne, Laboratoire d'Informatique Fondamentale et d'Intelligence Artificielle, Octobre 1983.

[LUX-83b] Augustin LUX. *Intelligence Artificielle et Vision par Ordinateur: Analyse Bibliographique de Systèmes de Vision.* Rapport de Recherche LIFIA 1 IMAG 397, Laboratoire d'Informatique Fondamentale et d'Intelligence Artificielle, Novembre 1983.

[LUX-85a] Augustin LUX. Algorithmique et contrôle en vision par ordinateur. Rapport de Thèse d'Etat, Laboratoire d'Informatique Fondamentale et d'Intelligence Artificielle, Institut National Polytechnique de Grenoble, Septembre 1985.

[LUX-85b] Augustin LUX and Jose Luis GORDILLO. Synthesizing vision programs from robot task specifications. In Olivier Faugeras and George Giralt, editors, *Robotics Research: The 3rd International Symposium*, M.I.T. Press, Cambridge, 1986.

[MACKW-73] Alan K. MACKWORTH. Interpreting pictures of polyhedral scenes. *Artificial Intelligence*, 4:121–137, Summer 1973.

[MALIK-85] Jitendra MALIK. Labelling line drawings of curved objects. In *'85 Image Understanding Workshop*, pages 209–218, D.A.R.P.A., December 1985.

[MARR-78] David MARR. Representing visual information - a computational approach. In Allen R. Hanson and Edward M. Riseman, editors, *Computer Vision Systems*, Academic Press, New York, 1978.

[MARR-79] David MARR and Tomaso POGGIO. A computational theory of human stereo vision. *Proceedings of the Royal Society of London*, B-204:301–328, May 1979.

[MARR-80] David MARR and Ellen C. HILDRETH. Theory of edge detection. *Proceedings of the Royal Society of London*, B-207:187–217, February 1980.

[MARR-82] David MARR. *VISION: a Computational Investigation into the Human Representation and Processing of Visual Information.* W. H. Freeman and Company, San Francisco, 1982.

[MAZER-85] Emmanuel MAZER and Jean François MIRIBEL. *LE LANGAGE LM: manuel de référence.* CEPADUES Editions, Toulouse, Janvier 1985.

[MERO-75] L. MERO and Z. VASSY. A simplified and fast version of the hueckel operator for finding optimal edges in pictures. In *4th International Joint Conference on Artificial Intelligence*, pages 650–655, I.J.C.A.I., Tbilisi, September 1975.

[MOHR-84] R. MOHR and B. WROBEL. La correspondance en stéréovision vue comme une recherche d'un chemin optimal. In *4ème Congrès Reconnaissance des Formes et Intelligence Artificielle*, pages 71–79, R.F.I.A., Paris, Janvier 1984.

[NAGAO-80] Makoto NAGAO and Takashi MATSUYAMA. *A Structural Analysis of Complex Aerial Photographs.* Plenum Press, New York, 1980.

[NAGAO-84] Makoto NAGAO. Control strategies in pattern analysis. *Pattern Recognition*, 17(1):45–56, January 1984.

[NINIO-81] Jacques NINIO. Random-curve stereograms: a flexible tool for the study of binocular vision. *Perception*, 10:403–410, October 1981.

[NISHI-81] H. Keith NISHIHARA. Intensity, visible-surface, and volumetric representations. *Artificial Intelligence*, 17:265–284, August 1981.

[OHTA-80] Yu-ichi OHTA. *A region-oriented image-analysis system by computer.* PhD thesis, Kyoto University - Department of Information Science, March 1980.

[OHTA-83] Yu-ichi OHTA and Takeo KANADE. *Stereo by intra- and inter-scanline search using dynamic programming.* Research Report CMU CS-83-162, Carnegie-Mellon University - Department of Computer Science, October 1983.

[OSHIM-81] Masaki OSHIMA and Yoshiaki SHIRAI. Object recognition using three-dimensional information. In *7th International Joint Conference on Artificial Intelligence*, pages 601–609, I.J.C.A.I., Vancouver, August 1981.

[PENTL-82] Alex P. PENTLAND. Local computation of shape. In *'82 AAAI Conference*, pages 22–25, A.A.A.I., Pittsburgh, August 1982.

[PIAGE-61] Jean PIAGET. *Les Mécanismes Perceptifs.* Presses Universitaires de France, Paris, 1961.

[POGGI-84] Tomaso POGGIO. *Vision by Man and Machine: How the brain processes visual information may be suggested by studies in computer vision (and vice versa).* Research Report A.I. MEMO 776, Massachusetts Institute of Technology - Artificial Intelligence Laboratory, March 1984.

[PRAZD-82] K. PRAZDNY. The role of the eye position information in algorithms for stereoscopic matching. In *'82 AAAI Conference*, pages 1–4, A.A.A.I., Pittsburgh, August 1982.

[ROBER-65] Larry G. ROBERTS. Machine perception of three-dimensional solids. In J. T. Tippett and al, editors, *Optical and Electro-optical Information Processing*, M.I.T. Press, Cambridge, 1965.

[ROSEN-76] A. ROSENFELD R. A. HUMMEL and S. W. ZUCKER. Scene labeling by relaxation operations. *I.E.E.E. Transactions on Systems, Man, and Cybernetics*, SMC-6(6):420–433, June 1976.

[SAINT-86] Christian de SAINTE MARIE. The necessity of learning while doing. In *1st European Working Session on Learning*, L.R.I. - Unité de recherche AL KHOWARIZMI, Orsay, Février 1986.

[SHAFE-82] Steven A. SHAFER and Takeo KANADE. *Using Shadows in Finding Surface Orientations.* Research Report CMU CS-82-100, Carnegie-Mellon University - Department of Computer Science, January 1982.

[SHIRA-73] Yoshiaki SHIRAI and Hirochika INOUE. Guiding a robot by visual feedback in assembly tasks. *Pattern Recognition*, 5:99–108, 1973.

[SHIRA-75] Yoshiaki SHIRAI. Analysing intensity arrays using knowledge about scenes. In Patrick H. Winston, editor, *The Psychology of Computer Vision*, McGraw-Hill Book Company, New York, 1975.

[SHIRA-78] Yoshiaki SHIRAI. Recognition of real-world objects using edge cues. In Allen R. Hanson and Edward M. Riseman, editors, *Computer Vision Systems*, Academic Press, 1978.

[SOUVI-83] Viviane SOUVIGNIER. PVV: un système d'interprétation d'images par prédiction et vérification. Rapport de Thèse 3ème Cycle, Laboratoire d'Informatique Fondamentale et d'Intelligence Artificielle, Institut National Polytechnique de Grenoble, Juin 1983.

[STEVE-81] Kent A. STEVENS. The visual interpretation of surface contours. *Artificial Intelligence*, 17:47–73, August 1981.

[SUGIH-79] Kokichi SUGIHARA. Range-data analysis guided by a junction dictionary. *Artificial Intelligence*, 12:41–69, May 1979.

[SUGIH-84] Kokichi SUGIHARA. An algebraic approach to shape-from-image problems. *Artificial Intelligence*, 23:59–95, May 1984.

[TEOH-84] W. TEOH and X. D. ZHANG. An inexpensive stereoscopic vision system for robots. In *1st I.E.E.E. Conference on Robotics and Automation*, pages 186–189, I.E.E.E., Atlanta, March 1984.

[THONN-85] Monique THONNAT. *Automatic morphological description of galaxies and classification by an expert system.* Research Report 387, Institut National de Recherche en Informatique et en Automatique - Centre de Sophia-Antipolis, March 1985.

[THORP-83a] Charles E. THORPE and Steven SHAFER. *Topological Correspondence in Line Drawings of Multiple Views of Objects.* Research Report CMU CS-83-113, Carnegie-Mellon University - Department of Computer Science, March 1983.

[THORP-83b] Charles E. THORPE. *An Analysis of Interest Operators for FIDO.* Technical Report CMU RI-TR-83-19, Carnegie-Mellon University - The Robotics Institute, December 1983.

[TIBER-83] Guy TIBERGHIEN. La mémoire des visages. In *L'Année Psychologique*, XXX, XXX, 1983.

[TROCC-86] Jocelyne TROCCAZ. *Modélisation du Raisonnement Géométrique pour la Programmation des Robots.* PhD thesis, Laboratoire d'Informatique Fondamentale et d'Intelligence Artificielle, Institut National Polytechnique de Grenoble, Mars 1986.

[TSOTS-84] John K. TSOTSOS. Knowledge and the visual process: content, form and use. *Pattern Recognition*, 17(1):13–27, January 1984.

[TSUJI-81] Saburo TSUJI and H. NAKANO. Knowledge-based identification of artery branches in cine-angiograms. In *7th International Joint*

Conference on Artificial Intelligence, pages 710–715, I.J.C.A.I., Vancouver, August 1981.

[TSUJI-83] H. GUO M. YACHIDA, S. TSUJI and Y. Q. CHOU. Robot vision for determining three-dimensional geometry of flexible wires. In *1st International Conference on Advanced Robotics*, pages 133–138, I.C.A.R., Tokyo, 1983.

[TSUKI-85] Toshifumi TSUKIYAMA and Thomas S. HUANG. Motion stereo for navigation of autonomous vehicles in a passageway. In *3rd Workshop on Computer Vision: Representation and Control*, pages 148–155, I.E.E.E., Bellaire, October 1985.

[ULLMA-79] Shimon ULLMAN. *The Interpretation of Visual Motion.* M.I.T. Press, Cambridge, 1979.

[ULLMA-80] Shimon ULLMAN. *Against direct perception.* Research Report A.I. MEMO 574, Massachusetts Institute of Technology - Artificial Intelligence Laboratory, March 1980.

[VERNO-84] David VERNON. Robot vision in automated electrical wire crimping. In *1er Colloque Image*, pages 863–868, C.I., Biarritz, Mai 1984.

[WALTZ-75] David G. WALTZ. Understanding line drawings of scenes with shadows. In Patrick H. Winston, editor, *The Psychology of Computer Vision*, McGraw-Hill Book Company, New York, 1975.

[WITKI-81] Andrew P. WITKIN. Recovering surface shape and orientation from texture. *Artificial Intelligence*, 17:17–45, August 1981.

[WITKI-82] Andrew P. WITKIN. Intensity-based edge classification. In *'82 AAAI Conference*, pages 36–41, A.A.A.I., Pittsburgh, August 1982.

[WOODH-78] Robert J. WOODHAM. *Reflectance Map Techniques for Analyzing Surface Defects in Metal Castings.* PhD thesis, Massachusetts Institute of Technology - Artificial Intelligence Laboratory, June 1978.

[YACHI-82] M. YACHIDA S. TSUJI and X. HUANG. WIRESIGHT - A computer vision system for 3-D measurement and recognition of flexible wire using cross stripe light. In *6th International Joint Conference on Pattern Recognition*, pages 220–222, I.J.C.P.R., Munich, October 1982.

[YUILL-84] Alan L. YUILLE and Tomaso POGGIO. *A generalized ordering constraint for stereo correspondence.* Research Report A.I. MEMO 777, Massachusetts Institute of Technology - Artificial Intelligence Laboratory, May 1984.

AUTORISATION de SOUTENANCE

VU les dispositions de l'article 15 Titre III de l'arrêté du 5 juillet 1984 relatif aux études doctorales

VU les rapports de présentation de Messieurs

. M. BERTHOD, Directeur de recherche
. R. MOHR, Professeur

Monsieur Yves DEMAZEAU

est autorisé à présenter une thèse en soutenance en vue de l'obtention du diplôme de DOCTEUR de L'INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE, spécialité "Informatique".

Fait à Grenoble, le 15 décembre 1986

D. BLOCH
Président
de l'Institut National Polytechnique
de Grenoble

P.O. le Vice-Président,