Chapitre 1

Introduction

1.1. Qu'est-ce que la vision ?

L'espace qui nous entoure a une structure tri-dimensionnelle (3D). Lorsque l'on demande à une personne de décrire ce qu'elle *voit*, elle n'éprouve aucune difficulté à nommer les objets qui l'entourent : téléphone, table, livre... Et pourtant l'information qui est réellement disponible sur la rétine des ses yeux n'est, ni plus ni moins, une collection de points (environ un million !). En chaque point ou *pixel* (picture element) il y a tout simplement une information qui donne une indication quant à la quantité de lumière et la couleur qui proviennent de l'espace environnant et qui ont été projetées à cet endroit de la rétine. Le téléphone, la table ou le livre *n'existent pas* sur la rétine. Guidé à la fois par l'information codée dans l'image (ou la rétine) et par ses propres connaissances, le processus visuel *construit* des percepts. Le téléphone ou le livre sont le réponses finales, résultant d'un processus *d'interprétation* qui fait partie intégrante du système de vision. De plus, il n'y a pas de correspondance terme à terme entre l'information sensorielle (la lumière et la couleur) et la réponse finale (des objets 3D). Le système de vision doit *fournir* les connaissances nécessaires afin de permettre une interprétation non ambiguë.

1.2. Comprendre la vision

Il n'est pas suffisant de constater qu'un problème est complexe. Encore faut-il essayer de le comprendre dans ces moindres détails et de proposer une solution.

La vision a suscité l'intérêt de nombreux scientifiques et philosophes depuis déjà très longtemps. Parmi ceux-ci, les neurobiologistes mènent des recherches théoriques et expérimentales afin d'essayer de comprendre l'anatomie et le fonctionnement du cerveau dans son ensemble. Ils ont découvert une structure très complexe qui est loin de leur avoir révélé tous ses secrets. La tâche des neurobiologistes semble être à la fois grandiose et illusoire. Grandiose, parce que le cerveau est une des plus complexes *inventions* de la nature. Il reste et restera pour longtemps le bastion encore inconnu que les sciences humaines se proposent de conquérir. Illusoire, car on ne connaît pas ses limites. Ces limites ne sont-elles par repoussées à chaque découverte ? David Hubel [99] a merveilleusement bien exprimé ce paradoxe : *Le cerveau peut-il comprendre le cerveau ?*

Avec la naissance de machines de calcul de plus en plus sophistiquées, un certain nombre de scientifiques se sont attaqués au problème de la vision d'un point de vue quantitatif : est-il possible de construire un *modèle computationnel* pour la perception visuelle ? Attention : il ne s'agit pas de fournir une explication de comment marche la vision biologique mais de **créer un modèle** qui, vu de l'extérieur, possède des propriétés semblables.

Ce modèle *artificiel* peut-il être d'une utilité quelconque quant à la vision biologique ? Peut-il constituer la base d'une nouvelle technologie – *des machines qui voient* ?

Il est certainement trop tôt pour répondre à ces questions et pour tirer des conclusions. Malgré les efforts non négligeables, il y a très peu de résultats convaincants. Nous pensons que deux démarches doivent être suivies simultanément :

- essayer d'élaborer une théorie de la **vision par ordinateur** qui doit nous guider à long terme ;
- tenter de résoudre des problèmes spécifiques dans le cadre de cette théorie : de tels résultats partiels permettraient de confirmer ou au contraire de mettre en cause certains aspects de la théorie.

1.3. Une théorie de la vision

L'élaboration d'une théorie scientifique demande trois étapes :

1. énoncer la théorie, spécifier et élaborer les concepts de base : ces concepts doivent exprimer le cadre formel qui est à la base de la théorie,

2. exprimer ces concepts sous forme mathématique,

3. réaliser un ensemble expérimental qui permette de vérifier la théorie.

Voici comment la vision par ordinateur peut s'énoncer briévement dans les termes de ce paradigme. La vision est un processus de traitement de l'information. Elle utilise des stratégies bien définies afin d'atteindre ses buts. L'entrée d'un système de vision est constituée par une séquence d'images. Le système lui-même apporte un certain nombre de connaissances qui interviennent à tous les niveaux. La sortie est une description de l'entrée en termes d'objets et de relations entre ces objets.

Deux types de stratégies sont mises en jeu : ascendantes et descendantes. Les stratégies ascendantes tentent de construire à partir de l'information sensorielle une représentation la plus abstraite possible (par exemple, un ensemble de primitives géométriques 3D). Les stratégies descendantes déduisent à partir de l'ensemble d'objets connus par le système une description compatible avec les primitives extraites de l'image. Il est alors possible de mettre en correspondance la représentation extraite de l'image avec les descriptions des objets afin de décrire les données sensorielles en termes de ces objets.

Les connaissances mises en jeu peuvent être de trois types : physiques, géométriques et sémantiques. Les lois physiques imposent des contraintes aux signaux lumineux qui partant d'une source, traversent la scène et se projettent sur l'image. La gravitation impose à la scène (et donc à l'image) une structure hétérogène : prépondérance de lignes verticales et horizontales pour ne citer qu'un exemple. La forme des objets (l'ensemble de ses surfaces) et la géométrie de la formation de l'image imposent des contraintes très strictes quant aux structures susceptibles d'être présentes dans l'image. A un niveau plus élevé, un objet peut être décrit par sa fonction dans le contexte d'un raisonnement symbolique. Cette fonction n'est pas directement mesurable dans l'image. On devrait pouvoir dériver des contraintes sur la forme et l'emplacement d'un objet à partir de sa fonction. Par exemple, le mot *chaise* désigne une classe d'objets réels (un objet réel est un objet qui occupe une place dans l'espace physique). Cependant il y a une grande variété de chaises quant à la forme et à la couleur pour ne citer que deux propriétés. Quelles sont les propriétés communes à toutes les chaises, **mesurables** dans l'image ?

L'étape suivante consiste à exprimer ces stratégies et connaissances dans le cadre d'un formalisme mathématique et à construire les algorithmes correspondants. Les performances de ces algorithmes doivent correspondre aux qualités exigées d'un tel système : *la reconnaissance visuelle doit être fiable et rapide.*

1.4. Le paradigme de David Marr

Vers la fin des années 70, David Marr [121] a proposé un modèle calculatoire pour le traitement et la représentation de l'information visuelle. Voici quels sont les principaux traits de ce paradigme :

• à partir d'une ou de plusieurs images un processus d'extraction de caractéristiques produit un description en termes d'attributs bi-dimensionnels ; ce niveau de représentation est appelé *première ébauche* (primal sketch) ;

• la première ébauche constitue l'entrée d'un certain nombre de processus plus ou moins indépendants qui calculent des propriétés tri-dimensionnelles locales relatives à la scène ; il s'agit d'une représentation centrée sur l'observateur, appelée *ébauche 2.5D* ; ces processus opèrent sur une séquence d'images (analyse du mouvement) sur une paire d'images (stéréoscopie) ou sur une seule image. Dans ce dernier cas il s'agit de processus d'inférence qui utilisent des connaissances géométriques (analyse des contours), géométriques et statistiques (analyse des textures), photométriques (analyse des ombrages) ou colorimétriques (analyse des reflets) ;

• l'ébauche 2.5D est mise en correspondance avec des connaissances 3D afin de construire une description de la scène en termes d'objets et de relations entre les objets ; il s'agit maintenant d'une représentation *centrée sur la scène* (la description ne dépend plus de la position de l'observateur).

1.5. Segmentation, reconstruction, reconnaissance

En pratique, le paradigme de David Marr se traduit par trois étapes de traitement : segmentation, reconstruction et reconnaissance.

La segmentation d'images étant la pierre de base de tout système de vision, de nombreux travaux lui ont été consacrés. La diversité des images, la difficulté du problème, les origines variées des chercheurs, l'évolution de la puissance de calcul des ordinateurs, et un certain empirisme dans l'évaluation des résultats ont conduit à l'introduction d'une multitude d'algorithmes.

Quelle que soit son origine, une image constitue une représentation d'un univers composé d'entités : objets dans une scène d'intérieur, cellules, surfaces sismiques, organes du corps humain ... Le but de toute méthode de segmentation est l'extraction d'attributs caractérisant ces entités. Les attributs étudiés correspondent à des points d'intérêt ou à des zones caractéristiques de l'image : contours et régions. La détection des contours (chapitre 2) implique la recherche des discontinuités locales de la fonction des niveaux de gris de l'image. La segmentation des contours (chapitre 3) consiste à approximer les contours par des représentations analytiques, telles que des droites ou des coniques. L'extraction de régions (chapitre 4) revient à déterminer des zones homogènes en niveaux de gris de l'image. Par exemple, dans le cas d'images réelles, les contours correspondent aux frontières des objets et les régions à leurs surfaces. Ces deux approches "contour" et "région" sont duales en ce sens qu'une région définit une ligne par son contour, et qu'une ligne fermée définit une région. Elles amènent cependant à des algorithmes complètement différents et ne fournissant pas les mêmes résultats. Cette dualité est cependant peu exploitée dans la plupart des méthodes existantes.

Un autre aspect de la segmentation est celui qui consiste à retrouver la géométrie des objets à partir des images. On obtient ainsi des représentations intrinsèques, aisément manipulables et utilisables, à partir de la réalité physique induite par l'image. De manière à obtenir ces caractéristiques géométriques, on est souvent conduit à définir une suite hiérarchique de représentations de l'information image permettant finalement d'obtenir des indices visuels servant à résoudre une tâche donnée. Dans les chapitres 10 et 11 nous illustrons ces principes dans le cas des images bi-dimensionnelles et des images volumiques.

La calibration est la première étape indispensable pour toute méthode de reconstruction (à moins que la calibration ait lieu en même temps que la reconstruction). Le chapitre 5 décrit en détail les modèles géométriques de plusieurs capteurs basés sur une caméra ainsi que plusieurs techniques de détermination des paramètres de ces capteurs (calibration). On étudiera ainsi la caméra matricielle, la caméra linéaire ainsi que les capteurs stéréoscopiques passifs et actifs. Le chapitre 12 aborde le problème de la caractérisation des cartes de profondeur obtenues avec un capteur stéréoscopique actif (caméra et faisceau laser).

Le chapitre 6 décrit en détail les principes de reconstruction tri-dimensionnelle à partir d'un système stéréoscopique. Plus particulièrement, le problème de mise en correspondance stéréo est abordé d'un point de vue géométrique et algorithmique. La reconstruction de surfaces polyédriques par vision stéréoscopique à partir d'images 2D peut être réalisée par une approche régions ou par une approche contours. Le chapitre 10 présente une approche de type géométrie algorithmique pour résoudre le problème de reconstruction polyédrique.

La reconnaissance consiste essentiellement à comparer des indices visuels bi- ou tri-dimensionnels avec les indices des objets à reconnaître. Les méthodes de reconnaissance sont souvent couplées avec des méthodes de localisation. Le chapitre 7 décrit quelques méthodes de localisation (position et orientation avec six degrés de liberté) à partir d'indices visuels 3D mis en correspondance avec des indices d'objets 3D. Le chapitre 8 décrit une méthode de localisation à partir d'indices visuels 2D mis en correspondance avec des indices d'objets 3D. Enfin le chapitre 9 montre comment on peut combiner les méthodes de localisaton avec des méthodes de recherche arborescente pour pouvoir reconnaître des objets rigides.

1.6. Quelques références bibliographiques

Le premier ouvrage consacré partiellement à la vision par ordinateur est celui de Duda et Hart, datant de 1973 [50]. A une première partie consacrée à la reconnaissance des formes, fait suite une deuxième partie qui introduit les bases théoriques d'une approche géométrique de l'interprétation d'une image.

Les ouvrages de Gonzales et Wintz (1977) et de Rosenfeld et Kak (1982, seconde édition) passent en revue l'état de l'art de ce qu'on appelle aujourd'hui la "vision bas niveau" [73], [160].

Pendant longtemps, l'ouvrage de référence en vision par ordinateur a été celui de Ballard et Brown (1982), [15]. Sans rentrer dans les détails mathématiques, ce texte fournit une vue synthétique des travaux de recherche dans les années 80.

L'ouvrage de Horn, publié en 1986, aborde quelques aspects de la vision d'un point de vue plus fondamental [97]. Les bases mathématiques de la formation d'une image, de la détection de contours et de régions, des propriétés photométriques ainsi que de la perception du mouvement sont clairement présentées.

L'utilisation de la vision par ordinateur pour la navigation des robots est le thème de l'ouvrage d'Ayache, publié en 1989 [8] et en 1991 [9] (version anglaise). On y trouve notamment les détails de l'utilisation du filtre de Kalman étendu pour intégrer l'information provenant de plusieurs cartes stéréoscopiques.

Enfin, l'ouvrage de Faugeras est, avec le nôtre, le plus récent [55]. Il propose une approche géométrique (géometrie projective et euclidienne) pour résoudre notamment le problème de reconstruction. Très clair et très détaillé, contenant de nombreux exemples ainsi que des exercices, ce texte rend compte de 10 années de travaux de recherche effectués par l'auteur et par son équipe de l'INRIA.