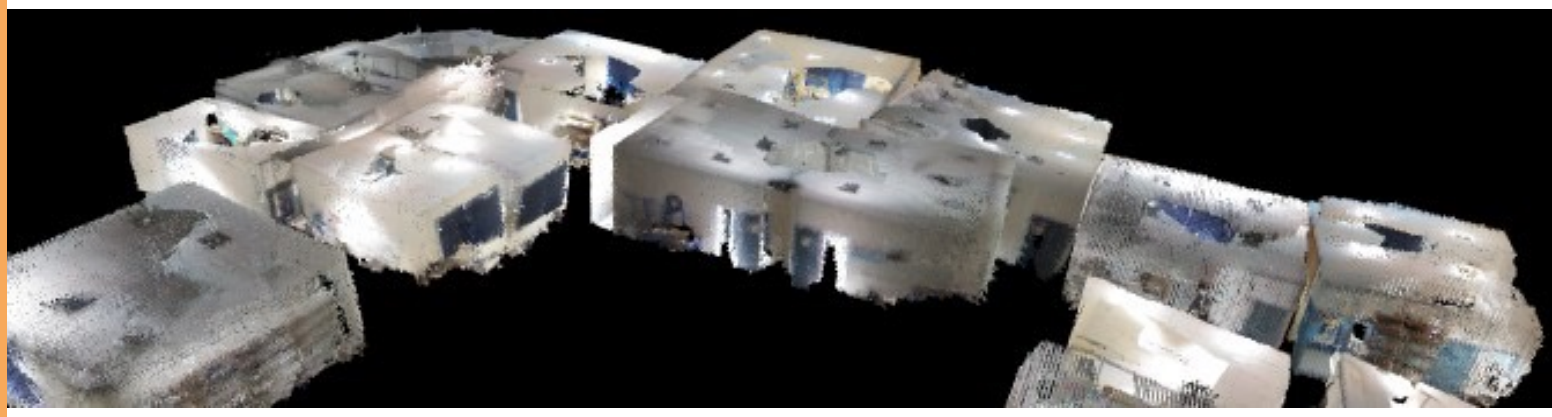

Cartographie 3-D du 1er étage du centre de recherche de l'Académie Militaire de Saint-Cyr Coëtquidan - Source J. Motsch.

La vision par ordinateur

Au-delà du système visuel humain ?

Jean Motsch
Maître de conférences à l'Académie Militaire de Saint-Cyr Coëtquidan

Déverrouiller son téléphone mobile en le regardant, générer un montage photo de votre chat en train de grandir, projeter un meuble que l'on envisage d'acheter dans une pièce pour voir s'il convient, trois opérations réalisables par nombre d'appareils nomades aujourd'hui. Leur point commun : ce sont des applications de la vision par ordinateur.

Une (pas si courte) introduction à la vision par ordinateur

Quelques éléments de définition

De manière naïve, les humains utilisent leurs yeux et leurs cerveaux pour voir et visualiser de manière sensible le monde qui les entoure. La vision par ordinateur cherche à faire de même, voire mieux, en utilisant une machine informatique. En développant un peu les deux définitions, on obtient que la vision (humaine) peut être décrite comme la perception du monde extérieur par la vue, un des cinq sens classiques. Plus précisément, la vision humaine concerne les rayonnements lumineux et leur interprétation cognitive. Par analogie, la vision par ordinateur s'occupe des méthodes permettant à une machine informatique de comprendre son environnement à partir d'images numériques ou de vidéos. Par extension, les imageurs peuvent aller au-delà du spectre visible, comme le spectre infrarouge, et d'autres capteurs peuvent être utilisés, comme des capteurs de profondeurs (LIDAR par exemple). Pour un ingénieur, il s'agira également de comprendre et d'automatiser les différentes tâches réalisées par le système visuel humain.

Parmi ces tâches se retrouvent des méthodes pour acquérir, traiter, analyser et comprendre les images numériques et l'extraction de données sur le monde concret pour produire des informations numériques ou symboliques. La multiplicité des tâches requises dans une application de vision par ordinateur a généré dans son histoire de nombreuses (sous-) disciplines connexes.

Une discipline à la croisée d'autres disciplines

La vision par ordinateur est connexe d'autres disciplines techniques et médicales tout en recouvrant nombre de sous-domaines. Parmi les domaines connexes, on trouvera :

• **Électronique et physique du solide** qui fournissent d'une part les capteurs (souvent matriciels) adaptés, et d'autre part les processeurs aux capacités de calcul autorisant les traitements les plus complexes en temps réel. Les *Systems On Chips* (SoC)

présents dans tous les dispositifs mobiles en sont le meilleur exemple ;

• **Traitement du signal** dont les outils pour les signaux 1-D peuvent facilement être étendus aux signaux 2-D ou N-D. Néanmoins, les spécificités des images sont prises en compte en **traitement d'image** avec des outils adaptés ;

• **Mathématiques (appliquées)** en particulier les outils statistiques (description et estimation), les méthodes d'optimisation, la géométrie (projective) et l'algèbre linéaire. De plus, les implémentations efficaces et robustes des méthodes de résolution ne sont pas oubliées ;

• **Neurobiologie** qui permet de proposer des méthodes fonctionnant par mimétisme du système visuel humain. Les réseaux de neurones et l'apprentissage profond tirent en partie profit des connaissances en ce domaine.

La variété des domaines connexes se retrouve dans les domaines conjoints à la vision par ordinateur et la frontière entre les disciplines est parfois difficile à cerner. Les domaines les plus proches de la vision par ordinateur sont le traitement d'image, l'analyse d'image et la vision par machine. Les outils fondamentaux sont souvent très proches voire identiques, et les disciplines sont souvent confondues. De même, la photogrammétrie qui permet d'obtenir un relief (cartographique) à partir de paires d'images est souvent considérée comme étant de la stéréoscopie par ordinateur. Enfin, une discipline inverse comme l'infographie se retrouve associée à la vision par ordinateur dans les applications de réalité augmentée. On peut donc essayer de distinguer les disciplines de la manière suivante :

• **Traitement d'image** et **analyse d'image** se concentrent sur les images 2-D et les opérations considérées comme de bas niveau comme les filtrages, la réduction de bruit, l'extraction de contours, les transformations géométriques. Ces disciplines vont jusqu'à fournir des primitives qui pourront être interprétées ultérieurement ;

• **Vision par ordinateur** (*stricto sensu*) inclut l'analyse 3-D de l'environnement à partir d'images 2-D. Il s'agit d'inverser la projection d'une scène 3-D en une ou plusieurs images 2-D. Autrement dit, comment reconstruire la structure 3-D de l'environnement, ou autre information, à partir d'une ou plusieurs images. Ce problème étant intrinsèquement **mal posé**, cette discipline nécessite de faire des hypothèses plus ou moins nombreuses et complexes sur la scène représentée dans les images ;

• **Vision par machine** regroupe les différentes disciplines et techniques associées aux applications industrielles que sont l'inspection automatique (dans le contrôle qualité de pièces), la surveillance de processus et le contrôle de robots industriels (comme les bras manipulateurs). Dans ce contexte, les capteurs ainsi que les asservissements sont intégrés au traitement des images dans des applications souvent soumises à des contraintes d'exécution en temps-réel. Les implémentations sont donc souvent optimisées, tant au niveau matériel qu'au niveau logiciel, avec souvent un contrôle fin des conditions d'illumination. Ceci est une différence majeure avec les autres disciplines qui, de manière générale, ne contrôlent pas les paramètres de l'environnement de la scène observée. Lorsque les robots sont autonomes, le terme de **vision robotique** est également utilisé ;

• **Imagerie (médicale)** qui va souvent au-delà de la simple prise de vue mais intègre des outils de mesure et de détection de pathologies ;

• **Reconnaissance de forme** qui utilise des méthodes statistiques ou des réseaux de neurones pour extraire de l'information à partir de signaux. Une grande partie de cette discipline est appliquée à des images.

La multiplicité des disciplines et des sous-disciplines ainsi que leurs proximités thématiques font que les contours de chacune d'entre elles sont flous. Cette situation sera renforcée lorsque l'on regarde les applications se retrouvant sous les auspices de la vision par ordinateur.

Des applications tous azimuts

Partant de données de grandes dimensions, plusieurs dizaines de millions de pixels en couleur par exemple, les applications de vision par ordinateur cherchent à fournir des informations numériques ou symboliques, éventuellement à l'aide de décisions, pour comprendre la scène observée. Dans ce contexte, cette compréhension (d'une image) consiste à démêler l'information symbolique sous-jacente, de dimension réduite, des données (images) volumineuses, à l'aide de modèles issus de la géométrie, de la physique, des statistiques et des techniques d'apprentissage.

Les applications de vision par ordinateur peuvent être placées dans une des catégories suivantes :

• Les applications de reconnaissance qui se chargent de :

- détecter un objet (une cible) ou une situation particulière ;
- classifier un objet en fournissant également sa position (2-D) ou sa pose (3-D) ;
- identifier une instance particulière d'un objet.

• Les applications d'analyse du mouvement qui :

- déterminent les mouvements 3-D d'une caméra à partir d'une séquence d'images ;
- suivent le mouvement d'un ensemble de points d'intérêt qui peut constituer une cible ;
- estiment le flot optique.

• Les applications de reconstruction de scène 3-D ;

• Les applications d'amélioration d'images.

Figure 1 : Quelques applications (d'en haut à gauche à en bas à droite) détection de visages, image panoramique obtenue avec un mouvement de rotation, description automatique d'une image (outil DenseCap) et reconnaissance automatique de mots (source REE 2022-2 p. 112).

Parmi les applications industrielles ou grand public de la vision par ordinateur, on retrouve :

• La reconnaissance optique de caractères ;

• La stabilisation (souvent verticale) des prises de vues ;

• L'authentification visuelle, souvent par les visages ;

• La surveillance du trafic routier ;

• La sécurité des véhicules, de la voiture au bateau à voiles de course ;

• Le comptage du nombre de personnes dans un espace donné ;

• L'annotation automatique d'images ou de vidéos ;

• La construction d'images nouvelles à partir d'une description textuelle.

Ainsi, il semble bien que l'association imageur et puissance de calcul permette d'obtenir des résultats spectaculaires qui semblent mettre au défi le système visuel humain. Néanmoins, l'état de l'art des applications n'est que le résultat d'une (assez) longue histoire technique dans laquelle les connaissances sur l'homme ont été parfois utilisées pour progresser. Donc, avant de faire un tour historique, il peut être nécessaire de faire un tour neurobiologique du système visuel humain.

Le système visuel humain : une référence ?

La vision est une tâche que l'être humain réalise sans aucune difficulté apparente. On peut donc aisément imaginer que l'on puisse construire un système de vision par ordinateur en imitant le comportement des différents éléments du système visuel humain ainsi que tous les processus qui s'y déroulent. Ainsi énoncé, ce mimétisme semble évident, mais la description qui suit, quoique succincte, montrera qu'il n'en est rien.

Une structure anatomique complexe

Le système visuel humain est principalement constitué de l'œil, des nerfs optiques, du chiasma optique, du tractus optique, du corps géniculé latéral, des radiations optiques et du cortex visuel. Dans une approche simplifiée, l'œil peut être assimilé à un système de prise de vue, avec la pupille comme diaphragme, la cornée et le cristallin comme objectif et la rétine comme surface photosensible. L'image obtenue est donc formée par une projection perspective inversée de l'environnement.

La transduction de l'énergie lumineuse en potentiel d'actions se fait sur la rétine avec une analyse locale en zones de contraste

"Dans une approche simplifiée, l'œil peut être assimilé à un système de prise de vue, avec la pupille comme diaphragme, la cornée et le cristallin comme objectif et la rétine comme surface photosensible."

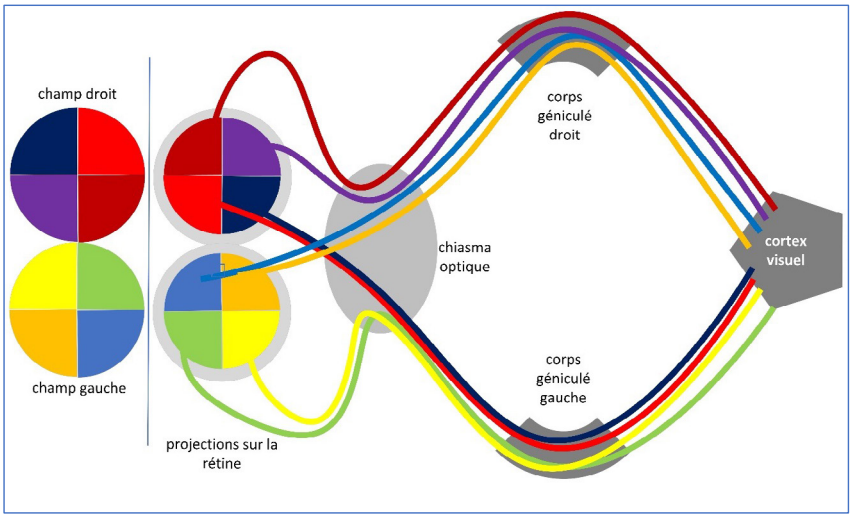

Figure 2 : Description schématique du câblage optique humain.

et le résultat de ce traitement est envoyé à travers le nerf optique au reste du système visuel. Les informations sont traitées par quadrant dans les vues gauche et droite comme illustré sur la figure 2. Il faut noter que chez l'être humain, le système visuel est le seul système sensoriel à être directement connecté, par le nerf optique, au cerveau, pour permettre de traiter rapidement l'information visuelle.

Des traitements successifs en suivant le nerf optique

Les nerfs optiques de chaque œil se rejoignent au niveau du *chiasma optique*, ce qui permet de redistribuer l'information visuelle. Ainsi, les quadrants (hémichamps) se retrouvent dans l'hémisphère opposé du cerveau. Le *tractus optique* qui suit distribue dans chaque hémisphère l'information visuelle à différents centres dont les *corps géniculés latéraux* (CGL) pour la perception visuelle consciente, l'hypothalamus pour la synchronisation du cycle éveil/sommeil, le *colliculus* supérieur qui est chargé de diriger les récepteurs de la tête vers des objets d'intérêt et le *pulvinar* qui gère les saccades (petits mouvements rapides des yeux).

La plupart des fibres du nerf optique s'arrêtent au niveau du CGL. Celui-ci évalue la distance des objets et ajoute aux objets principaux une indication de vitesse avant de transmettre les impulsions à la zone V1 du cortex visuel. Cette indication de vitesse permettra de prédire le déplacement des objets suivis. Le CGL est également relié de manière plus parcimonieuse aux zones V2 et V3 du cortex visuel. Les aires corticales effectuent les opérations suivantes :

• V1 : détection de contours pour comprendre l'organisation spatiale, construction d'une carte de points d'intérêt pour orienter l'attention ;

• V2 : relié au pulvinar et V1, avec un rôle proche, prise en compte des faux-contours, détermination de la profondeur à partir de la disparité binoculaire et distinction entre avant et arrière-plan ;

• V3 : prise en charge du mouvement global des objets (en module et direction) ;

• V4 : reconnaissance des formes simples ;

• V5 : intégration de mouvements locaux des objets à leur mouvement global et analyse du mouvement propre ;

• V6 : analyse du mouvement des objets relativement au fond de la scène.

Les opérations de reconnaissance de formes complexes, des objets ou des visages se font dans le *gyrus temporal inférieur*. Associé à l'hippocampe, il permettra la mémorisation de ceux-ci.

On peut ajouter à cette description sommaire l'hypothèse de la coexistence de deux voies visuelles neuro-physiologiquement distinctes. D'une part, la voie du *système magnocellulaire* qui serait spécialisée dans trois informations : le mouvement, la profondeur et la forme globale. Cette voie traiterait les basses fréquences spatiales des cellules périphériques de la rétine, sensibles en basse lumière. D'autre part, la voie du *système parvocellulaire* qui s'occuperait de la couleur, des détails des formes et des textures, en

Le système visuel humain en quelques chiffres

• Même si le système visuel humain voit en couleur à la lumière du jour, il y a environ 91 millions de bâtonnets (pour la lumière) pour 4,5 millions de cônes (pour les couleurs). Cela n'empêche pas que dans la fovéa, la zone la plus sensible de la rétine, on trouve au contraire deux cents fois plus de cônes que de bâtonnets !

• Il y a environ 100 milliards de neurones dans le cerveau humain et chaque neurone est connecté en moyenne à 10 000 autres neurones. Il pourrait y avoir jusqu'à un million de milliards de synapses. La capacité mémoire du cerveau humain est évaluée entre 1 et 1000 To. Au repos, un cerveau consomme environ 20 % de l'énergie métabolique soit en moyenne 12 W ;

• La vue représente environ 80% des perceptions d'un être humain. 50 % du cortex cérébral est utilisé pour le traitement des informations. On considère que le cerveau humain traite les images environ soixante mille fois plus rapidement que le texte.

La puissance du système visuel humain est qu'il fournit des résultats excellents, en un temps limité et avec une consommation énergétique très faible. Qui dit mieux ?

"La principale distinction entre vision par ordinateur et traitement d'image est la volonté de retrouver la structure tridimensionnelle du monde à partir d'images et d'utiliser cette pierre angulaire pour accéder à une compréhension complète de la scène. "

traitant les hautes fréquences spatiales et traiterait plutôt les informations du centre de la rétine dont celles issues de la fovéa, sa zone la plus sensible.

On le voit, les traitements semblent se faire dans un ordre de complexité croissante, en partant d'opérations de bas niveau jusqu'à atteindre une description symbolique de la scène observée. Même avec cette description simplifiée, il semble impossible de décrire exactement le système visuel humain à l'aide d'un modèle numérique.

D'un sujet de stage à l'intelligence artificielle : une (petite) histoire de la vision par ordinateur

La petite histoire retient que concevoir un système de vision par ordinateur fut posé comme un sujet de stage au M.I.T. en 1966. *The Summer Vision Project* [1] devait utiliser les étudiants travaillant l'été pour construire *une part significative d'un système visuel*. L'objectif principal en est la segmentation de l'image pour séparer les objets, les zones du fond et le bruit (*chaos*). Un objectif intermédiaire est d'être capable d'analyser une scène contenant des objets ne se recouvrant pas comme des balles, des briques de couleurs ou textures (identiques ou différentes) et des cylindres.

Les sous-problèmes à résoudre énoncés dans le projet sont pertinents à plus d'un titre :

• Étudier des propriétés de surfaces observées, ponctuelles et locales ;

• Connecter les régions dans l'image présentant des propriétés constantes ;

• Établir une liste de propriétés à considérer pour les régions dans l'image, comme l'aire, les contours, la boite englobante, les moments, les régions connexes ;

• Étudier les propriétés de courbes élémentaires : segment droit, morceau convexe, morceau elliptique, élément lisse ;

• Segmenter les arêtes d'un cylindre.

Même si penser résoudre une application de la vision par ordinateur en partant de zéro en un été semble bien naïf aujourd'hui, certains des problèmes évoqués dès le *Summer Vision Project* restent d'actualité.

Une évolution sur plus de 50 ans

Des débuts aux années 1970

La vision par ordinateur apparaît à ce moment-là comme étant une composante d'un agenda beaucoup plus ambitieux qui cherchait à imiter l'intelligence humaine. Il s'agissait également de doter les robots de comportements intelligents. Comme le *Summer Vision Project* le rappelle, la vision par ordinateur est vue par les pionniers de l'intelligence artificielle et de la robotique comme une étape élémentaire sur le chemin visant à résoudre des problèmes de plus haut niveau comme le raisonnement et la planification.

La principale distinction entre vision par ordinateur et traitement d'image est la volonté de retrouver la structure tridimensionnelle du monde à partir d'images et d'utiliser cette pierre angulaire pour accéder à une compréhension complète de la scène. Il s'agit donc d'extraire des primitives dans l'image comme des contours, des lignes afin de déterminer une structure en blocs ou en cylindres généralisés, c'est-à-dire des solides de révolution avec des contours clos. Les relations entre les différents objets sont également modélisées à l'aide de ressorts simulant des forces de rappel et des positions relatives à ajuster.

Parmi les autres considérations des chercheurs de cette période on retrouve des approches dites qualitatives ou quantitatives, suivant la nature des informations extraites des images. Ainsi, la compréhension des mécanismes expliquant les intensités et leurs dégradés à partir des mécanismes de formation d'images prenant en compte l'orientation des surfaces 3-D et les ombres fournissent une description 2½-D qualitative. De même, les approches quantitatives s'intéressent pour la première fois à apparier des primitives dans des paires d'images pour faire de la stéréoscopie. C'est aussi le cas du calcul du flot optique reposant sur les intensités lumineuses. Et, de manière conjointe, les chercheurs commencent à estimer à la fois la structure 3-D de la scène ainsi que les mouvements de la caméra.

À la fin de cette période, David Marr, dans son ouvrage séminal [2], résume assez bien la pensée qui régit la manière dont la vision par ordinateur est perçue. En particulier, Marr introduit la notion de trois niveaux de description d'un système de traitement d'informations (visuelles). On peut les appréhender de la manière suivante :

1 https://people.csail.mit.edu/brooks/idocs/AIM-100.pdf

2 "Vision: A Computational Investigation into the Human Representation and Processing of Visual Information"

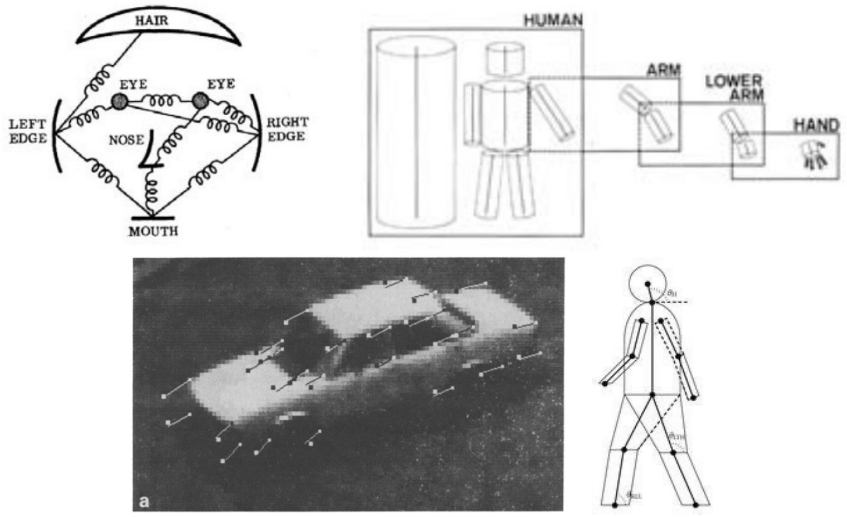

Figure 3 : Exemples des années 1970 à 1980 (d'en haut à g. à en bas à dr.): modèle pictural de la tête de Fischler & Elschlager, modèle articulé humain de Marr, quelques vecteurs de mouvement obtenus par le calcul du flot optique par Nagel entre trames successives d'une vidéo, modèle pictural du corps humain de Borgefors.

• **Les considérations calculatoires :** quel est l'objectif de la tâche à remplir et quelles sont les contraintes connues ou qui peuvent être apportées pour attaquer le problème ?

• **Les représentations et les algorithmes :** quelles sont les données fournies en entrée, attendues en sortie, et, surtout, quelles sont les représentations intermédiaires (primitives, points d'intérêt, par exemple) ? Quels sont les algorithmes mis en œuvre pour calculer les résultats requis ?

• **L'implémentation matérielle :** comment associer les représentations et les algorithmes au matériel disponible ? De manière inverse, comment peut-on tirer profit des contraintes matérielles pour choisir les représentations et algorithmes ? Un exemple actuel est l'utilisation de processeurs graphiques (*GPU*) pour faire de l'apprentissage profond.

En pratique, le paradigme développé par Marr se traduit souvent par trois étapes de traitement :

• **Segmentation :** c'est l'extraction de primitives ou d'attributs caractérisant les entités présentes dans l'image ;

• **Reconstruction :** c'est le passage 2-D vers 3-D, souvent en prenant comme point de départ les caractéristiques de la caméra, sa calibration ;

• **Reconnaissance :** l'identification par association d'attributs entre objets connus et à reconnaître.

Force est de constater que les idées de Marr sont suffisamment générales pour garder toute leur pertinence.

Les années 1980

Cette période se concentre sur le développement d'outils mathématiques plus sophistiqués pour analyser de manière quantitative des images.

Parmi ces outils, les pyramides d'images sont largement utilisées pour faire de la fusion d'images et de la mise en correspondance *coarse-to-fine*. Le concept de traitement *scale-space* en sera la version continue. Ces approches seront couronnées par l'arrivée des ondelettes.

L'utilisation de la stéréo comme indice quantitatif de forme sera étendue à toute une série de méthodes appelées *shape-from-X* : la forme par analyse des ombres (*shape-from-shading*), la stéréo photométrique, la forme par analyse de la texture (*shape-from-texture*) et la forme par analyse du flou (*shape-from-focus*).

Les chercheurs se rendent également compte qu'un cadre commun peut unifier, ou au moins décrire, les pro-

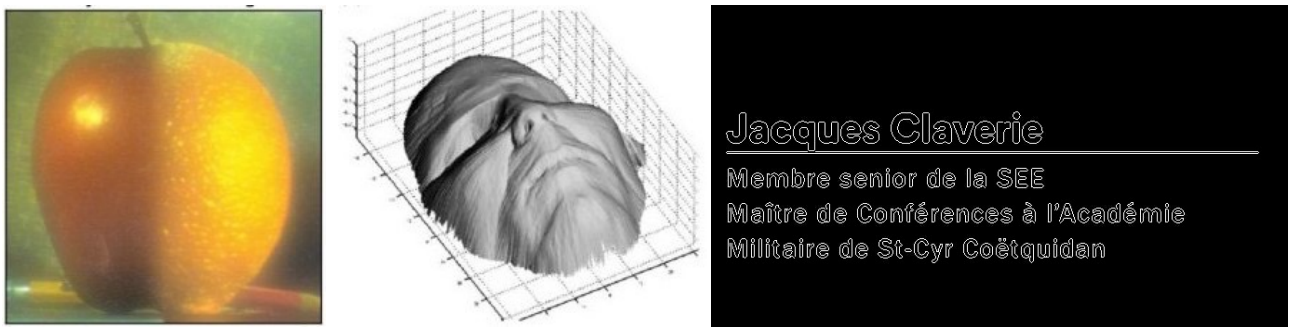

Figure 4 : Exemples des années 1980 à 1990 (de g. à dr.) mélange pyramidal pomme-orange, reconstruction d'un modèle 3-D à partir des ombres d'une image (shape-from-shading par Faugeras & al.), détection de contours (source REE 2022-2 p. 112).

blèmes de stéréoscopie, de flot optique, de *shape-from-X* et de détection de contours. Ce cadre est celui des *problèmes d'optimisations variationnelles* qui peuvent être rendus robustes (ou bien posés) en utilisant des méthodes de régularisation. C'est également à cette période que les mêmes problèmes sont formulés en utilisant des champs de Markov aléatoires (*Markov Random Fields*) qui permettent d'utiliser de meilleures méthodes de recherche d'optima (globaux), comme le recuit simulé.

Enfin, cette période ne néglige pas le développement du traitement de données de profondeur (3-D) pour leur acquisition, leur fusion, leur représentation et leur reconnaissance.

Les années 1990

Dans le domaine de la 3-D, la prise en compte des invariants projectifs et les reconstructions projectives se développent. N'ayant pas besoin de calibrer la caméra, les modes reconstruits le sont dans un espace projectif, non euclidien (on ne peut pas mesurer de distances, entre autres), ce qui est suffisant pour nombre d'applications. Les premiers modèles 3-D épars générés automatiquement apparaissent. Les algorithmes de stéréo multi-vues permettent de produire de leur côté des surfaces 3-D complètes.

Les techniques de flot optique s'améliorent ainsi que les algorithmes denses de mise en correspondance. Le progrès majeur dans ce domaine qui nécessite de faire une optimisation globale est la formulation du problème sous forme de graphe et sa résolution par des techniques d'élagage.

Les algorithmes de suivi progressent en se concentrant sur les contours actifs, tels que les snakes, les filtres particulaires et les ensembles de niveaux (*level-sets*). Les techniques directes pour suivre des visages ou bien des corps entiers ne sont pas oubliées.

Il faut relever que c'est aussi une période qui voit poindre les techniques d'apprentissage statistique, avec en pionnier l'application de l'analyse en composantes principales (*eigenfaces*) pour la reconnaissance de visages. De même, des courbes complexes peuvent être suivies en utilisant des systèmes dynamiques linéaires. Enfin, le lien entre infographie et vision par ordinateur se renforce dans cette décade pour obtenir de meilleurs rendus d'images de synthèse utilisant des images réelles. On trouve pêle-mêle des techniques de mélange de vues comme le *morphing*, les premières images panoramiques ainsi que la création des premiers mondes 3-D réalistes.

Les années 2000

Cette période est marquée par trois phénomènes pour la vision par ordinateur. D'abord, l'explosion de la photographie numérique, désormais accessible au grand public, avec des capteurs et des mémoires de plus en plus performants. Ensuite, la puissance de calcul disponible, même pour le grand public, autorise des applications de plus en plus gourmandes en ressources matérielles. Enfin, Internet permet la circulation de données de manière massive, en particulier des images. Et ce n'est qu'un début.

Parmi les algorithmes qui profitent de l'émergence de ce nouveau marché, l'obtention d'images à haute dynamique (*HDR*) permet d'augmenter les capacités du capteur en prenant plusieurs images avec des paramètres différents puis de les fusionner tout en conservant les tons (*tone mapping*). Les photographes argentiques se souviendront de l'art du *bracketing*. Dans la même veine, on trouve la fusion d'images prise avec et sans flash.

La reconnaissance d'objet profite de son côté de l'essor des techniques reposant sur la caractérisation (*features*) de morceaux (*patch*) de l'image. Ces *features* ont comme principale propriété d'être invariantes par rotation et changement d'échelle. Elles reposent souvent sur des histogrammes directionnels normalisés dont le calcul est optimisé pour réduire les durées de traitement. Leur appariement

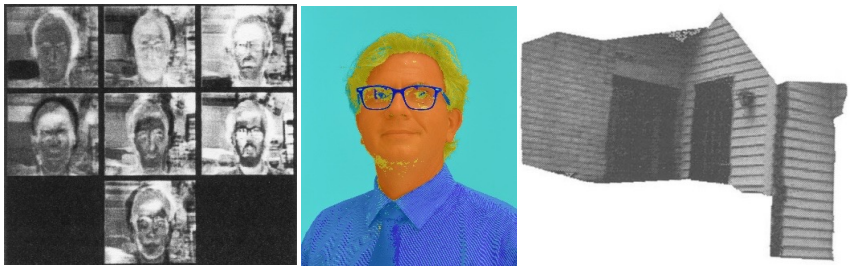

Figure 5 : Exemples des années 1990 à 2000 (de g. à dr.) eigenfaces de Turk & Pentland (les sept premiers vecteurs d'une base orthonormées de l'espace des visages), segmentation couleurs d'une image, obtention d'un modèle 3-D dense à partir du mouvement de la caméra (structure-from-motion de Tomasi & Kanade).

L'auteur

Jean Motsch. Ancien élève de l'ENS de Cachan et de l'ENSPS, lauréat de l'agrégation de physique appliquée, docteur en traitement du signal et de l'image, il est maître de conférences à l'Académie Militaire de Saint-Cyr Coëtquidan, en service détaché de l'INSA Rennes (section 61). Membre du Centre de recherche de Coëtquidan (CReC), ses activités concernent la vision par ordinateur, en particulier en vision robotique pour des applications de cartographie ou de détection d'objectifs, dans un contexte embarqué et collaboratif. Jean Motsch est également directeur du semestre (M1) Smart Robotics.

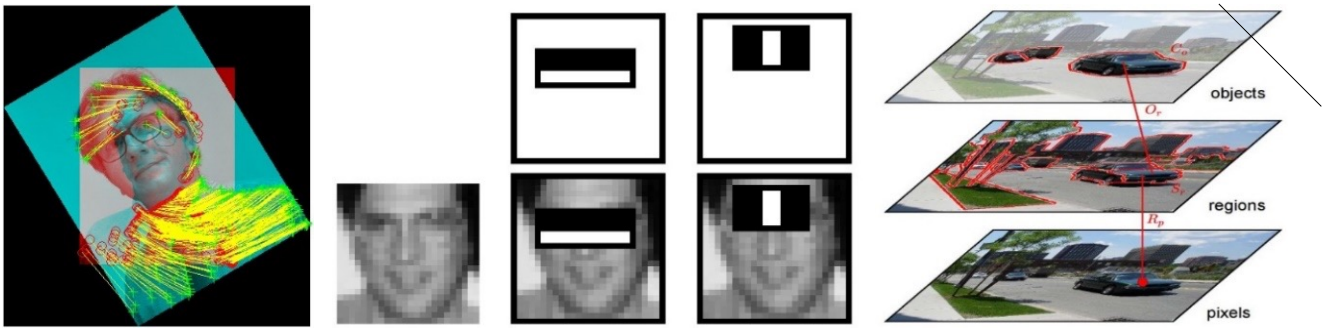

Figure 6 : Exemples des années 2000 à 2010 (de g. à dr.) appariements de points d'intérêt pour du recalage d'images, illustration de la pertinence des ondelettes de Haar pour la détection de visage (méthode de Viola & Jones), segmentation basée région (méthode de Gould & al.).

"L'arrivée massive des smartphones a renforcé le poids des images (numériques) dans la vie quotidienne, tout en mettant dans chaque poche une puissance de calcul sans commune mesure avec celle disponible dans les laboratoires des années 1980."

(*matching*) est en lui-même une tâche complexe et des techniques apparaissent pour fournir un résultat probable en un temps raisonnable.

En arrière-plan, on voit le développement d'algorithmes toujours plus efficaces pour résoudre des problèmes complexes d'optimisation globale en allant au-delà des élagages de graphes.

La grande tendance de cette fin de période est l'application massive de techniques sophistiquées d'apprentissage machine pour résoudre des problèmes de vision par ordinateur. Cette tendance coïncide avec la disponibilité croissante d'une très grande quantité de données (images) partiellement étiquetées sur Internet. Ceci rend possible l'apprentissage des catégories d'objet sans nécessiter une supervision humaine incompatible avec les volumes de données à traiter.

Depuis 2010

L'arrivée massive des *smartphones* a renforcé le poids des images (numériques) dans la vie quotidienne, tout en mettant dans chaque poche une puissance de calcul sans commune mesure avec celle disponible dans les laboratoires des années 1980.

Le principal phénomène observé est l'explosion de l'application des méthodes d'apprentissage machine massif dans tous les domaines, vision par ordinateur comprise. Ces méthodes d'apprentissage, souvent supervisées, se retrouvent sous le vocable de réseaux de neurones (*neural networks* ou *neural nets*) et d'apprentissage profond (*deep learning*). Il faut noter que ces deux expressions d'une part cachent une grande diversité de structures et d'autre part, la profondeur concerne la forme des réseaux concernés.

Dans le contexte de la vision par ordinateur, et plus particulièrement dans les applications de reconnaissance d'objets (leur classification), l'apprentissage marque un

Figure 7 : Exemples depuis les années 2010 (de g. à dr.) : étiquetage automatique d'une image, cartographie intérieure 3-D par méthode RGBD-SLAM, traduction en temps réel de texte présent dans une scène (source REE 2022-2 p. 113).

changement de paradigme et obtient des résultats bien meilleurs que les méthodes classiques. Plus précisément, il s'agit de réseaux de type *DCNN* (*deep convolutional neural nets*), des réseaux de neurones convolutifs profonds. Pour simplifier, la reconnaissance d'objets procède en deux actions : d'abord, extraire des primitives *ad hoc* à l'application recherchée puis, classifier les primitives. Souvent, les primitives retenues, ainsi que suggéré par Marr, sont choisies de manière *artisanale*, à partir de la connaissance du problème à résoudre. Les primitives ne sont pas les mêmes pour la détection de visages que pour l'identification des véhicules blindés. Les *DCNN* rendent inutiles cet artisanat : la partie profonde convolutive apprend les filtres permettant d'extraire les primitives adaptées au problème à résoudre.

Cette apparente panacée présente néanmoins quelques inconvénients majeurs : d'abord, il faut des volumes de données très grands pour bien apprendre. Ensuite, la durée de l'apprentissage peut se compter en semaines, même si l'utilisation de processeurs graphiques réduit d'un ordre de grandeur les temps de calcul. Enfin, les réseaux de neurones eux-mêmes ont besoin d'être paramétrés à l'aide d'heuristiques aux fondations peu profondes.

Néanmoins, le *deep learning* doit aussi sa popularité à plusieurs points forts. D'abord, l'apprentissage par transfert, qui permet d'utiliser la partie profonde d'un réseau pour extraire des primitives génériques. Il ne reste qu'à apprendre la couche de classification avec un gain de temps d'apprentissage évident. On peut ainsi réutiliser un réseau *champion* pour le spécialiser. Ensuite, si l'apprentissage est long, l'utilisation des réseaux est beaucoup plus rapide. Il est ainsi possible de disposer de systèmes de reconnaissances en temps réel embarqués. Enfin, les applications grand public étant nombreuses, les industriels sont poussés à fournir des *frameworks* pour généraliser les tâches de vision par ordinateur.

Et ensuite ?

Par son ubiquité dans les applications grand public et sa pertinence pour les industriels, les outils de vision par ordinateur sont disséminés dans la société. En termes économiques, le marché global de la vision par ordinateur évolue rapidement, avec des revenus passant d'un peu moins de 6 milliards de dollars en 2014 à environ 12 milliards en 2022, avec une projection à 20 milliards en 2030.

Néanmoins, ce marché ne comporte pas que des applications grand public ou ludiques. Les applications de reconnaissance de visage associées à des outils de vidéosurveillance sont, à juste titre, mises à l'épreuve de l'acceptation sociale. Le droit à la vie privée vient en effet se confronter à des impératifs de sécurité sur la voie publique. En particulier si les décideurs oublient que les systèmes automatiques font toujours des erreurs.

Enfin, d'un point de vue scientifique, la vision par ordinateur présente encore de nombreux défis. L'explicabilité des résultats en reconnaissance, quelle que soit la méthode utilisée, est un champ qui demande à être exploré afin de permettre de considérer ces outils autrement que comme des boîtes noires. ■

Bibliographie

• Richard Szeliski, « Computer Vision : Algorithms and Applications », 2ème édition, 1232p, Springer, 2022

• Peter Corke, « Robotics, Vision and Control », 2ème édition, 697p, Springer, 2017

• Richard Hartley et Andrew Zisserman, « Multiple View Geometry in Computer Vision », 2ème édition, Cambridge University Press, 2003.

Résumé

Environ 50 ans après ses débuts, les applications de vision par ordinateur se retrouvent disséminées dans tous les secteurs d'activités, de la vie quotidienne à l'usine, en passant par le divertissement, le commerce, les transports et la défense. Comme domaine scientifique, la vision par ordinateur en rejoint de nombreux autres comme le traitement d'image tout en couvrant sous un même vocable un grand nombre de sous-domaines comme la stéréoscopie ou la vision robotique. L'augmentation incessante de la puissance de calcul disponible, d'une part, et l'évolution, plus récente, des imageurs lui a permis de proposer des modèles théoriques de plus en plus fins et des algorithmes de plus en plus complexes. Désormais, la généralisation des dispositifs nomades, téléphones mobiles en particulier, distribue les outils de vision par ordinateur jusque dans nos poches. Tout ceci énoncé, le système visuel humain devient-il obsolète ? ■

Abstract

After 50 years of existence, computer vision applications are present in all business areas, from consumer market to industrial plants, including entertainment, sales, transportation and defense. As a science domain, computer vision borders several domains, as image processing, and encompasses lot of sub-domains like stereoscopy, or robot vision. Following the never-ending increase in computing power available, and the evolution, more recently, of imaging sensors, computer vision offers theoretical models with more details, and algorithms of larger complexity. From now on, the general use of mobile devices allows computer vision tools even in our pockets. Once all that said, do we still need human vision? ■