Comment donner un sens à l'image numérique ? par Jean Ponce. De la description à la recherche d'images : entretien avec Daniel Teruggi, propos recueillis par Dominique Chouchan

Jean Ponce

les cahiers de l'INRIA

Rendez-vous avec les sciences et technologies de l'information et de la communication

VISION ARTIFICIELLE

Comment donner un sens à l'image numérique ?

Pourra-t-on un jour doter les ordinateurs de capacités visuelles comparables à celles du cerveau humain ? Autrement dit, une machine saura-t-elle interpréter ce qu'elle « voit » ? Cette question est au cœur de la vision artificielle.

Un être humain est capable d'identifier les individus et objets qui se trouvent dans son champ visuel. Réussir à ce qu'un ordinateur puisse faire de même représente un défi immense : il s'agit de parvenir à ce que la machine interprète chaque élément d'une image, fixe ou animée. Dans notre jargon informatique, cela s'appelle l'analyse de scènes. Au-delà de leur aspect fondamental, les recherches en la matière ont des retombées dans divers secteurs, de la robotique aux effets spéciaux, en passant par la recherche de photos dans des bases en contenant plusieurs millions.

© I. LAPTEV

Fig. 1 : Le repérage des éléments de cette image (personne, verre, bougie, *etc.*) et des actions qui s'y déroulent (l'homme qui boit), évident pour un être humain, est loin de l'être pour une machine.

L'une des difficultés majeures tient à l'énorme variabilité des images d'une même catégorie d'objets : variabilité des points de vue, de l'arrangement des sources lumineuses, des couleurs, des textures, des formes, *etc.* Par exemple, deux chaises ou tables peuvent avoir des styles ou structures très différents. Dans le cas d'individus, l'apparence varie selon l'expression du visage, la posture, les vêtements, *etc.*, même si l'arrangement géométrique des yeux, du nez et de la bouche ne change guère d'une personne à l'autre, pas plus que le motif d'ombres et de lumières du visage. En outre, un objet peut être partiellement caché par un obstacle et l'avant d'un solide opaque en masque systématiquement la partie arrière. Enfin, chaque objet ou individu est souvent noyé dans un « fouillis » constitué par l'arrière-plan de la scène (fig. 1).

Nous sommes ainsi confrontés à un double problème : celui de la modélisation (comment représenter une image ?) et celui de la reconnaissance (quelles méthodes de classification et d'apprentissage ?). Pour la reconnaissance de visages, la première idée qui vient à l'esprit s'appuie sur une modélisation et une technique de classification très simples. Une image est représentée par la matrice des niveaux de gris (ou des couleurs) de ses pixels*. Quant à la classification, elle consiste à mesurer la distance entre la matrice associée à une image-test et celles associées à des visages de référence stockés en mémoire. L'image-test est alors étiquetée par le nom de la plus proche, au

sens de la distance choisie, parmi les images de référence. Mais si cette démarche donne des résultats acceptables dans des conditions favorables (point de vue et illumination fixes), il n'en va pas de même dans les situations de grande variabilité évoquées plus haut.

Les premiers travaux sur l'interprétation des scènes visuelles ont en fait été menés au début des années 1960, par exemple par l'Américain Lawrence G. Roberts au *Massachusetts Institute of Technology* (MIT). L'accent a longtemps été mis sur une approche structurelle (syntaxique) fondée sur une description globale des images en termes de formes primitives (cylindres ou cônes par exemple). Mais il est extrêmement difficile d'extraire ces formes d'images réelles et, surtout, l'analyse de scènes diffère radicalement de celle du langage naturel : même si l'interprétation d'un texte ne se réduit pas à une simple analyse syntaxique, les lettres, mots, *etc.*, n'en forment pas moins un vocabulaire bien établi. Ce n'est pas le cas dans le monde visuel où les primitives (les « mots ») sont à découvrir.

Un tournant s'est opéré à la fin des années 1990 sous l'impulsion, entre autres, de Cordelia Schmid (Inria, France) et de David Lowe (université de Colombie britannique, Canada)[1,2]. La communauté de vision artificielle se tourne alors vers des primitives locales plus proches de l'image brute et beaucoup plus faciles à extraire. L'idée est de déterminer autour de chaque pixel (ou d'un ensemble parcimonieux de points d'intérêt*) une ellipse dont la forme, la taille et l'orientation s'adaptent automatiquement au point de vue d'une caméra. D'où la possibilité de caractériser l'apparence de chaque ellipse par un vecteur (SIFT*) dont les composantes mesurent la distribution des niveaux de gris dans la partie de l'image correspondante.

Cette méthodologie permet d'obtenir une description locale des images à la fois discriminative (efficace pour l'identification d'objets) et relativement insensible aux variations de point de vue et d'illumination. En outre, des méthodes classiques (dites de discrétisation) permettent ensuite de passer des vecteurs SIFT à un petit nombre de vecteurs représentatifs, assimilables en quelque sorte à des « mots visuels ». Nous sommes par là même en mesure d'adapter à la vision des méthodes de classification héritées du traitement du langage naturel[3].

Un document textuel peut en effet lui aussi être décrit par un vecteur, l'histogramme représentant la fréquence des mots qui le composent, hors toute considération d'ordre et de syntaxe : on parle de « sac de mots ». Typiquement, un système tel que Google recherche les documents (les pages web) correspondant à une requête textuelle en combinant des méthodes d'indexation efficaces avec la recherche des plus proches voisins de cette requête dans l'espace des sacs de mots. Transposée dans le visuel, cette méthode permet la recherche d'objets dans une scène (fig. 2).

La stratégie des sacs de mots offre également le moyen de discriminer dans un échantillon de textes, par exemple des courriels, ceux qui sont acceptables et ceux à rejeter (comme les spams). Les premiers forment un ensemble de prototypes dits « positifs » dans l'espace vectoriel des histogrammes, les seconds un ensemble de prototypes « négatifs ». Tout nouveau document est ainsi classé comme acceptable ou non par comparaison à ces prototypes, sur la base de sa distance aux prototypes en question (méthode des plus proches voisins). En vision artificielle, il s'agit plutôt de décider si une image contient un certain type d'objet ou pas. Il faut pour cela affiner la méthode de classification, ce qu'autorisent certains résultats issus du domaine de l'apprentissage statistique.

© J. SIVIC ET A. ZISSERMAN

Fig. 2 : En haut, la même camionnette apparaît dans deux plans différents d'un film mais avec une localisation et un éclairage différents ainsi que la présence d'une femme sur une image de droite. Au centre, des primitives visuelles (ellipses) sont extraites de ces images. En bas, on voit le résultat de l'appariement automatique de ces primitives pour la recherche de la camionnette (désignée par le rectangle jaune) dans l'image de droite : l'ordinateur est capable d'identifier la camionnette à droite en dépit des différences précitées.

* Le pixel est l'unité de surface d'une image numérique.

* Par point d'intérêt, on désigne une zone de l'image particulièrement saillante (frontière d'un objet, changement de texture, de couleur, *etc.*).

* SIFT est l'acronyme de *Scale-Invariant Feature Transform*, ce qui en français désigne l'opération consistant à extraire de l'image des caractéristiques invariantes par changement d'échelle.

© J. MAIRAL

Fig. 3 : Les deux images de gauche contiennent entre autres une bicyclette. Grâce à la technique de classification que nous avons mis au point, l'ordinateur est capable de classer les pixels qu'il suppose appartenir à une bicyclette avec une certaine confiance (du rouge au jaune selon que le degré de confiance est élevé ou faible).

Des algorithmes plus puissants que la méthode des plus proches voisins ont en effet été mis au point ces dernières années. Tel est le cas des algorithmes baptisés « machines à vecteurs de support » (MVS), fondés sur des développements théoriques des années 1990 dus à Vladimir Vapnik, aujourd'hui chercheur dans les laboratoires NEC (Princeton). Le principe est de construire une surface séparant les prototypes positifs des négatifs dans l'espace vectoriel des histogrammes (ici l'espace des vecteurs SIFT). Tout nouveau document (toute nouvelle image) reçoit alors l'étiquette associée au côté de la surface où se trouve l'histogramme des « mots » correspondants. Bien qu'elle ignore toute information spatiale, cette technique donne d'excellents résultats en classification d'images[4].

Mais le Néerlandais Jan J. Kœnderink, l'un des pères de la vision artificielle, va plus loin. A ces représentations (modèles), qu'il qualifie de désordonnées, il substitue des modèles seulement « localement » désordonnés, de manière à préserver l'arrangement spatial global de chaque image, perdu par la méthode précédente. La version la plus simple de cette approche consiste à remplacer l'histogramme unique des sacs de mots par un ensemble d'histogrammes, chacun étant associé aux mailles d'un quadrillage *a priori* superposé à l'image. Ces histogrammes sont ensuite mis bout à bout pour former un seul vecteur, lequel est donné en entrée d'une machine à vecteurs de support. Cette approche améliore de manière significative les résultats de classification et de détection[5,6].

Autre piste : passer du « sac de mots » à une décomposition en « lettres » visuelles. Cette piste est explorée à l'Inria dans notre équipe[7]. L'image est découpée en petits rectangles de quelques dizaines de pixels. Les « lettres » sont symbolisées par les colonnes d'une matrice (une sorte d'« alphabet ») optimisée pour la tâche de classification. Le vecteur de niveaux de gris correspondant est alors représenté comme une combinaison linéaire d'une fraction réduite de ces lettres. Une fois l'alphabet appris, l'ordinateur est en mesure de discriminer les pixels appartenant à une classe d'objets, par exemple des bicyclettes, des pixels appartenant à une seconde classe, par exemple le fond de l'image. Cette méthode est étonnamment performante : elle permet de capturer le concept visuel d'un objet à partir d'informations purement locales (fig. 3).

Il va de soi que l'interprétation d'images animées (vidéos) accentue encore les difficultés, très brièvement évoquées ici. A ce jour, il reste un énorme travail avant de pouvoir doter la machine de capacités visuelles comparables à celles de l'être humain : approfondir le problème de la décomposition d'une scène en ses composantes sémantiques, examiner les aspects tridimensionnels de l'analyse de scènes et mieux exploiter les modèles probabilistes issus de l'apprentissage statistique. L'évaluation des méthodes d'analyse de scènes est d'ailleurs devenue une composante importante de la recherche dans ce domaine, en particulier dans le cadre du réseau d'excellence européen Pascal sur la modélisation statistique des interfaces multimodales.

Jean Ponce, professeur à l'Ecole normale supérieure (ENS), est responsable scientifique de l'équipe Willow commune à l'Inria, à l'ENS et au CNRS. Ses recherches portent sur la vision artificielle. Il est l'auteur du livre *Computer Vision : A Modern Approach* (Prentice Hall, 2002), traduit en chinois, japonais et russe.

[1] C. Schmid et R. Mohr, Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 872-877, 1996.

[2] D. Lowe, Proc. International Conference on Computer Vision, pp. 1150-1157, 1999.

[3] J. Sivic et A. Zisserman, Proc. International Conference on Computer Vision, pp. 1470-1477, 2003.

[4] J. Zhang, M. Marszalek, S. Lazebnik et C. Schmid, The International Journal of Computer Vision, pp. 213-238, 2007.

[5] N. Dalal et B Triggs, Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 886-893, 2005.

[6] S. Lazebnik, C. Schmid et J. Ponce, Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2169-2178, 2006.

[7] J. Mairal, F. Bach, J. Ponce, G. Sapiro et A. Zisserman, Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, 2008.

D.R.

L'Institut national de l'audiovisuel (Ina) possède 1,4 million d'heures d'archives de radio-télévision. Autant dire que la recherche d'une image dans ce fonds s'apparente à celle d'une aiguille dans une botte de foin.

Entretien avec Daniel Teruggi

De la description à la recherche d'images

Comment peut-on gérer un fonds d'archives aussi volumineux que celui de l'Ina de manière à le rendre exploitable ?
Daniel Teruggi : La première chose est d'assurer sa conservation, la plus grande part étant encore sous forme analogique. Nous mettons donc en œuvre un programme de numérisation : 600 000 heures sont à ce jour numérisées et accessibles en ligne par les professionnels, dont 25 000 également accessibles par le public*. La seconde concerne la navigation dans cet océan de sons et d'images, dont l'hétérogénéité des descriptions tient entre autres à la diversité des époques de production. Ces descriptions, essentiellement textuelles, sont souvent très incomplètes. L'une de nos activités consiste à développer des outils permettant de les améliorer, en particulier dans le cadre de programmes tels que Quaero*, sur les technologies d'analyse automatique, de classification et d'utilisation de documents multimédias et multilingues ou Vitalas*, sur l'accès intelligent à des archives multimédias.

Plus précisément, sur quels types d'outils portez-vous vos efforts ?
D. T. : Nous travaillons en parallèle sur l'optimisation de technologies existantes et sur la conception de nouvelles. Un exemple de technologie déjà mise en œuvre : la transcription automatique de parole en texte. Les algorithmes actuels sont très efficaces pour l'expression orale contemporaine, mais beaucoup moins pour des documents plus anciens : ne serait-ce que dans les années 1960, le français parlé était très différent de celui d'aujourd'hui, ce qui impose d'affiner la technique si l'on veut pouvoir extraire le texte signifiant. La reconnaissance de visages constitue un autre problème, très difficile, sur lequel nous nous penchons. Nous travaillons également beaucoup sur la segmentation automatique : l'idée est de segmenter une émission (comme le journal télévisé) de manière à en faire une description structurelle et sémiologique. En matière d'approches nouvelles, nous nous intéressons particulièrement à la recherche d'images similaires au regard d'un certain nombre de critères, lesquels ne correspondent pas nécessairement au sens général de l'image : par exemple, la recherche de toutes les images où figure un coucher de soleil, et ce quel que soit le type de scène ou de source (documentaire, journal sportif, fiction…).

D'où le projet que vous mettez en place avec l'Inria et Microsoft Research* ?
D. T. : La question sur laquelle nous réfléchissons porte sur l'identification de situations : comment, dans une image, déterminer s'il y a une foule, une personne avec un chapeau, tel objet ou tel animal, *etc.* ? Il faut pour cela concevoir des modèles simplifiés de cette image qui puissent être appliqués à grande échelle et permettre ensuite d'en retrouver d'autres avec des situations similaires. Exemple : un sociologue réalise une étude sur la cigarette dans les images publicitaires présentées à la télévision. Nos archives contiennent 200 000 publicités. L'idée est de donner au chercheur les moyens de parcourir ce fonds et d'en extraire de manière automatique les images avec fumeur (s), sur la base d'une scène de référence. Il s'agit en quelque sorte d'un moteur de recherche doté d'une certaine « intelligence ». Les applications d'une telle technologie concernent une très grande diversité d'usages.

Propos recueillis par Dominique Chouchan

Daniel Teruggi est directeur de la recherche à l'Ina. Il poursuit parallèlement son activité de compositeur au sein du Groupe de recherches musicales de l'Ina, qu'il avait rejoint en 1983 après des études de musique en Argentine puis en France.

* Ces **images** sont accessibles sur le site www.ina.fr (onglet en bas à droite de la page d'accueil)

* Les programmes **Quaero** et **Vitalas** associent des partenaires publics et privés, français et européens.

* Ce projet du **Centre commun de recherche Inria-Microsoft Research** (Orsay) associe l'Ina et trois équipes de l'Inria.

Réalisation rédactionnelle : Dominique Chouchan - Réalisation graphique : E-Graphics