Catégorisation de Scènes Naturelles : l'Homme vs la Machine

Guyader Nathalie[1], Alan Chauvin[2], Le Borgne Hervé[1]

[1]LIS, 46 Av. Félix-Viallet, 38031 Grenoble Cedex, France
[2]LPE, BP 47, 38040 Grenoble Cedex 9, France
{nguyader,hleborgn}@lis.inpg.fr, alan.chauvin@lis.inpg.fr

Résumé

Dans cet article nous développons une méthode, pour la catégorisation sémantique des images de scènes naturelles, qui repose sur la mise en relation d'une mesure comportementale et d'un modèle inspiré du fonctionnement du système visuel humain. Dans une première partie, un groupe de sujets doit associer des images représentant des scènes naturelles. Les distances évaluées à partir des associations nous permettent de calculer des matrices de similitudes entre toutes les images. Une méthode d'analyse de données en grande dimension : l'Analyse en Composante Curviligne (ACC), permet d'extraire de ces matrices les catégories sémantiques présentes dans la base d'images étudiée. Dans une deuxième partie, nous extrayons grâce à un modèle des cellules complexes du cortex visuel primaire un vecteur d'attributs des images, qui après analyse par ACC sera comparé aux mesures chez l'humain. Cette comparaison permet de raffiner notre modèle.

I. Introduction

Dans cet article, nous proposons une approche de la catégorisation d'images à partir d'un modèle du système visuel et de données issues de la psychophysique. Le but est d'étudier et de comprendre comment l'être humain identifie son environnement visuel et d'en tirer une méthode exploitable par une machine. Cette prise en compte de la perception visuelle chez l'homme est une chose nouvelle dans le domaine des technologies de l'information et fait appel à différentes disciplines que sont la psychologie, les neurosciences et les sciences de l'ingénieur. On s'est principalement attaché ici à la perception et la catégorisation des « scènes naturelles ». Les raisons de ce choix sont d'une part, que les scènes constituent notre environnement quotidien et d'autre part que leur perception concentre nombre de problèmes que doit résoudre un système de reconnaissance d'images. Ce type de stimulus est relativement peu utilisé dans la littérature (que se soit en neurosciences ou psychologie cognitive) ; en effet, les stimuli utilisés chez l'homme comme l'animal sont généralement des objets appauvris, décontextualisés et souvent isolés. La compréhension des mécanismes de l'analyse des scènes visuelles débouche également sur d'importants enjeux socio-économiques. Ainsi le développement croissant de nombreuses bases d'images distribuées à travers Internet, les musées, les agences de presse ou de publicité et la demande croissante de moteurs de recherche efficace requiert des outils automatiques de catégorisation, de labellisation et d'indexation d'images. Automatiser ces tâches est actuellement un défi technologique qui mérite d'être relevé.

Dans une première partie, la catégorisation est réalisée par l'intermédiaire d'une expérience en psychophysique. Dans une seconde partie, les mêmes images sont classées à partir des descripteurs modélisant les cellules du système visuel. Les résultats des deux méthodes sont ensuite comparés afin d'apporter des améliorations au modèle.

II. Expérience de catégorisation d'images

La base d'images étudiée a été choisie afin de couvrir la majorité des environnements que l'on rencontre dans la nature. Nous nous sommes inspirés des catégories sémantiques mises en évidence par l'expérience de « Computer Scaling » de Rogowitz et al. [1] ; nous avons pris 105 images recouvrant une dizaine de catégories différentes contenant: des animaux, des personnages, des scènes d'intérieur, des paysages comme des plages ou des montagnes, des scènes urbaines. Etant pour le moment, uniquement focalisés sur l'information de structure, nous avons exclu la composante couleur pour l'association d'images. Le but de l'expérience est d'obtenir une matrice de similitude entre toutes les images de la base.
Lors de chaque essai, une image de référence est présentée sur un écran à coté de huit autres. Le choix de l'image de référence et des images cibles est contrebalancé entre les sujets afin de s'assurer que chaque image de référence est opposée à l'ensemble de la base. Le sujet choisit l'image qui est la plus ressemblante à celle de référence. Les deux images sont ensuite affichées côte à côte afin que le sujet juge de leur degré de ressemblance. Ce jugement de similitude n'existe pas dans l'expérience originale de « Computer Scaling ». L'introduire permet d'obtenir à la fois des séparations franches entre catégories mais aussi des continuums entre certaines catégories, par exemple des personnes en bord de mer seront entre la classe des personnes et celle des scènes ouvertes. Une Analyse en Composantes Curvilignes (ACC) [2] permet d'obtenir en 2 dimensions une « organisation perceptive » de ces images. Nous remarquons très clairement une organisation en « clusters » sémantiques. (Cf. figure 1).

Figure 1: Projection en 2D de la matrice des similitudes « humaines ».

Le modèle présenté ci-après n'étant basé que sur les orientations présentes dans les images, il ne peut extraire des catégories comme les personnages ou les animaux. C'est pourquoi nous ne regarderons ici que les résultats de classification obtenus sur les catégories : scènes ouvertes (plage, désert, champ), montagnes, forêts, villes et intérieurs. On travaille donc avec 62 images.

III. Modèle de catégorisation d'images

Des travaux préliminaires du LIS ont montré que l'analyse de scènes basée sur le spectre d'énergie global de l'image était suffisante pour extraire certaines catégories perceptives [3]. De plus, les performances de notre système visuel sont remarquables pour la catégorisation d'images; c'est pourquoi nous nous sommes inspirés des principes qui le gouvernent. Nous avons cherché à reproduire ce que l'on connaît de la chaîne des traitements de notre système visuel; les traitements portant in fine sur le spectre d'énergie.

Notre travail a consisté en la modélisation d'une partie des fonctions rétiniennes et corticales. On modélise les traitements effectués par la rétine et par les cellules simples et complexes du cortex visuel. Dans le cortex visuel de l'homme, les cellules simples fournissent un filtrage des images sur le modèle d'une série d'ondelettes. Les cellules complexes élaborent les énergies locales de chaque image; leurs sorties peuvent être considérées comme les composantes d'un vecteur d'attributs de l'image. Le système visuel élabore un traitement de ce vecteur, capable de mener à la reconnaissance de scènes [4].
A l'image du système visuel, après un pré-filtrage rétinien, nous appliquons un filtre « cortical » qui code l'image par son énergie dans sept bandes de fréquences spatiales et sept orientations. Les sorties des filtres à différentes fréquences et différentes orientations sont considérées comme les composantes d'un vecteur d'attributs à 49 dimensions. Ce vecteur caractéristique subit un premier traitement non linéaire de normalisation, dont l'effet est une relative indépendance par rapport au flou (normalisation dans chaque bande de fréquence) et par rapport aux niveaux locaux d'énergie dans l'image. Par une ACC qui trouve la variété intrinsèque des données dans l'espace à 49 dimensions, notre base d'images est projetée dans un espace à 2-dimension. Comme lors des projections réalisées à partir de la matrice de similitude « humaine », les images s'organisent en nuages de points représentatifs de certaines catégories sémantiques. Certaines images se retrouvent mal classées : par exemple des images d'arbres possédant des orientations verticales bien marquées se retrouvent classées parmi les villes. C'est pourquoi nous essayons de plaquer l'organisation obtenue ici sur celle obtenue en figure 1.

IV. Comparaison des deux méthodes

Afin d'obtenir une représentation 2D de l'organisation des images obtenue par notre modèle la plus semblable à celle obtenue par l'expérimentation, nous allons chercher le vecteur $\Omega = (\omega_1, \ldots, \omega_{49})$ de poids associés à chaque filtre qui minimisera la fonction de coût suivante :

$$C = \sum_{i,j} \left(D_{ij}^2 - E_{ij}^2\right)^2 = \sum_{i,j} \left(D_{ij}^2 - \sum_{k=1}^{49} w_k \left(data(i,k) - data(j,k)\right)^2 \right)^2$$

où D est la matrice des distances euclidiennes calculées sur le représentation 2D des données perceptives et E celle des distances euclidiennes sur la représentation 2D obtenue par notre modèle de filtre de Gabor. *Data(i, :)* est la description de l'image i par les filtres de Gabor.
L'optimisation se fait par une méthode de descente de gradient. On obtient alors 49 coefficients qui serviront de pondération aux réponses des différents filtres de Gabor. On projette alors par ACC, les images caractérisées d'une part par les Gabor simples et d'autre part par les Gabor pondérés. On remarque alors une séparation meilleure des images en classes sémantiques.

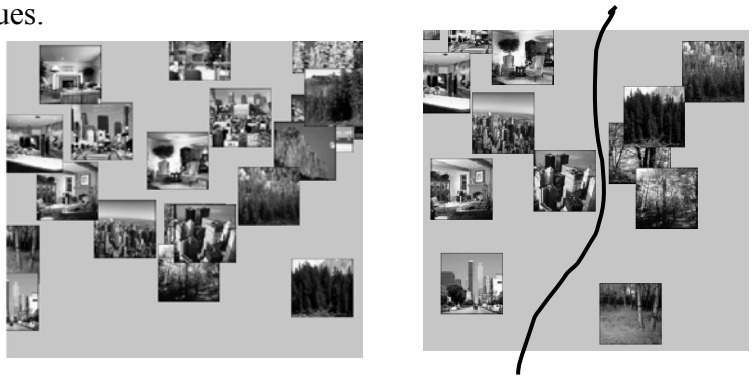

Figure 2: Organisation de images par projection des attributs extraits avec des Gabor simples et avec les Gabor pondérés.

Il est intéressant de visualisation le vecteur des différents poids ayant permis l'optimisation décrite ci-dessus. On remarque qu'il est important de pondérer de manière plus forte les filtres dans les orientations obliques (Cf. Figure3).

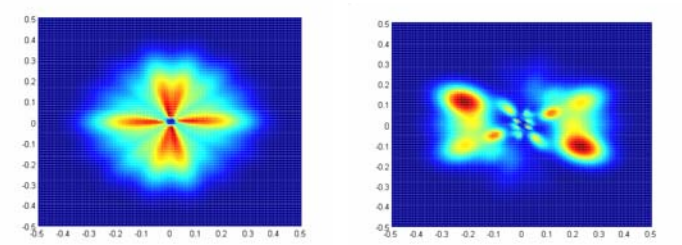

Figure 3 : Réponses moyennes des Gabor sur les 62 images considérées (à gauche), visualisation des pondérations associées à chaque filtre après optimisation (à droite).

Pour vérifier la validité de notre modèle, on le teste sur un classifier simple : on regarde les différentes classes (ici 5). On calcule alors le vecteur moyen de chaque classe avant et après optimisation. On regarde ensuite la distance euclidienne qui sépare chaque image de la base à chacun des vecteurs caractéristiques des différentes classes. On associe à l'image la classe dont le vecteur caractéristique est le plus proche. D'où les matrices de classification :

	Sans optimisation	Avec optimisation
Scènes ouvertes	60%	75%
Villes	66%	88%
Montagnes	50%	60%
Intérieurs	70%	92%

V. Conclusion

Notre modèle a été modifié afin de plaquer l'organisation donnée par le système d'analyse sur celle faite par le sujet humain. La pondération de nos attributs améliore significativement nos résultats de classification. Nous montrons ainsi qu'il est indispensable de prendre en compte, dans des algorithmes de catégorisation d'images, la perception du sujet humain. Les résultats sont concluants et peuvent aboutir à d'importants changements dans les méthodes d'analyse de scènes.

V. Bibliographie

[1] B. Rogowitz, T. Frese, J. Smith, C.A. Bouman, and E. Kalin, "Perceptual image similarity experiments", Human Vision and Electronic Imaging III, *Proc. of SPIE*, vol. 3299, San Jose, CA, January 26-29, 1998.

[2] Demartines P. & Hérault J. *Curvilinear Component Analysis : a Self-Organising Neural Network for Non-Linear Mapping of Data Sets*, IEEE Trans. On Neural Networks, 8, 1, 148-154, 1997.

[3] Guérin-Dugué A., Oliva A. *Classification of Scene Photographs from Local Orientations Features*, Pattern Recognition Letters, 21, pp 1135-1140, 2000.

[4] Guyader N. & Herault J. *Représentation espace-fréquence pour la catégorisation d'images*. Colloque GRETSI 01, Toulouse.

[5] Chauvin A., Herault J. & Marendaz C. *Natural scene perception: visual attractors and image processing*. NCPW7 2001, Brighton, UK.