Représentation espace-fréquence pour la catégorisation d'images

Nathalie GUYADER, Jeanny HERAULT

Laboratoire des Images et des Signaux

INPG, 46 Av. Félix Viallet, 38031, GRENOBLE Cedex 01

{guyader, herault}@lis.inpg.fr

Résumé – La méthode présentée vise à reproduire la chaîne des traitements de notre système visuel. Après un préfiltrage rétinien (égalisation adaptative des contrastes locaux, blanchiment spectral), nous appliquons un filtre "cortical" qui code l'image par son énergie dans sept bandes de fréquences spatiales et sept orientations. Dans un mode "global", chaque image est codée par un vecteur à 49 dimensions et une classification supervisée en catégories sémantiques est effectuée. Dans un mode "local", la même procédure est appliquée à 16 imagettes à l'intérieur de chaque image, ce qui permet de définir des pourcentages d'appartenance à plusieurs catégories. Le premier, plus efficace en temps de calcul, voit ses ambiguïtés levées par le second qui s'avère plus précis, mais plus calculatoire.

Abstract *– This work aims at modelling the processes in our visual system. After a retinal preprocessing (adaptive equalization of local contrasts, spectral whitenning), we apply a "cortical" filter which codes an image by means of its energy within seven spatial frequency bands and seven orientations. In a "Global" mode, each image is coded by a 49-dimension vector and a supervised classification into semantic categories is performed. In a "Local" mode, the same procedure applies to 16 patches in each image, allowing to define belonging percentages to various categories. The first mode is computationally more efficient, but leads to ambiguities. These ambiguities can be resolved by the second mode, which is more accurate but more computationally demanding.*

1. Introduction

Nous considérons une image de scène $i(x,y)=i(\mathbf{x})$ comme une collection d'objets ou de surfaces réfléchissant la lumière, juxtaposés ou se recouvrant partiellement. Ces éléments peuvent varier en éclairement selon $E(\mathbf{x})$, en position dans l'image, en taille (distance), en attitude (orientation ou perspective). Ils sont caractérisés par leur coefficient de réflexion $\rho(\mathbf{x})$. L'image (achromatique) produite est en général de la forme $i(\mathbf{x})=\rho(\mathbf{x})\cdot E(\mathbf{x})$, bien qu'il puisse exister des sources ponctuelles.

1.1 Propriétés et statistique des scènes

Pour un même type de scène, il existe un certain nombre de facteurs de variabilité et de constantes qu'il nous faut identifier en vue de la catégorisation.

L'éclairement dans une scène est une fonction soit *globale* liée à la direction de la source et à la lumière diffuse ambiante, soit *locale* (ombres portées, ombres propres) liée à la disposition ou à la posture des objets. Excepté le cas des ombres propres, cette fonction varie lentement avec la position dans l'image.

On pourra aussi tenir compte du fait qu'il existe un bruit photonique lié à l'intensité et un flou isotrope qui peut varier avec la position (éloignement d'un objet). En tout état de cause, c'est le *coefficient de réflexion* qui est la principale constante à évaluer, mais l'illumination de la scène est aussi un paramètre important.

Si ***la position*** des objets peut varier dans une certaine limite (le ciel est généralement en haut), leur *nature* est une "constante" typique d'une scène.

Le spectre d'énergie moyen des images est une fonction en 1/f. Dans une image particulière, il est lié à la nature et à la position relative des objets, mais indépendant des positions absolues.

L'inversion droite-gauche conserve la nature de la scène mais peut changer sa description selon les descripteurs choisis.

La taille, l'orientation, ou l'attitude des objets peuvent varier, il faut identifier ces grandeurs ou s'en affranchir dans la description des objets pour pouvoir catégoriser la scène, par exemple identifier les zones de perspective.

La composition d'une image est variable : par exemple, comment catégoriser une image de plage avec un bosquet et des immeubles? Cela peut être vu, selon la taille de ces composantes, comme une forêt ou comme une ville.

Pour catégoriser une image, il nous faut résoudre toutes ces variabilités et identifier les constantes. Les performances de notre système visuel sont remarquables dans ce type de tâche et notre but est de nous inspirer des principes qui le gouvernent.

1.2 La perception visuelle

L'architecture de notre système visuel est un modèle intéressant car les fonctions que nous connaissons résolvent une partie des problèmes de variabilité cités plus haut, celles que nous ne connaissons pas (encore) et que nous étudions dans le cadre des sciences cognitives devraient résoudre le reste.

La rétine est le siège d'un filtrage spatio-temporel passe-haut après un processus de compression adaptative. Il en résulte une *égalisation des contrastes* dans l'image (d'où une relative insensibilité aux variations d'éclairement) et un blanchiment spectral qui compense la statistique en 1/f des images (Beaudot, 1996; Hérault, 2001).

Le cortex visuel primaire (aire V1) est constitué de filtres passe-bande orientés (de type ondelette de Gabor). Le module de l'énergie locale donne la *signature spectrale* des objets, *indépendamment de leur position*. Ces filtres sont en

interactions locales mutuelles (détection d'attributs, Purpura & al., 1994) ou régionales (influence du contexte), souvent sous le contrôle descendant de processus attentionnels (Li, 1999; Chauvin, 2001).

Les aires "supérieures", sont divisées en deux voies qui traitent soit les formes et la couleur (What), soit les positions et les mouvements (Where). La première répond sélectivement à des formes complexes ou à des types d'objets.

Les performances globales du système visuel sont impressionantes : nous sommes capables de *catégoriser* une image en moins de 100 ms, la présence d'un type d'objet est détectée en 150 ms (Van Rullen & Thorpe, 1998), cependant, il faut plus de temps pour reconnaître -identifier- une scène particulière (mécanismes d'exploration, processus "top-down"...). Nous supposons que cette rapidité de catégorisation ne peut se faire qu'à partir d'indices de bas niveau, dont la statistique est évaluée sur toute l'image (Guérin & Oliva, 2000).

2. Outils et méthode

2.1 Traitement rétinien

Les photorécepteurs de la rétine réalisent une compression de niveau selon l'équation : $r(\mathbf{x})=\frac{i(\mathbf{x})}{i(\mathbf{x})+i_0(\mathbf{x})}$. Ils s'adaptent à l'intensité moyenne de l'image dans leur voisinage car le terme $i_0(\mathbf{x})$ est en fait une moyenne locale $\bar{i}(\mathbf{x})$ par lissage de l'image d'entrée : $i_0(\mathbf{x})\approx\bar{\rho}(\mathbf{x})\,\bar{E}(\mathbf{x})$. Comme l'éclairement varie lentement, $\bar{E}(\mathbf{x})\approx E(\mathbf{x})$ et :

$r(\mathbf{x})=\frac{\rho(\mathbf{x})\,E(\mathbf{x})}{\rho(\mathbf{x})\,E(\mathbf{x})+\bar{\rho}(\mathbf{x})\,\bar{E}(\mathbf{x})}\approx\frac{\rho(\mathbf{x})}{\rho(\mathbf{x})+\bar{\rho}(\mathbf{x})}$ Il en résulte une égalisation des contrastes entre les zones d'ombre et celles de fort éclairement. Les circuits neuronaux de la rétine réalisent un filtrage passe-haut de l'image des photorécepteurs, ce qui constitue une sorte de blanchiment spectral à cause de la forme en 1/f du spectre moyen des images. La figure 1 donne un exemple des traitements réalisés.

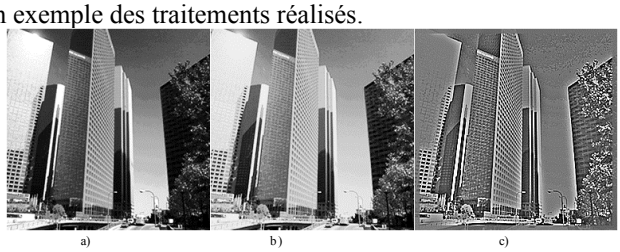

FIG. 1 : L'image originale (en a) est compressée localement dans les zones d'ombres par les photorécepteurs (en b), puis filtrée passe haut par les circuits rétiniens (en c). On notera l'égalisation des contrastes dans toutes les zones.

2.2 Modèle de spectre

Soit une image $i(\mathbf{x})=\sum_i o_i(\mathbf{x}-\mathbf{x}_i)$ composée d'une série d'objets o_i à des positions $\mathbf{x}_i$ dont les spectres 2D s'écrivent. $O_i(f)=|O_i(f)|\exp\left(j\,2\pi\,f^T x_i+j\,\varphi_i(f)\right)$ En conséquence, le module du spectre de l'image s'écrit :

$$|I(f)|^2=\sum_i|O_i(f)|^2+2\sum_{i,j>i}|O_i(f)||O_j(f)|\cos\left(2\pi\,f^T\Delta x_{ij}+\varphi_{ij}(f)\right)$$

Le premier terme représente le spectre moyen des objets, indépendamment de leur position et le second contient la structure de l'image liée aux positions relatives des objets $\Delta\mathbf{x}_{ij}$, les termes de différences de phase $\varphi_{ij}(\mathbf{f})$ s'annulant rapidement avec le module de la fréquence. La figure 2 donne en a: le spectre global de la configuration d'objets A, B, C (en encart), en b: le spectre moyen des objets et en c: le spectre de la structure représentée par les centres de gravité de objets. On note que les hautes fréquences sont typiques du spectre moyen des objets et que les basses fréquences les sont de la structure de l'image.

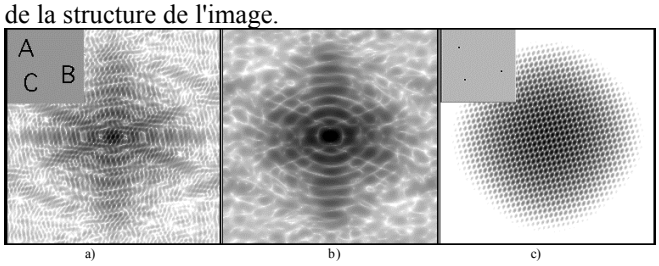

FIG. 2 : Spectre d'une image (a), spectre moyen de ses objets (b) et spectre de sa structure (c).

Typique de la structure d'une image, le module du spectre constitue donc un descripteur de choix en vue de la catégorisation.

2.3 Traitement "cortical"

Arrivée dans l'aire V1 du cortex visuel, l'image rétinienne est décomposée en un certain nombre de primitives par le filtrage des neurones du cortex ; lesquels sont sensibles à diverses bandes de fréquences spatiales, temporelles, et à certaines orientations des stimuli. Les cellules du cortex sont de deux types : les cellules simples modélisables par des filtres de Gabor en phase et en quadrature et les cellules complexes qui intègrent l'énergie des stimuli visuels en sortie d'un filtre (Jones & Palmer, 1987). Nous modélisons ici les cellules complexes par des ondelettes de Gabor.

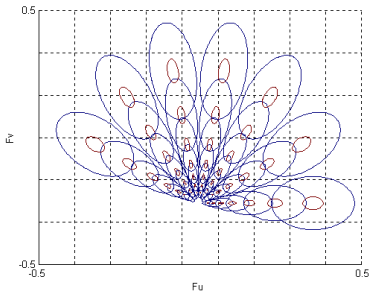

FIG. 3 : Rosace de 49 filtres de Gabor, 7 bandes de fréquences et 7 orientations avec un fort recouvrement des filtres.

Les images seront ici décomposées par 49 ondelettes de Gabor représentées en figure 3, soit 7 bandes de fréquences et 7 orientations différentes. En s'inspirant de la biologie des cellules du cortex visuel qui ont une largeur de bande moyenne de 1.2 octave (De Vallois, 1988), la largeur de la bande radiale relative des filtres de Gabor est fixée à 1 octave.

La largeur de la bande transversale relative est de 180°/7. On désire en effet, avoir toutes les caractéristiques de l'image et donc couvrir au mieux le domaine spectral. Ayant choisi 7 orientations, il est logique de prendre une largeur de bande transversale de 180°/7.

On parle d'échantillonnage du spectre en log-polaire car les fréquences centrales des filtres sont en progression géométrique de raison 1.5 : $f_k = 1.5^k f_0$ (à l'image des cellules simples du cortex visuel), et les filtres sont fonctions de l'orientation.

Chaque image est caractérisée par une matrice 7 x 7 : cette matrice est l'index choisi pour la catégorisation. Chaque composante de la matrice correspond à un échantillon du module du spectre selon l'orientation et la fréquence (énergie en sortie d'un filtre).

Dans cette nouvelle représentation, zoom et rotation deviennent des translations plus facilement exploitables. Ainsi, en considérant les spectres log-polaire comme des images de 7 × 7 pixels, l'intercorrélation entre une image originale et une image déformée (image originale tournée et zoomée) donne un maximum décalé du centre du facteur de zoom et de rotation.

De plus, nous tenons compte de l'invariance des différentes classes par rapport à l'inversion droite-gauche d'une image : équivalence d'une transformation globale $\theta \rightarrow \pi - \theta$.

2.4 Implémentation

2.4.1 Extraction des caractéristiques

Nous testons deux types de descripteurs : les descripteurs globaux (image en entier), et les descripteurs locaux (les images sont divisées en 16 imagettes).

Les images que nous souhaitons catégoriser subissent tout d'abord le prétraitement rétinien que nous venons de décrire. Ensuite, nous appliquons une fenêtre d'apodisation de Hanning sur l'image pour éviter ces effets de bords et nous calculons les énergies dans les différentes bandes de fréquences spatiales et orientations. Une image est donc codée comme un vecteur à 49 dimensions.

Plusieurs types de normalisation des vecteurs ou matrices caractéristiques ont été testés.

La normalisation qui procure les meilleurs résultats lors de la catégorisation des images est la normalisation par bande de fréquence. Cette normalisation permet de comparer par la suite des images floues avec des images nettes. En effet, le flou est une fonction isotrope de la fréquence G(f) et la normalisation par bande de fréquence supprime ce terme.

$$\textit{si flou}: \quad E(f_i,\theta_j) = \frac{E(f_i,\theta_j)G(f)}{\sum_j E(f_i,\theta_j)G(f)} = \frac{E(f_i,\theta_j)}{\sum_j E(f_i,\theta_j)}$$

2.4.2 Validation des descripteurs par catégorisation

Notre système de catégorisation est obtenu par un apprentissage supervisé. La catégorisation se fait sur des mesures de dissimilarités.

Nous travaillons sur une base de 250 images de scènes en niveaux de gris. Les images appartiennent à cinq catégories sémantiques différentes que sont les plages, les villes, les forêts, les montagnes et les intérieurs. Chaque classe possède des caractéristiques spectrales propres.

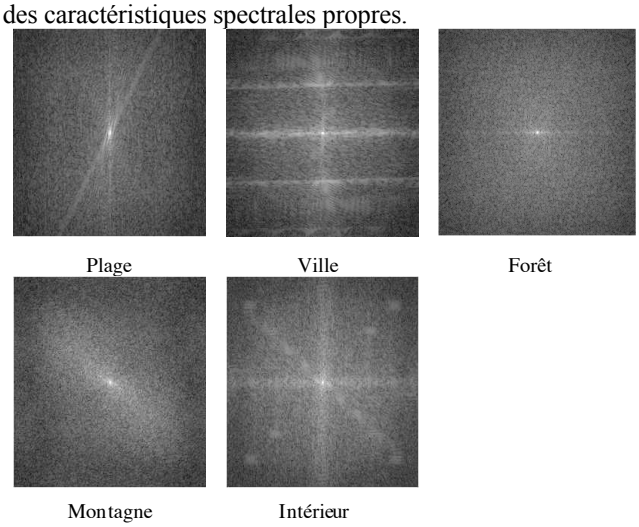

FIG. 4 : Exemple du module des spectres d'une image des différentes catégories sémantiques.

Chaque classe est caractérisée par :

1 - son centre ou vecteur caractéristique moyen ,

2 - ses directions principales (Analyse en Composantes Principales par classe avec conservation de 90% de la variance).

La décision d'appartenance d'un vecteur test (index d'une image) à une classe se fait soit par comparaison de sa distance euclidienne ou de Mahalanobis aux différents centres des classes, soit par sa distance euclidienne aux différents plans principaux.

La distance de Minkowski (norme L^p) est également testée. Dans ce cas, on calcule la distance de Minkowski entre le vecteur test et tous les autres vecteurs de la base. La distance qui le sépare du nuage représentatif d'une classe est la distance moyenne avec les points de cette classe. On compare ensuite les distances du vecteur test à chacune des classes.

3. Spectres globaux

On donne les résultats ci-dessous pour les images ayant servi de base d'apprentissage.

Il faut ici noter l'importance du prétraitement rétinien, sans lequel les résultats seraient largement moins concluants.

Les tableaux 1 et 2 donnent les pourcentages d'images (sur 50 images par classe) bien classées.

La distance qui procure les meilleurs résultats est celle aux plans principaux des différentes classes (cette distance tient compte de la distance qui sépare le vecteur test au centre de la classe ainsi que de la direction de ce vecteur par rapport au nuage de points représentatifs des catégories).

Certains pourcentages sont faibles mais :

1 - on sait que les classes ne sont pas franchement séparées (Hérault, 1997) ; une image peut contenir des informations de plusieurs catégories.

2 – notre index n'est pas pour autant mauvais, en effet si l'on regarde la classification artificiel versus naturel (soit plage, forêt et montagne contre ville et intérieur) les pourcentages sont améliorés :

TAB.1 : Catégorisation par comparaison des différentes distances (5 classes sémantiques).

Comparaison des :	Pourcentage de catégorisation correct :
- distance euclidienne aux centres des classes (D.E)	P : 60% F : 50% M : 82% V : 41% I : 25%
- distance de Mahalanobis aux centres classes (D.M)	P : 80% F : 24% M : 28% V : 62% I : 26%
- distance euclidienne aux plans principaux (D.P.P)	P : 48% F : 74% M : 84% V : 80% I : 70%
- distance de Minkowski (D.Mink) p=6	P : 46% F : 48% M : 76% V : 58% I : 68%

TAB.2 : Pourcentage de bonne catégorisation naturel/artificiel

- D.E	90%
- D.M	86%
- D.P.P	93%
- D.Mink	89.2%

Afin d'améliorer la catégorisation des images nous nous intéressons aux spectres locaux.

4. Spectres locaux

Les images sont maintenant divisées en 16 imagettes, chacune subissant les traitements décrits ci-dessus.

F	F	F	F	?	?	?	?
F	F	F	V	M	M	P	P
P	P	P	P	P	P	P	P
P	P	P	P	?	?	P	P

FIG. 5 : Exemple de catégorisation de 2 images appartenant à la classe des plages par une méthode locale. (F : forêt, P : plage, V : ville, M : montagne, ? : non significatif).

Chaque imagette est associée, avec les mêmes méthodes que précédemment, à une classe. Pour catégoriser l'image on donne un poids plus fort aux imagettes centrales et lors d'ambiguïtés (c'est-à-dire si 2 patchs appartiennent à une catégorie et 2 autres à une autre catégorie) on regarde les classes d'appartenance des imagettes autour. La première image présentée à la figure 5 est une image catégorisée sur le globale comme une forêt ; le fait d'appliquer une méthode locale nous donne la réelle composition de la scène (moitié plage, moitié arbre et un artefact ville). La deuxième image est elle catégorisée « montagne » par la méthode globale, la méthode locale la place de manière privilégiée dans la classe des « plages » (voir les 4 imagettes centrales).

La méthode des spectres locaux se rapproche de notre mode de perception : l'appartenance à une catégorie n'est pas une certitude, une image peut contenir des zones relatives à différentes catégories et la décision d'appartenance se fera soit selon un vote majoritaire (en l'absence d'une idée-guide), soit en fonction d'un contexte de recherche (selon le type d'images auquel on s'intéresse).

Dans le cas d'un vote majoritaire, la méthode augmente notablement le pourcentage de bonnes catégorisations pour les scènes ouvertes, P : 82%, C : 80%.

Dans notre approche, le type de descripteurs est volontairement limité aux spectres d'énergies, il est bien entendu nécessaire d'introduire d'autres descripteurs comme par exemple des indices de profondeur, de perspective, de couleur...

Références

BEAUDOT W. H. A. (1996). Sensory coding in the vertebrate retina: towards an adaptive control of visual sensitivity. Network: Computation in Neural Systems 7 317–323.

CHAUVIN A., MARENDAZ C., HERAULT J. (2001). Perception des scènes naturelles : Simulation d'interactions 'scène-objet' par cartes de saillance. Sciences de la Vision et Applications, sous presse.

DE VALOIS R.L., & DE VALOIS K. K. (1988). Spatial Vision. *Oxford University Press*. New York.

GUERIN-DUGUE A., OLIVA A. (2000). Classification of Scene Photographs from Local Orientations Features, Pattern Recognition Letters, vol. 21, pp. 1135-1140.

JONES J.P., PALMER L.A., (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J. Neurophysiol. 58(6) : 1233-1258.

HÉRAULT J. (2001). De la rétine biologique aux circuits neuromorphiques. in "Traité IC2, Les Systemes de Vision", J-M Jolion ed. Hermès.

HERAULT , J., OLIVA, A., & GUERIN-DUGUE, A. (1997). Scene Categorisation by Curvilinear Component Analysis of Low Frequency Spectra. *Proceedings of the 5th European Symposium on Artificial Neural Network*., Bruges, Belgium. pp. 91-96.

LI Zhaoping. (1999). Visual segmentation by contextual influences via intra-cortical interactions in the primary visual cortex. Network: Comput. Neural Syst. 10, 187–212.

OLIVA A., TORRALBA A. B., GUERIN-DUGUE A. & HERAULT J. (1999). Super-Ordinate Representation of Scenes from Power Spectrum Shapes. CIR-99, The Challenge of Image Retrieval.

PURPURA KP, VICTOR JD, KATZ E. (1994). Striate cortex extracts higher-order spatial correlations from visual textures. Proc Natl Acad Sci U S A 1994 Aug 30;91(18):8482-6.

VAN RULLEN R. & THORPE S. (1998). Ultra-Rapid Visual Categorisation of natural scenes: Order of spiking in ganglion cells as a code *Perception*, Vol. 27, pp S153.