Représentation d'images par chaînes de symboles : Application à la recherche par le contenu *

Isabelle SIMAND, Jean-Michel JOLION

LIRIS, FRE CNRS 2672 INSA
Bât. J. Verne
INSA Lyon, 69621 Villeurbanne cedex - France
{Isabelle.Simand, Jean-Michel.Jolion}@liris.cnrs.fr

Résumé – Cet article introduit une nouvelle représentation d'images basée sur des chaînes de symboles. Mêlant les notions de points d'intérêt, de contraste, et d'ordre, cette signature offre une description à la fois concise et précise de l'image. La comparaison de deux images consiste en la comparaison des symboles qui composent les chaînes signature. D'où l'élaboration d'une distance entre symboles dérivée de la distance d'édition.

Abstract – This paper introduces a new representation of images based on strings of symbols. This signature, both precise and compact, is based on notions such as interest points, contrast and order. The comparison between two images consists in the comparison of the symbols which compose the strings. It leads to the development of a distance between symbols such as the edit distance.

1 Introduction

Le domaine de l'indexation a connu un essor très rapide ces dernières années, conduisant à un ensemble de méthodologies et de techniques de plus en plus stables. Les travaux exposés ici renouent avec l'approche exploratoire initiale, en s'appuyant, entre autres, sur les outils de la reconnaissance des formes. En effet si les caractéristiques extraites dans les images sont souvent numériques, celles qui constituent notre code d'images sont symboliques. Par conséquent nous conservons toute la puissance de l'approche structurelle de la reconnaissance des formes. En particulier, la phase de comparaison pourra faire intervenir une distance entre symboles, telle que la distance de Levenstein. L'apport d'une technique de ce type sera illustré dans la suite.
Cette étude trouve sa source dans divers travaux sur le parcours visuel de l'oeil humain, qui est particulièrement attiré par les zones de fort contraste [1], ce qu'illustre d'ailleurs la loi de Naka-Rushton. En outre, dans [2], Thorpe montre l'importance de la notion d'ordre dans l'analyse des images, ce qui justifie l'usage d'une liste triée de symboles. Nos travaux ont donc pour but de traduire ces deux éléments (ordre et contraste) dans une nouvelle définition d'un code d'images pour la classification, la reconnaissance et l'indexation d'images.

2 Architecture générale

Lors de l'analyse d'une image, l'oeil effectue une succession de focus d'attention aussi appelée saccade. L'extraction de points d'intérêt dans une image, si elle se base sur un critère pertinent, constitue une bonne traduction de ce comportement, sans en être une modélisation. En outre la représentation, peu coûteuse, d'une image par un tel ensemble de points a aujourd'hui fait la preuve de ses performances (ce depuis les travaux de Schmid et Mohr [3]). C'est pourquoi notre code reprend cette approche, combinée avec les notions d'ordre et de contraste. La figure 1 illustre les différentes étapes de notre méthode, de l'image à sa signature.

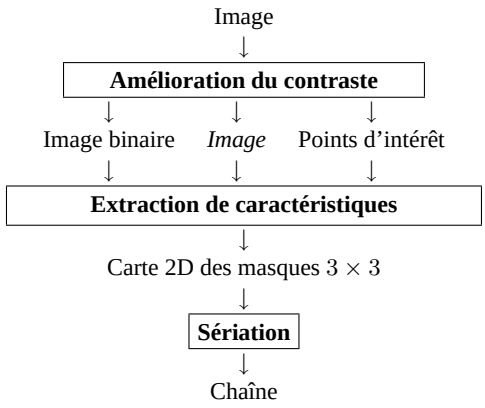


FIG. 1: *Schéma général de l'extraction du code à partir d'une image.*

2.1 Binarisation

La première étape du codage consiste en la binarisation de l'image, qui est préalablement convertie en niveaux de gris si nécessaire. Nous utilisons à ce stade une méthode développée il y a quelques années [4], et basée sur le rehaussement de contraste. Elle vérifie les étapes suivantes :

- représentation multirésolution d'une image sous la forme d'une pyramide,
- calcul d'une pyramide de contraste par ratio de deux niveaux successifs,

*Cette étude est financée par France Télécom R&D dans le cadre du contrat ECAV3 42568725.

FIG. 2: *Plusieurs niveaux de la pyramide initiale (à gauche) et sa binarisation obtenue en 7 itérations.*

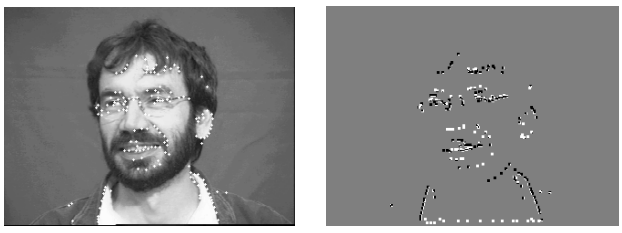
- réhaussement itératif du contraste de l'image à tous les niveaux de résolution,
- reconstitution de l'image réhaussée par projection top-down.

Le point de convergence de ce processus est une image binaire dont l'information de contraste a été préservée, tout en réduisant l'entropie, ce qui permet de conserver un maximum de détail. La figure 2 illustre cette convergence.

2.2 Points d'intérêt

La première étape de binarisation permet aussi l'extraction des points d'intérêt, point de départ de notre signature. En effet, ces points sont les maxima locaux de contraste, calculés lors de la première itération du mécanisme de binarisation. Ces points peuvent être extraits à plusieurs niveaux de résolution, dans le but de récupérer davantage d'information. La figure 3 est un exemple d'extraction de ces points.
Jouant le rôle d'un filtre sur l'image binaire, ces points d'intérêt et leur voisinage immédiat seront les seules informations retenues pour constituer le code d'une image. Ce code comprend deux composantes :

- un symbole binaire, centré sur un point d'intérêt, extrait de l'image binaire
- une double composante couleur associée à ce symbole, extrait à partir de l'image originale.



(a) (b)

FIG. 3: *(a) Image originale avec superposition des points d'intérêt, (b) masques binaires ,extraits de la version binaire de (a), dans le voisinage des points d'intérêt.*

2.3 Sériation

L'étape précédente fournit donc un ensemble de points, répartis dans l'image. A l'heure actuelle, comparer deux distributions 2D (deux graphes) reste encore très complexe, surtout pour plusieurs milliers de points (sommets). Le passage en 1D, suppose l'utilisation d'un procédé de sériation, qui peut être de différentes natures : parcours du graphe, parcours de l'image, etc. Ici, la sériation s'effectue simplement à partir d'un critère scalaire. En effet la liste de points est ordonnée par ordre décroissant de la valeur de contraste, obtenue lors de l'extraction des points. Cela ne permet malheureusement pas de conserver l'information spatiale contenue dans la distribution 2D de points, mais les autres méthodes de sériation développées lors de cette étude n'ont pas encore permis de mettre en évidence l'apport d'une telle information [5].

3 Un nouveau code

3.1 Définition

Le code associé à chaque image est donc une chaîne ordonnée d'entités, chacune constituée d'un symbole et d'une double composante couleur. Ces symboles, ou masques, composés de pixels noirs et blancs, sont de taille 3 par 3 pixels. Il en existe 512 différents, ce qui fournit une diversité conséquente, tout en restant raisonnable à étudier en détail.
Considérée comme un complément, l'information couleur devait être grossière. Pour chaque masque sont calculés deux couleurs : l'une correspondant aux pixels noirs N et l'autre aux pixels blancs B du masque. Elles résultent de la moyenne seuillée des composantes couleurs de ces deux zones N et B. On se ramène alors à seulement 8 valeurs pour chacune des deux couleurs extraites, ce qui permet déjà d'améliorer les résultats (cf. section 4), en gardant un code compact. En effet, sa taille est de 9 bits pour un masque et 2x3 bits pour la couleur, soit 15.n bits pour une chaîne de longueur n points. Pour une image I, on a :

$$I \rightarrow (\begin{smallmatrix} & & \\ & & \\ & & \end{smallmatrix}\ \begin{smallmatrix} c_{0.2} \\ c_{0.1} \end{smallmatrix}, \begin{smallmatrix} & & \\ & & \\ & & \end{smallmatrix}\ \begin{smallmatrix} c_{1.2} \\ c_{1.1} \end{smallmatrix}, \begin{smallmatrix} & & \\ & & \\ & & \end{smallmatrix}\ \begin{smallmatrix} c_{2.2} \\ c_{2.1} \end{smallmatrix}, \begin{smallmatrix} & & \\ & & \\ & & \end{smallmatrix}\ \begin{smallmatrix} c_{3.2} \\ c_{3.1} \end{smallmatrix}, \begin{smallmatrix} & & \\ & & \\ & & \end{smallmatrix}\ \begin{smallmatrix} c_{4.2} \\ c_{4.1} \end{smallmatrix}, \begin{smallmatrix} & & \\ & & \\ & & \end{smallmatrix}\ \begin{smallmatrix} c_{5.2} \\ c_{5.1} \end{smallmatrix} \ldots)$$

3.2 Comparaison de deux codes d'images

Le développement d'un code symbolique et non numérique permet d'avoir accès à l'ensemble des outils de la reconnaissance de formes, qui sont à la fois performants et originaux dans le domaine de l'indexation d'images. En particulier, une distance de type Levenstein, qui fait intervenir des coûts élémentaires de substitution, de suppression et d'insertion entre symboles, est dans notre cas tout à fait appropriée. En effet, pour savoir si deux images se ressemblent, il s'agit de déterminer la similarité de leur signature. Autrement dit, calculer quel est le coût pour passer d'une chaîne à l'autre, ce coût devant être d'autant plus élevé que les images comparées sont différentes. C'est exactement ce qu'offre ce genre de distances. Qui plus est, la notion d'ordre est conservée si les coûts élémentaires sont calculés entre deux entités de même rang dans la chaîne.
La distance de Levenstein étant trop coûteuse pour de longues chaînes, nous avons mis au point d'autres distances qui s'en inspirent. A commencer par une distance de distributions. Dans

ce cas on construit pour chaque chaîne un histogramme à 512 entrées qui contient les occurences des masques qui la composent. La différence entre les deux histogrammes ainsi obtenus représente alors la valeur de similarité entre les deux images comparées. A noter qu'on peut faire de même pour la couleur avec un histogramme à 8 entrées. Ici malheureusement, la notion d'ordre disparaît totalement.
Nous avons donc élaboré une troisième distance, à mi-chemin entre les deux précédentes : ici, l'incrément lors de la construction des histogrammes H_1 et H_2 correspond au coût de substitution entre deux masques de même rang. En outre, ce coût est pondéré par la position des masques dans les chaînes, afin de favoriser les masques placés en tête. Plus formellement, cela se traduit par :

$$H_1 \leftarrow 0\,; H_2 \leftarrow 0\,; W \leftarrow 0$$
$$\text{for } i : 1 \rightarrow \min(l(s_1), l(s_2))$$

$$\text{do} \begin{cases} H_1(m_1^i) = H_1(m_1^i) + \omega_i \times c(m_1^i, m_2^i) \\ H_2(m_2^i) = H_2(m_2^i) + \omega_i \times c(m_1^i, m_2^i) \\ W = W + \omega_i \times c(m_1^i, m_2^i) \end{cases}$$

$$d(s_1, s_2) = \textstyle\sum_i |H_1(i) - H_2(i)|/W$$

où $l(s_1)$ et $l(s_2)$ sont les longueurs des chaînes s_1 et s_2 ; $c(m_1^i, m_2^i)$ le coût de substitution entre deux masques de rang i ;
et ω_i la pondération du type $K - i$ (K constante ou égale à $\min(l(s_1), l(s_2))$+1). W permet de normaliser la valeur obtenue afin qu'elle soit indépendante de la longueur des chaînes comparées.
Une nouvelle fois, on peut procéder de même avec la couleur.

3.3 Distance entre deux symboles

Ainsi, pour connaître la similarité de deux chaînes, il est nécessaire d'établir le coût de substitution d'un masque par un autre, ces coûts étant alors stockés dans une matrice symétrique 512x512.
Le coût de base est la distance de Hamming, qui cumule le nombre de bits différents entre deux mots de 9 bits qui codent deux masques quelconques. Cependant, pour prendre en compte l'information symbolique contenue dans un masque, nous avons établit une distance dite structurelle. Pour se faire, les masques ayant les mêmes caractéritiques de forme ont été regroupés, et la distance de Hamming a alors été utilisée pour définir des distances moyennes inter- et intra-classes.
Par ailleurs, les coûts de substitution entre deux couleurs élémentaires ont été stockées dans une matrice 8x8, compte tenu du nombre de leurs composantes R, G et B qui diffèrent.
D'autre part, ces coûts peuvent aussi faire l'objet d'un apprentissage par l'exemple [6].

4 Expérimentations

Afin d'aborder au mieux les différents aspects offerts par la vaste thématique de l'indexation, nous avons mis au point différentes bases d'images permettant de répondre à un ensemble de questions (classification, reconnaissance d'une classe particulière, similarité à une image requête, robustesse à l'ajout de bruit, etc.). Par manque de place nous ne présenterons qu'une partie de nos résultats, dont le détail est présenté dans [6].

FIG. 4: *Echantillon d'images de la base 1. Sept catégories : bateaux, voitures, avions, outils, trains, légumes, objets vikings*

La figure 4 représente un échantillon d'images, issues de sept classes de 30 à 50 images, regroupées selon un critère sémantique tel que "voiture". Cette première base a pour objectif d'expérimenter la classification d'un ensemble d'images parmi un nombre prédéfini de classes.
La figure 5 montre l'avantage de notre méthode par rapport à une classification de référence basée sur les histogrammes couleurs. Avec tous les pixels de l'image, cette méthode permet d'obtenir 68.7% de bonne classification, quand la nôtre peut atteindre près de 74% avec seulement 2500 symboles. Cette figure présente aussi le résultat que l'on obtient si on optimise globalement la matrice de coûts élémentaires sur cette base d'image. Cette optimisation consiste en un "apprentissage", sur une base ayant les mêmes classes, des coûts de substitution entre deux masques quelconques : si les chaînes comparées appartiennent à la même classe, ces coûts sont réduits, sinon, ils sont augmentés. Par ailleurs, différents tests ont été menés afin de montrer l'apport de la couleur dans ce type de classification. En effet, la couleur peut être combinée aux masques suite à une étape de fusion de données que nous ne détaillerons pas ici. En outre, comme nous l'avons souligné plus haut, il est possible

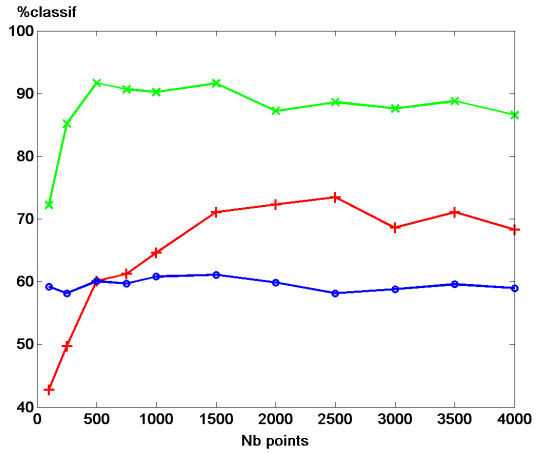


FIG. 5: *Pourcentage de bonne classification en fonction du nombre de points extraits. Valeurs obtenues avec les histogrammes couleurs (•), avec notre méthode prenant en compte l'ordre(+), avec l'optimisation (×).*

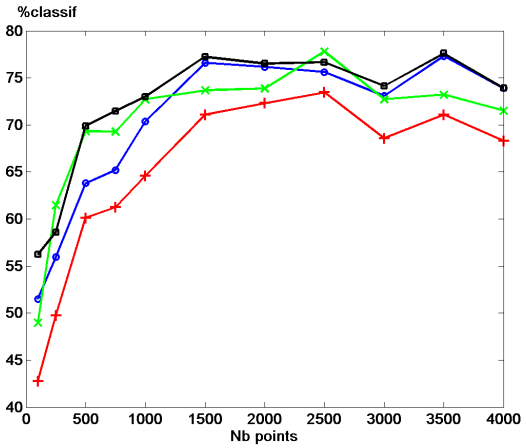


FIG. 6: *Pourcentage de bonne classification en fonction du nombre de points extraits. Valeurs obtenues avec les masques binaires seuls (+), avec les masques binaires et la couleur (•), avec les masques binaires extraits à deux niveaux de résolutions (×), avec les masques binaires et la couleur extraits à deux niveaux de résolution (□).*

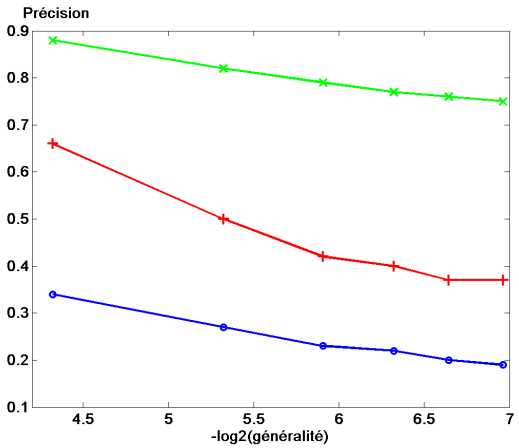


FIG. 7: *Précision moyenne pour 300 requètes en fonction de la généralité. Valeurs obtenues par Laurent* et al. *(+), avec notre méthode (masques et couleur, un seul niveau de résolution)(•), avec optimisation (×).*

d'extraire des points d'intérêt à divers niveaux de résolution, qui peuvent ensuite se combiner pour améliorer les résultats. La figure 6 illustre l'apport de ces deux informations supplémentaires.

Une autre base, composée initialement de 300 images réparties en 20 classes, à laquelle on ajoute progressivement 300 à 1500 images quelconques, permet de mettre en évidence la robustesse de la signature. On parle aussi de généralité:

$$\text{Généralité} = \frac{\text{Nombre d'images pertinentes}}{\text{Nombre d'images dans la base}} \qquad (1)$$

Cette base est initialement composée de 20 classes de 15 images chacune. La généralité pour une image donnée est donc $\frac{15}{300}$. Chacune de ces 300 images est successivement utilisée comme requète au sein de la base, afin de retrouver les 14 images les plus similaires. On calcule alors la précision moyenne obtenue. Ensuite, des paquets d'images quelconques sont progressivement ajoutés, réduisant ainsi la généralité. Le but est de tester la robustesse de la signature face à l'addition de "bruit" dans la base. Pour se faire, on calcule la nouvelle précision moyenne après chaque ajout de paquet, et ce jusqu'à l'addition de 1570 images (à comparer avec les 14 images similaires à une requète donnée).
Cette base ayant déjà été utilisée pour ce genre de tests nous avons comparé notre approche avec celle développée par Laurent *et al.* [7]. La figure 7 illustre les résultats obtenus.
Les résultats que l'on obtient sans optimisation sont inférieurs à ceux de Laurent *et al.*, dont la méthode est, rappelons le, beaucoup plus complexe que la nôtre. En revanche, lorsqu'on optimise la matrice de coûts sur la base des 300 images, la précision moyenne est toujours supérieure à 0.7.

5 Conclusion

La méthode présentée ici propose une signature d'image originale, basée sur une approche structurelle. Cette démarche, qui se distingue des approches numériques traditionnelles, offre de ce fait des outils innovants en indexation d'image tels que ceux de la reconnaissance de formes. En outre, ce codage symbolique semble plus à même de refléter le contenu sémantique d'une image, ce que confirment les résultats encourageants obtenus.
Différents points restent cependant à améliorer (la couleur, pour l'instant très grossière) et à prospecter : une étude plus précise des chaînes permettra de déterminer des longueurs optimales ; la mise en place d'une méthode de comparaison à la volée offrira un résultat plus rapide ; etc.

Références

[1] P.J. Parkhurst et E. Niebur. Scene content selected by active vision. *Spatial Vision*, 16(2):125-154, 2003.

[2] S.J. Thorpe, A. Delorme, et R. VanRullen. Spike-based strategies for rapid processing. *Neural Networks*, 14:715-725, 2001.

[3] C. Schmid et R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans. on PAMI*, 19(5):530-535, 1997.

[4] J.M. Jolion. Analyse multirésolution du contraste dans les images numériques. *Traitement du Signal*, 11(3):245-255, 1994.

[5] J. Ros, C. Laurent, J.M. Jolion et I. Simand. Comparing String Representations and Distances in a Natural Images Classification Task, L. Brun and M. Vento (Eds), *GbR-PR'05*, 2005, LNCS 3434, 71-83.

[6] J.M. Jolion, I. Simand et P. Prabhat. Développement d'une nouvelle représentation de formes contenues dans des images par analyse du parcours visuel et points saillants. Rapport technique ECAV-3, LIRIS, INSA Lyon, 2004.

[7] C. Laurent, N. Laurent, M. Maurizot et T. Dorval. In depth analysis and evaluation of saliency-based color image indexing methods using wavelet salient features. *Multimedia Tools and Application*, 2004.