Sélection de singularités locales par stimulation de carte GHSOM

Grégoire Lefebvre[1], Christophe Garcia[1], Jean-Marc Salotti[2], Julien Ros[1]

[1]France Telecom R&D - TECH/IRIS/ICM
4, rue du Clos Courtel 35512 Cesson Sévigné Cedex - FRANCE

[2]Institut de Cognitique - Université Victor Segalen
146, rue Léo Saignat 35512 33076 Bordeaux Cedex - FRANCE

{prénom.nom}@orange-ftgroup.com, salotti@idc.u-bordeaux2.fr

Résumé – Dans cet article, nous nous intéressons à la sélection des singularités présentes dans les images. Nous élaborons une architecture reposant sur une carte auto-organisatrice en extension hiérarchique (GHSOM) pour reconnaître différents concepts visuels à partir d'une approche locale et éparse pour construire une représentation sous la forme de «sac de caractéristiques». Les signatures locales activent les neurones de la carte GHSOM produisant des activations neuronales hiérarchiques. Ces stimuli révèlent la classe d'appartenance de l'image. Dans le cadre d'une application de classification d'images naturelles, nous obtenons des résultats convaincants avec un taux de classification supérieure à 82%. L'utilisation de la méthode pour une application biométrique offre plus de 95% de bonnes identifications.

Abstract – In this paper, we focus on singularity selection for classifying visual concepts. We design a scheme that relies on a Growing Hierarchical Self-Organizing Map (GHSOM). Robust local signatures are first extracted and projected into a specialized GHSOM network in order to obtain a "bag of features" representation. The extracted signatures activate several neurons producing hierarchical neural activations. These stimuli reveal the image content. For natural scene recognition, we obtain persuasive results with more than 82% of classification rate. For face recognition, this method offers more than 95% of good identifications.

1 Introduction

Selon les études psycho-visuelles de Hoffman [4], le système visuel humain exécute des mouvements saccadiques entre des régions saillantes pour capturer le contenu des images. De nombreux travaux en vision par ordinateur s'inspirent de cette observation pour décrire l'information visuelle des images dans une optique d'indexation, de classification ou de détection d'objets. Contrairement aux approches globales, pour lesquelles une signature unique est calculée en considérant tous les pixels de l'image avec la même importance, ces approches locales représentent le contenu de l'image par un ensemble de signatures locales centrées autour de points saillants. La détection de ces points d'intérêt [5] se focalise ainsi dans les zones perceptuellement significatives de l'image. Cette représentation parcimonieuse propose de décrire localement une image autours de points saillants par les contours présents dans cette région. Ici, nous proposons une étude comparative entre trois descripteurs possédant les propriétés d'invariance en rotation et en changement d'échelle : SIFT (*Scale-Invariant Feature Transform* [6]), SURF (*Speed Up Robust Features* [1]) et RFD (*Regularity Foveal descriptor* [8]). De plus, les travaux de Tversky [9] montre que comparer deux images revient à détecter des concepts d'appartenance et d'exclusion entre ces régions saillantes. Notre méthode tente de reproduire cette extraction de concepts par la constitution d'un histogramme d'activation neuronale basés sur l'information d'un réseau GHSOM (*Growing Hierarchical Self-Organizing Map* [3]) soumis à la stimulation de descripteurs locaux. Le réseau de neurones GHSOM est un arbre de cartes SOM ayant la propriété de s'adapté aux données d'apprentissage par un processus d'élargissement ou d'expansion de ses feuilles SOM. Une sélection d'information est alors effectuée par l'apprentissage de la carte GHSOM et la distinction entre les concepts visuels est estimée par un classifieur supervisé SVM (*Support Vector Machine*).

Cet article est organisé comme suit : la section 2 présente notre système de sélection de singularités locales par stimulation de carte GHSOM. Ensuite, les résultats expérimentaux de la section 3 illustrent les performances d'une telle architecture dans le cadre d'une application de classification d'images naturelles et de reconnaissance de visages. Finalement, nous dressons quelques conclusions.

2 Sélection de caractéristiques par GHSOM

2.1 Architecture du système

Classiquement, un processus de classification se déroule en trois étapes : une étape de prétraitements, une seconde d'ex-

traction de caractéristiques et finalement une étape de classification. Dans cette étude, nous nous intéressons principalement aux deux premières phases, la dernière étant réalisée par un classifieur SVM. Pour construire le vecteur caractéristique $\mathbf{H}_{\mathcal{I}}$ pour chaque image $\mathcal{I}$, on procède comme suit (cf. Figure 1) :

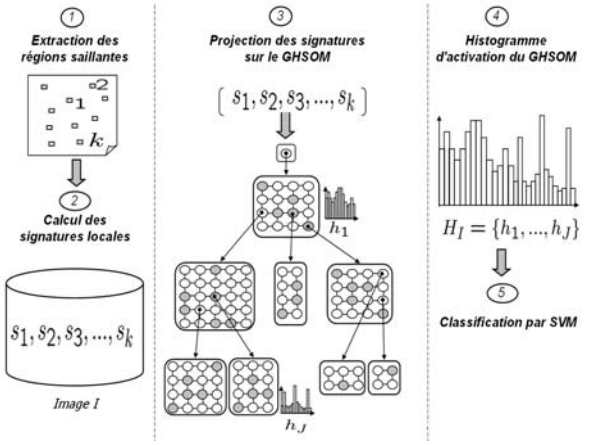

FIG. 1 – Architecture du système proposé

- Pour chaque image $\mathcal{I}$ d'apprentissage :
  - Le détecteur [5] extraits des points saillants.
  - Pour chaque région d'intérêt $k \in \mathcal{I}$:
    - Nous calculons la signature locale s_k avec selon l'expérimentation le descripteur SIFT, SURF ou RFD pour composer une base d'apprentissage.
- Les signatures locales issues de cette base permettent l'apprentissage du réseau GHSOM pour toutes les catégories, formant ainsi un dictionnaire visuel.
- Pour chaque image $\mathcal{I}$ de la catégorie $\mathcal{C}$:
  - Pour chaque signature s_k :
    - Le réseau GHSOM, précédemment appris, reçoit la signature et détermine le neurone BMU (*Best Matching Unit*) c pour chaque feuille SOM de l'arbre GHSOM. Le BMU est défini par :

$$c = \arg\min_i ||s_k - m_i||, \forall i = \{1, \ldots, u\}, \quad (1)$$

      avec u le nombre de neurones dans la feuille SOM considérée et m_i ses poids neuronaux.
    - Chaque histogramme d'activation h_j correspondant à chaque feuille est mis à jour avec l'erreur de quantification du BMU c :

$$h_j[c](t+1) = h_j[c](t) + ||s_k - m_c||, \quad (2)$$

      avec $j = \{1, \ldots, J\}$ l'indice de la feuille SOM et t le temps variant de 0 à $k-1$.
  - Chaque histogramme d'activation h_j, calculé à partir de toutes les signatures de l'image $\mathcal{I}$, est concaténé dans un histogramme global d'activation $\mathbf{H}_{\mathcal{I}}$.
  - Ce vecteur caractéristique $\mathbf{H}_{\mathcal{I}}$ est introduit dans un classifieur SVM pour un apprentissage supervisé.

La stimulation du réseau GHSOM par une signature locale révèle une erreur de quantification pour chaque BMU de chaque feuille SOM de l'arbre. La somme de ces erreurs incrémente l'histogramme d'activation GHSOM ; cette valeur représente la correspondance entre la signature et le réseau GHSOM. Ces activations neuronales permettent d'apprendre l'adéquation entre une image de test et les images d'apprentissage par une classifieur SVM, dévoilant ainsi la classe d'appartenance de l'image considérée.

2.2 Apprentissage GHSOM

Pour construire une représentation robuste sous la forme de «sac de caractéristiques», nous projetons les signatures locales sur des cartes topologiques afin de sélectionner les informations pertinentes. L'extension du modèle SOM par le réseau GHSOM [3] permet une décomposition de l'espace d'apprentissage en sous-parties, par une stratégie d'expansion hiérarchique. Tandis que la taille d'un réseau SOM est fixé *apriori*, la structure GHSOM s'adapte à la distribution des données pendant la phase d'apprentissage. En effet, le réseau GHSOM est un arbre de cartes SOM dont la taille des branches et la configuration des feuilles varient en fonction des données. Ainsi, le processus d'apprentissage est composé de deux phases : une stratégie d'élargissement et une expansion, correspondant respectivement à deux critères τ_1 et τ_2.

L'initialisation débute par la création de deux cartes SOM :
- Premièrement, une carte SOM singleton (composé d'une seule cellule) est crée au niveau 0 de l'arbre GHSOM. L'unique cellule est définit par le vecteur $m_0 = [\mu_{0_1}, \mu_{0_2}, \ldots, \mu_{0_d}]^T$, $m_0 \in \Re^d$, égal à la moyenne de toutes les vecteurs d'entrée.
- En second, une carte SOM 2×2 est initialisée au niveau 1 avec quatre vecteurs aléatoires $m_i \in \Re^d, m_i = [\mu_{i_1}, \ldots, \mu_{i_d}]^T, i \in \{1, 2, \ldots, 4\}$.

La stratégie d'élargissement des cartes SOM commence avec le calcul de l'erreur moyenne de quantification (EMQ) $\mathbf{emq}_0$ de la cellule 0 de la première carte. Ce calcul est donné par l'équation 3 avec K le nombre de signatures s_k :

$$\mathbf{emq}_0 = \frac{1}{K} \sum_{j=0}^{K-1} ||m_0 - s_k|| \quad (3)$$

Pour définir la deuxième feuille SOM, une règle d'apprentissage est appliquée au voisinage du BMU c par :

$$m_i(t+1) = m_i(t) + \lambda(t)\phi_{ci}(t)[s_k - m_i(t)], \quad (4)$$

où $\lambda(t)$ désigne le taux d'apprentissage décroissant avec le temps avec $0 < \lambda(t) < 1$. $\phi_{ci}(t)$ représente la fonction de voisinage. Classiquement, une fonction gaussienne est utilisée :

$$\phi_{ci} = \exp - \frac{||r_c - r_i||^2}{2\delta(t)^2}. \quad (5)$$

Ici, la norme euclidienne est choisie et r_i est la position 2D pour le i^{eme} neurone de la carte. $\delta(t)$ spécifie la largeur du voisinage décroissant au cours du temps t de $\frac{\sqrt{2}}{2}u$ to 0.5.

$$\mathbf{EMQ}_l = \frac{1}{u} \cdot \sum_i \mathbf{emq}_i \quad (6)$$

Ensuite, l'erreur EMQ de la carte est calculée par l'équation 6 après un nombre fixé d'itérations $\zeta = 500 \times u$. Dans cette formule, $\mathbf{EMQ}_l$ est l'erreur EMQ du niveau l, $\mathbf{emq}_i$ est l'erreur EMQ de la cellule i et u le nombre de cellule dans la carte. Par analogie avec l'équation 3, $\mathbf{emq}_i$ est calculé comme étant la distance moyenne entre les référents m_i et les entrées projetées sur la cellule i. Evaluer l'erreur EMQ de la carte est utilisé pour le critère d'élargissement (cf. critère 7). L'idée sous-jacente est que chaque couche du GHSOM explique la déviation standard des données projetées dans la couche supérieure. Ainsi, les cartes SOM de la couche étudiée grandissent jusqu'à ce que l'erreur EMQ de la couche précédente soit réduit par le seuil τ_1.

$$\mathbf{EMQ}_1 \geq \tau_1 \cdot \mathbf{emq}_0 \quad (7)$$

Ainsi, aussi longtemps que le critère 7 reste vrai, soit une ligne, soit une colonne est ajouté à la carte étudiée. Suivant la figure 2, cette insertion est guidée par la cellule e possédant la plus grande erreur moyenne de quantification $\mathbf{emq}_e$ et son voisin p le plus dissimilaire. Alors, on insert une ligne et les poids des nouvelles cellules sont simplement choisis comme la moyenne des cellules précédentes deux à deux. Il résulte une carte SOM de 9 cellules présentée à droite de la figure.

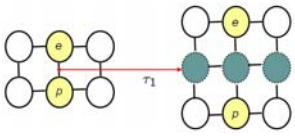

FIG. 2 – Insertion d'une nouvelle ligne.

La stratégie d'expansion commence lorsque l'élargissement de la première carte est achevé (le critère 7 n'est plus satisfait). Toutes les cellules sont alors examinées et celles possédant la plus grande erreur EMQ définissent de nouvelles cartes SOM pour le niveau inférieur. Un nouveau paramètre τ_2 décrit la granularité de la profondeur de l'arbre GHSOM. Plus précisément, chaque cellule i vérifiant le critère 8 subit l'expansion.

$$\mathbf{emq}_i > \tau_2 \cdot \mathbf{emq}_0 \quad (8)$$

Si on considère la figure 3, le réseau GHSOM décrit huit cartes SOM disposées sur quatre niveaux. Soit la cellule e de la cinquième carte satisfaisant le critère 8. A ce stade, une nouvelle carte 2×2 est attaché au neurone e. Pour définir les poids initiaux de cette carte, prenons la cellule r comme exemple. Ainsi, le vecteur associé à r sera la distance moyenne entre les vecteurs des cellules a, b, c avec le vecteur référent de e. L'apprentissage de la carte SOM se poursuit alors, ainsi que la procédure d'élargissement.

Finalement , le paramètre τ_1 contrôle la largeur de chaque carte SOM, et τ_2 spécifie la qualité désirée de représentation des données d'observation.

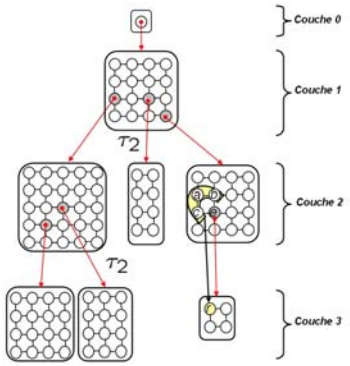

FIG. 3 – Processus d'expansion

3 Expérimentations

3.1 Les bases d'images

FIG. 4 – Les bases d'images FERET et SIMPLIcity

3.1.1 FERET

Pour tester notre approche, nous avons utilisé la base de visage FERET[1]. Pour construire la galerie (i.e. la base de connaissance), nous avons choisi le corpus comprenant 46 individus à reconnaître. Deux images servent à tester l'identité d'une personne, le reste étant alloué à l'apprentissage. Nous avons créé un modèle de visage 200×200 par individu dans lequel nous extrayons les points d'intérêt et calculons les signatures locales relatives. Cette approche fait partie des techniques analytiques de reconnaissance de visages, par opposition aux approches holistiques considérant le visage dans sa globalité.

3.1.2 SIMPLIcity

La base SIMPLIcity[2] contient 1000 images de taille 384×256 extraites de la base commerciale COREL. Cette base contient dix catégories : *african people, beaches, buildings, buses, dinosaurs, elephants, flowers, food, horses and mountains.* Ces concepts sont assez variés et les frontières interclasses ne sont pas triviales. Pour tester notre approche de classification, nous divisons la base d'images en deux parties égales : 500 images d'apprentissage et 500 images de test.

[1] http ://www.feret.org
[2] http ://wang.ist.psu.edu/ jwang/test1.tar

3.2 Configuration du système

Pour déterminer une configuration idéale du système, nous comparons trois types de paramétrage pour GHSOM :
- «élargissement» : $\tau_1 = 0,10$ et $\tau_2 = 0,35$;
- «expansion» : $\tau_1 = 0,35$ et $\tau_2 = 0,10$;
- «équitable» : $\tau_1 = 0,10$ et $\tau_2 = 0,10$;

Le descripteur SURF est utilisé dans sa forme originel comme proposé par Bay, sa dimension est de 64. Le paramétrage classique du descripteur SIFT offre une signature de taille 128, et le RFD est utilisé avec 8 orientations, 3 exposants de Hölder et 16 sous-régions pour une dimension de 384. La classification finale est réalisée par un SVM au noyau gaussien : $K(X_i, X_j) = exp(-\alpha||X_i - X_j||^2)$, $\alpha = 0,02$.

3.3 Comparaison des performances

	élargissement	expansion	équitable
RBF	50%	69.6%	61.9%
MLP	73.9%	94.6%	94.5%
SVM	69.6%	95,7%	92.4%

TAB. 1 – Classification FERET selon 3 stratégies GHSOM

Sur la base FERET, nous avons choisi d'illustrer le pouvoir de discrimination de la méthode sur la problématique de reconnaissance de visages pour un usage biométrique. Nous réalisons ainsi une étude comparative autour du pouvoir de sélection d'information RFD pour trois configurations distinctes de GHSOM. Il est intéressant également d'observer la différence de performance avec les classifeurs RBF (*Radial Basis Function*) et MLP (*Multi-Layer Perceptron*). Le tableau 1 montre que la meilleure configuration retenue est une architecture favorisant l'expansion de la carte GHSOM ($\tau_1 = 0.35$ et $\tau_2 = 0.10$) suivi d'un classifieur SVM à noyau gaussien. Ce choix d'architecture nous permet ici d'obtenir un taux de reconnaissance de $95,7\%$. Nous conservons par conséquent ce paramétrage du système pour la section suivante.

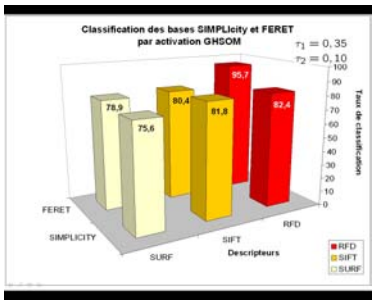

FIG. 5 – Classification des bases SIMPLIcity et FERET

Dans une seconde application, nous démontrons le pouvoir de généralisation de la méthode pour une classification d'images naturelles où les concepts visuels à apprendre sont variés. Pour la base SIMPLIcity, nous obtenons $82,4\%$ de bonne classification (cf. figure 5). Le descripteur RFD révèle également ici son pouvoir de structuration de l'information de singularité saillante par comparaison avec les descripteurs SIFT et SURF. Selon [7] et [2], les performances des systèmes actuelles varient de $37,5\%$ à $84,1\%$. Néanmoins, dans ces papiers, l'expérimentation est réalisée sur la méthode de validation croisée *leaving-one-out*. Celle-ci consiste à tester chaque image avec un classifieur entrainé sur les 999 images restantes de la base. Cette approche est donc plus simple, que notre protocole découpant les images en deux ensembles équitables (500 images pour l'apprentissage et 500 images pour le test). Par conséquent, on estime que notre système propose une des meilleures classification sur cette base.

4 Conclusion

Cet article propose un système original de classification d'images naturelles, utilisant l'information des singularités contenues dans les régions de forte saillance. Sur la base des trois principales propriétés des réseaux GHSOM, qui sont : la réduction de dimension, la préservation de la topologie et l'émergence de caractéristiques invariantes, notre architecture offre des résultats prometteurs avec un classifieur SVM. On envisage plusieurs perspectives à ces travaux pour ajouter des contraintes géométriques afin de représenter au mieux la distribution des informations. De même, nous jugeons pertinent d'investiguer un nouveau noyau SVM prenant en compte la topologie des réseaux GHSOM.

Références

[1] Bay H., Tuytelaars T., and Van Gool L. SURF : Speeded Up Robust Features. In *9th European Conference on Computer Vision*, 2006.

[2] Deselaers T., Keysers D., and Ney H. Features for Image Retrieval – A Quantitative Comparison. In *DAGM'04 : 26th Pattern Recognition Symposium*, Tübingen, Germany, September 2004.

[3] Dittenbach M., Merkl D., and Rauber A. The growing hierarchical self-organizing map. In *IJCNN '00*, volume 6, page 6015, 2000.

[4] Hoffman J.E. and Subramaniam B. The Role of Visual Attention in Saccadic Eye Movements. *Perception & Psychophysics*, pages 787–795, 1995.

[5] Laurent C., Laurent N., Maurizot M., and Dorval T. In Depth Analysis and Evaluation of Saliency-based Color Image Indexing Methods using Wavelet Salient Features. *Multimedia Tools and Application*, 2006.

[6] Lowe D.G. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2) :91–110, 2004.

[7] Maree R., Geurts P., Piater J., and Wehenkel L. Random subwindows for robust image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 34–40, San Diego, USA, June 2005.

[8] Ros J., Laurent C., and Lefebvre G. A cascade of unsupervised and supervised neural networks for natural image classification. In *CIVR*, pages 92–101, 2006.

[9] Tversky A. Features of similarity. *Psychological Review*, 4(84) :327–352, 1977.