Classification de textures par le modèle « sac de phrases »

THUC TRINH LE[1], RONAN FABLET[1], JEAN-MARC BOUCHER[1]

[1] Lab-STICC, Institut Mines-Telecom, Telecom Bretagne CS 83818 29238 Brest Cedex 3, France

[1]{Thuc.Le, Ronan.Fablet, JM.Boucher}@telecom-bretagne.eu

Résumé - Dans cette étude, nous proposons un nouveau descripteur de texture basé sur la caractérisation de l'organisation spatiale des points d'intérêt d'une image texturée. Le descripteur SURF, qui a des propriétés d'invariance au changement de contraste et aux transformations géométriques, détecte d'abord les points d'intérêts et associe à chaque point un descripteur visuel. Le modèle « sac-de mots » calculé dans un voisinage local orienté d'un point d'intérêt est utilisé pour extraire une nouvelle signature locale, définissant une phrase. Une description globale de la texture est obtenue par la construction d'un dictionnaire à partir de ces nouvelles signatures menant au concept de « sac des phrases ». Nous appliquons la méthode à la classification de texture en utilisant la base UIUC. Cette application démontre l'intérêt de l'approche proposée vis-à-vis du modèle « sac-de-mot ».
Mots-clés— Classification de texture, sac de mot, point d'intérêt.

Abstract - In this paper, we propose a new description of texture based on Bag-of-Phrases. The SURF algorithm, which has invariance properties to contrast change and geometric deformations, extracts image keypoints and associates a visual descriptor to each point. Bags-of-Features computed in a local neighborhood around each keypoint are used as new local signature, defining a phrase. A global description of texture is obtained by building a codebook from these new features leading to bag-of-phrases. We address an application to UIUC texture classification and demonstrate the relevance of the proposed approach compared to Bag-of-Features.

Index Terms— Texture classification, bag of features, keypoint.

1 Introduction

L'analyse de la sémantique de l'image est une problématique importante pour de nombreux thèmes de recherche en robotique [1] et vision par ordinateur [2]. Les caractéristiques locales d'image [2] telles que les points d'intérêt conduisent à plusieurs améliorations significatives pour la reconnaissance d'image, notamment l'apprentissage des modèles de classification des signatures visuelles de ces points d'intérêts [3, 4, 5]. Ces signatures visuelles peuvent-être construites à partir des distributions locales de l'orientation des gradients d'image tels que les descripteurs SIFT [6] et SURF [7], ces derniers étant invariants et robustes aux changements de contraste et aux transformations géométriques, et améliorant les performances de reconnaissance d'image lorsque la base d'apprentissage est petite [1, 8]. L'approche « sac de mots » (SdM) [3] exploite les statistiques sur les occurrences de différents mots visuels dans l'image. Cette approche, cependant, ignore toutes les informations concernant la disposition spatiale des caractéristiques locales dans une image texturée.

La caractérisation conjointe des motifs visuels et de la disposition spatiale des ensembles de points d'intérêts a été traitée dans des travaux antérieurs [2, 9, 10]. Il a été montré que la caractérisation résultant de la disposition spatiale des ensembles visuels des points d'intérêts améliore les performances de reconnaissance. D'un point de vue statistique, les points d'intérêts peuvent être considérés comme une collection aléatoire associées à des processus ponctuels spatiaux (voir [11, 12]). Dans l'article [5], les statistiques spatiales du second ordre ont été associées au modèle SdM, et dans les articles [13, 14], un modèle probabiliste spatial, à savoir le modèle de Cox log-gaussien, a été utilisé pour fournir une caractérisation conjointe de l'information visuelle et spatiale amenée par ces collections aléatoires.

Dans cet article, le modèle « sac de mots » ne s'applique que dans un voisinage, potentiellement anisotrope, de chaque point d'intérêt, donnant de nouvelles caractéristiques locales qui contiennent l'information spatiale. Toutes ces caractéristiques sont ensuite assemblées en catégories de la même manière que dans l'approche SdM, donnant un sac de phrases (SdP), qui constitue une nouvelle description de l'image et qui peut être utilisé dans la classification de texture. Dans la suite, la section 2 présente un bref aperçu de la méthode SdM. La section 3 détaille l'approche proposée. Dans la section 4, nous présentons les applications à la classification de texture et comparons les résultats à l'état de l'art. Nous concluons dans la section 5.

2 Classification avec le modèle « sac de mots visuels »

Le modèle « sac de mots visuels » est une méthode éprouvée en représentation d'image, qui s'inspire de la représentation d'un document par l'histogramme des occurrences des mots le composant. Ce modèle ne tient pas compte des relations contextuelles des mots et classe le document en fonction de cet histogramme. En vision par ordinateur, le modèle « sac de mots » peut-être appliqué à la classification des images en traitant les points d'intérêts de l'image comme les mots visuels. Les principales étapes de la méthode sont :

- Détecter et décrire tous les points d'intérêt.
- Attribuer les descripteurs de ces points à un ensemble de groupes prédéterminés (un vocabulaire) avec un algorithme de quantification vectorielle.
- Construire un modèle « sac de mots visuels », qui compte le nombre de points assigné à chaque groupe.
- Traiter le « sac de mots visuels» comme le vecteur de caractéristiques d'un classifieur multi-classes et déterminer la classe à laquelle l'image est attribuée.

La détection et la description des points d'intérêt sont obtenus par le SURF (Speeded Up Robust Features) [7], partiellement inspiré par le descripteur SIFT, qui est le plus couramment utilisé pour l'extraction de points d'intérêt. Le principal avantage de SURF est que l'extraction des points d'intérêts se fait plus rapidement, notamment grâce à l'utilisation d'images intégrales. Il est peu sensible au changement d'intensité, de mise à l'échelle et de rotation, ce qui fait de lui un descripteur très robuste. La taille du vecteur de caractéristiques est ramenée en général à 64 afin d'améliorer la vitesse de calcul. Après avoir calculé les vecteurs de caractéristiques, un algorithme k-moyenne est appliqué à l'ensemble des images pour former un dictionnaire. Une image est finalement représentée par un histogramme de fréquence d'apparition de mots visuels dans le dictionnaire.

3 Le modèle « sac de phrases visuels »

Comme mentionné précédemment, l'histogramme global de mots visuels ne contient aucune information spatiale entre les points d'intérêts. Nous proposons de localiser le modèle « sac de mots » dans un voisinage orienté autour chaque point d'intérêt, puis d'analyser globalement les relations de cette description locale, ce qui est une façon d'introduire de l'information spatiale. Un voisinage local autour de chaque point d'intérêt est défini par une ellipse E (x, y, θ, R, g) de centre (x, y), avec l'orientation principale θ, le rayon R du cercle ayant la même surface que l'ellipse et un facteur d'aplatissement g [fig.1]. L'utilisation du facteur R permet de s'assurer que la zone de l'ellipse reste constante lorsque le facteur aplatissement g varie. Le choix d'une ellipse provient du désir de prendre en compte le fait que les textures présentent généralement des structures spatiales orientées.

Dans chaque ellipse un modèle SdM local est construit, où chaque barre de l'histogramme correspond à un mot visuel [fig.2]. L'orientation principale de chaque point d'intérêt à l'intérieur de l'ellipse, calculée par le détecteur SURF, est aussi un autre élément qui peut être ajouté. Cette orientation est utile pour discriminer des textures orientées de celles qui sont isotropes. [fig. 3].

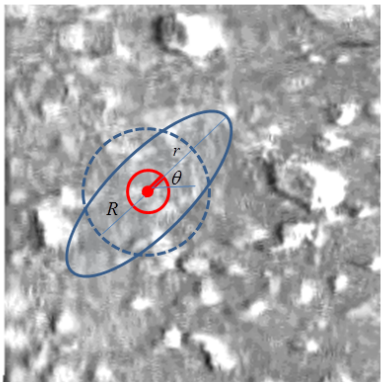

Figure 1: Voisin local d'un point d'intérêt

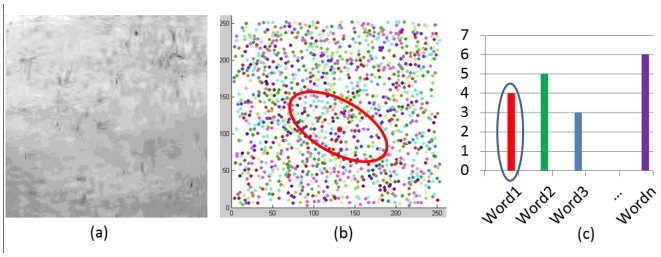

Figure 2: Le modèle "sac de mots" local d'un point d'intérêt: (a) l'image origine, (b) l'image représentée par un ensemble de mots visuels, (c) histogramme local de point d'intérêt en (b), chaque barre de l'histogramme correspond à un mot visuel dans l'ellipse

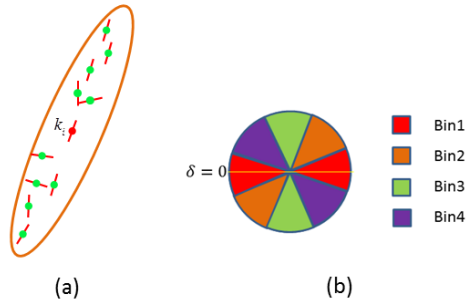

Figure 3: L'histogramme local d'orientation: (a) voisinage local d'un point, (b) quantification des orientations basée sur la différence entre l'orientation de l'ellipse et l'orientation individuelle de chaque point d'intérêt.

Après ces opérations, on obtient un histogramme local qui contient les deux informations d'orientation et de label du mot visuel et qui constitue une phrase. On forme ensuite un dictionnaire de phrases. Cela peut être fait en tenant compte d'une quantification scalaire ou bien vectorielle. Dans le cas de la quantification scalaire, chaque barre de l'histogramme, correspondant à un mot visuel, est regroupée indépendamment et une phrase est définie comme un triplet (w, f, b) où w est le label du mot visuel, f est la fréquence du mot w dans l'ellipse et b est le bin d'orientation. En quantification vectorielle, une phrase est définie comme un vecteur k-dimensionnel, où k est le nombre de mots visuels, correspondant à un histogramme entier. Ce dernier est traité comme un point dans l'espace visuel des mots, puis, un algorithme k-moyenne est appliqué. Une approche plus générale peut être appliquée par l'extraction de toutes les n-phrases de l'histogramme local, où n est le nombre de barres dans l'histogramme local (n variant de 1 à k), comme par exemple une paire,

un triplet… On utilise ensuite un critère pour choisir le meilleur candidat, comme la fréquence ou l'entropie. Dans cet article, nous comparons ces trois cas.

Après avoir défini le vocabulaire de phrases, on procède à la caractérisation d'une image en comptant le nombre de phrases dans chaque image, ce qui conduit à un histogramme global de phrases utilisé dans la classification.

Dans l'application à la texture, nous choisissons une classification par SVM non linaire avec une noyau Chi-2 :

$$K(x,y) = 1 - \sum_{i=1}^{k} \frac{(x_i - y_i)^2}{\frac{1}{2}(x_i + y_i)}$$

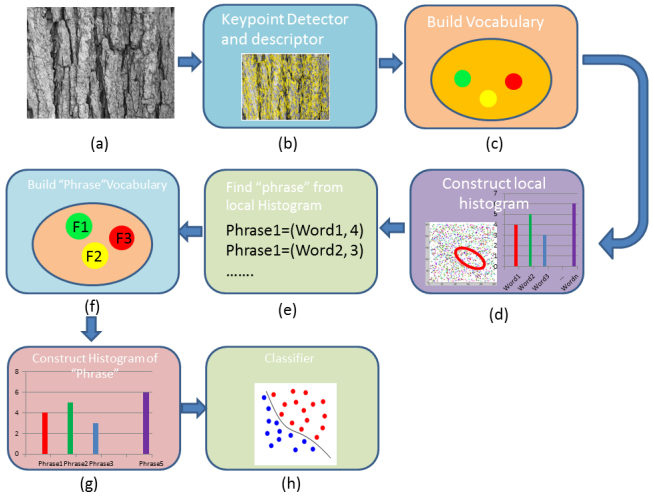

Figure 4: La méthode proposée: (a) L'image originale, (b) extraction des caractéristiques, (c) construction la dictionnaire des mots visuels, (d) construction de l'histogramme local, (e) extraction des phrases, (f) construction du dictionnaire de phrases, (g) construction de l'histogramme global de phrases, (f) classification de l'image.

4 Evaluation de l'approche:

4.1 Base de texture UIUC

L'évaluation utilise la base de textures UIUC qui comporte 25 classes avec des textures différentes. Chaque classe est composée de 40 images de taille 640x480 pixels. Des exemples sont présentés en Fig.5. Chaque classe comporte des échantillons avec des variations de contraste, d'échelle et rotation. L'apprentissage des modèles de classification est réalisé à partir d'un sous-ensemble d'images de chaque classe choisies aléatoirement, pour N_t images d'apprentissage par classe. Cette étape d'apprentissage est répétée 20 fois pour évaluer les performances de classification, en termes de taux moyen de bonne classification.

4.2 Résultats de classification

Dans [5], nous avons déjà comparé l'approche statistique descriptive de cooccurrence (SSC) à celle classique de descripteurs de Gabor, de matrices de cooccurrence et du modèle « sac de mots » [5], montrant les meilleures performances de SSC sur cette base de données. Nous en avions conclu que l'apport d'une information spatiale dans le modèle SSC améliorait les performances de classifications par rapport aux autres méthodes, notamment SdM. Nous comparons donc ici seulement l'approche « sac de phrases » (SdP) à SdM et SSC.

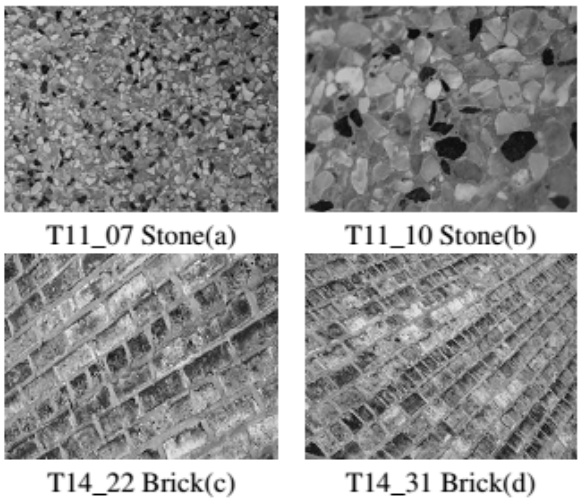

Figure 5: Les textures exemples de la donnée UIUC.

- Sac de mots [4] : L'algorithme SDM exploite les statistiques d'occurrence des différentes catégories de mots visuels dans l'image. Différents nombres de catégories de mots visuels k = {60, 120, 150} sont étudiés.
- Statistiques descriptives de cooccurrence (SSC) [5]: Dans cette application de classification d'images texturées, comme pour SdM, les nombres de catégories de mots visuels k = {60, 120, 150} sont étudiés. En outre, le nombre de paires de points d'intérêt k $*$ = {60} est considéré dans la procédure de réduction de la dimension de la taille du descripteur proposé. Les procédures d'adaptation d'échelle et de correction de bord sont appliquées.

Un ensemble de points d'intérêt d'échantillonnage aléatoire est exploité pour chaque étape de groupement hiérarchique. Les paramètres du voisinage local sont fixés à R = 30 et g = 0,5. Nous avons constaté expérimentalement que le choix de R n'est pas trop critique et peut être pris dans une large gamme de valeurs sans changement notable du taux de classification.

Les résultats de la classification en fonction du nombre d'images d'apprentissage N_t sont présentés dans la Table 1. On peut voir que la quantification vectorielle donne de meilleures performances que la quantification scalaire lorsque le nombre d'images d'apprentissage est très faible, mais la différence diminue quand le nombre d'images d'apprentissage augmente. SdM a des performances plus faibles que le SdP dans les 3 cas, ce qui prouve que la nouvelle information spatiale qui a

été ajoutée en considérant l'interaction entre les phrases locales est utile.

La différence de taux de classification entre SdP et SSC est de 1 à 3% en quantification vectorielle. On peut donc dire que les deux méthodes fournissent des résultats voisins. Le vecteur de descripteur fourni par SSC est de dimension Nrk*, où Nr est le nombre de rayons utilisés, typiquement 20, et k* le nombre de classes, ici 60. Il doit être comparé à la taille du vocabulaire, ici 200. Donc SdP fournit des résultats de classification sensiblement similaires à SSC, mais donne un descripteur moins complexe. Dans le dernier cas, quand toutes les n-phrases sont extraites, les résultats sont moins bons quand N_t est petit, car le nombre de bons candidats à sélectionner est insuffisant, mais lorsque N_t augmente, le taux de bonne classification est amélioré et devient identique à celui de SSC avec N_t=20.

5 Conclusion

Une méthode, appelée « sac de phrases », sur la base du regroupement global de « sacs de mots visuels » locaux a été élaborée, en introduisant une relation spatiale entre les caractéristiques locales. Les résultats de la classification sur la base de données de texture UIUC montrent qu'il fournit de meilleures performances que SdM. Cette méthode peut atteindre les mêmes performances que SSC qui a besoin d'un vecteur de descripteur plus complexe. Elle n'utilise pas de modèles de processus ponctuels spatiaux, conduisant à un algorithme plus simple.

Table 1: Taux moyens et écarts types de bonne classification des approches proposées pour les textures UIUCtex en comparaison aux méthodes SdM et SSC.

N_t		1	5	10	15	20
SdM		67.2 ±2.8	76.4 ±2.1	81.1 ±1.5	86.4 ±1.2	91.3 ±1.1
SSC		75.6 ±1.3	91.7 ±0.9	94.3 ±0.8	96.5 ±0.5	97.3 ±0.3
Méthode proposée	quantification scalaire	65.9 ±1.3	81.5 ±1.2	89.6 ±0.9	94.4 ±0.6	96.8 ±0.8
	quantification vectorielle	72.0 ±2.1	87.9 ±1.4	94.4 ±0.9	96.8 ±0.9	96.8 ±0.5
	général	70.5 ±2.6	88.4 ±1.5	94.3 ±0.9	96.1 ±0.8	97.1 ±0.4

Références

[1] M. Cummins P. Newman, "Probabilistic localization and mapping in the space of appearance," *IJRR*, vol. 27, no. 6, pp. 647–665, 2008.

[2] S.Lazebnik C.Schmid J.Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, IEEE, Ed., 2006, pp. 2169–2178.

[3] J.Sivic, "Efficient visual search of videos cast as text retrieval," *IEEE Trans. on PAMI*, vol. 31, no. 4, pp. 591–
605, April 2009.

[4] G. Csurka C. Bray C. Dance L. Fan, "Visual categorization with bags of keypoints," in *ECCV*, 2004, pp 1–22.

[5] H.G. Nguyen R. Fablet J.M. Boucher, "Spatial statistics of visual keypoints for texture recognition," in *ECCV*, 2010, pp. 764–777.

[6] D.G. Lowe, "Distinctive image features from scale invariant keypoints," *International Journal of Computer Vision*, pp. 91–110, 2004.

[7] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Speeded Up Robust Features", ETH Zurich, Katholieke Universiteit Leuven

[8] J. Zhang M. Marszalek S. Lazebnik C. Schmid, "Local features and kernels for classification of texture and object categories: a comprehensive study," *IJCV*, vol. 73, no. 2, pp. 213–238, 2007.

[9] H.Ling S.Soatto, "Proximity distribution kernels for geometric context in category recognition," in *ICCV*, 2007, pp. 1–8.

[10] J. Liu M. Shah, "Scene modeling using co-clustering," in *ICCV*, 2007, pp. 1–7.

[11] M. Schlather, "On the second-order characteristics of marked point process," *Bernouilli*, vol. 7, pp. 99–117, 2001.

[12] D. Stoyan H. Stoyan, *Fractals, random shapes and point fields*, Wiley, 1994.

[13] H.G. Nguyen R. Fablet J.M. Boucher, "Visual textures as realizations of multivariate log-gaussian cox processes," in *CVPR*, 2011, pp. 2945–2952.

[14] H.G. Nguyen R. Fablet A. Ehrhold J.M. Boucher, "Keypoint-based analysis of sonar images: application to seabed recognition," *IEEE Trans.on GRS*, vol. 50, no. 4, pp. 1171–1184, April 2012.