HAL
open science

Mots visuels issus de graphes locaux multi-niveaux pour la reconnaissance d'objets

Svebor Karaman, Jenny Benois-Pineau, Rémi Mégret, Aurélie Bugeau

To cite this version:

Svebor Karaman, Jenny Benois-Pineau, Rémi Mégret, Aurélie Bugeau. Mots visuels issus de graphes locaux multi-niveaux pour la reconnaissance d'objets. RFIA 2012 (Reconnaissance des Formes et Intelligence Artificielle), Jan 2012, Lyon, France. pp.978-2-9539515-2-3. hal-00656516

HAL Id: hal-00656516

https://hal.science/hal-00656516v1

Submitted on 17 Jan 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mots visuels issus de graphes locaux multi-niveaux pour la reconnaissance d'objets

Svebor Karaman[1], Jenny Benois-Pineau[1], Rémi Mégret[2] et Aurélie Bugeau[1]

[1] Univ. Bordeaux, LaBRI, UMR 5800,
351, Cours de la Libération, 33405 Talence Cedex, France
{Svebor.Karaman, Jenny.Benois-Pineau, Aurelie.Bugeau} @labri.fr

[2] Univ. Bordeaux, IMS, UMR 5218,
351, Cours de la Libération, 33405 Talence Cedex, France
Remi.Megret@ims-bordeaux.fr

Résumé

Dans cet article, nous nous intéressons au problème ouvert à ce jour en indexation et recherche d'images à savoir la reconnaissance des objets. Depuis l'apparition de l'approche par des « Sacs-de-descripteurs » et ensuite des « Sac-de-mots », la « déstructuration » de la description des images en utilisant des ensembles non structurés de caractéristiques a été contrée par l'introduction de différents groupements de descripteurs locaux ou encore par l'introduction de la topologie. Ainsi la reconnaissance d'objets peut être vue à ce jour comme le retour, à un autre niveau et avec d'autres outils, à la démarche structurelle. Les caractéristiques structurelles que nous proposons pour la reconnaissance d'objets sont les graphes locaux multi-niveaux emboîtés établis sur des ensembles de points SURF avec la triangulation de Delaunay. Cette représentation conserve l'invariance aux transformations géométriques du plan-image inhérente aux descripteurs SIFT/SURF. Une approche de type sac de mots visuels est appliquée sur ces graphes, donnant naissance à une représentation de sacs de mots issus de graphes locaux. La construction des graphes locaux opère par niveaux successifs, depuis les graphes de Delaunay élémentaires – les points SURF isolés – en augmentant le nombre de nœuds à chaque couche. Pour chaque niveau de graphes un dictionnaire visuel distinct est établi. Les expériences entreprises sur les ensembles de données SIVAL et Caltech-101 indiquent que les graphes multi-niveaux ont des performances complémentaires sur chaque niveau et que leur combinaison améliore les performances par rapport à l'approche par sacs de mots visuels.

Mots Clef

Représentation par primitives visuelles, caractéristiques structurelles, Sac-de-Mots-Visuels, graphes de mots visuels, triangulation de Delaunay, noyau dépendant du contexte.

Abstract

In this paper we are interested on open problem in CBIR such as object recognition in images. Since the introduction of approaches as "Bag-of-Features" or "Bag-of-Visual Words", the de-structuring of image description was counterbalanced by introduction of different strategies for spatial grouping of features and introduction of topology. Hence the object recognition can be seen today as coming back to structural methods in pattern recognition. The structural features we propose for object recognition are nested multi-layered local graphs built upon sets of SURF feature points with Delaunay triangulation. This representation conserves the invariance to affine transformations of image plane which the initial SIFT/SURF features have. A Bag-of-Visual-Words framework is applied on these graphs, giving birth to a Bag-of-Graph-Words representation. The multi-layer nature of the descriptors consists in scaling from trivial Delaunay graphs - isolated feature points - by increasing the number of nodes layer by layer. For each layer of graphs its own visual dictionary is built. The experiments conducted on the SIVAL and Caltech-101 data sets reveal that the graph features at different layers exhibit complementary performances on the same content and that the combination of all existing layers, yields significant improvement of the object recognition performance compared to single level approaches.

Keywords

Feature representation, Structural features, Bag-of-Visual-Words, Graph Words, Delaunay triangulation, Context Dependent Kernel.

1 Introduction

La recherche d'objets d'intérêt dans des images et des vidéos est l'un des domaines de recherche les plus actifs. Une des techniques les plus populaires abordant cette tâche se fonde sur l'utilisation des caractéristiques locales, tels que les points d'intérêt, par exemple les descripteurs SIFT [1] or SURF [3]. Ces descripteurs locaux présentent l'avantage d'être robustes et discriminants. Les points d'intérêt SIFT sont détectés sur les minima et maxima locaux des images de Différences de Gaussiennes (DoG) calculées à différents échelles. Le descripteur SIFT est formé d'histogrammes d'orientation locaux calculés dans des voisinages du point d'intérêt. SURF reprend un principe similaire, mais accélère les calculs en approximant les mesures à l'aide d'ondelettes de Haar et d'images intégrales.

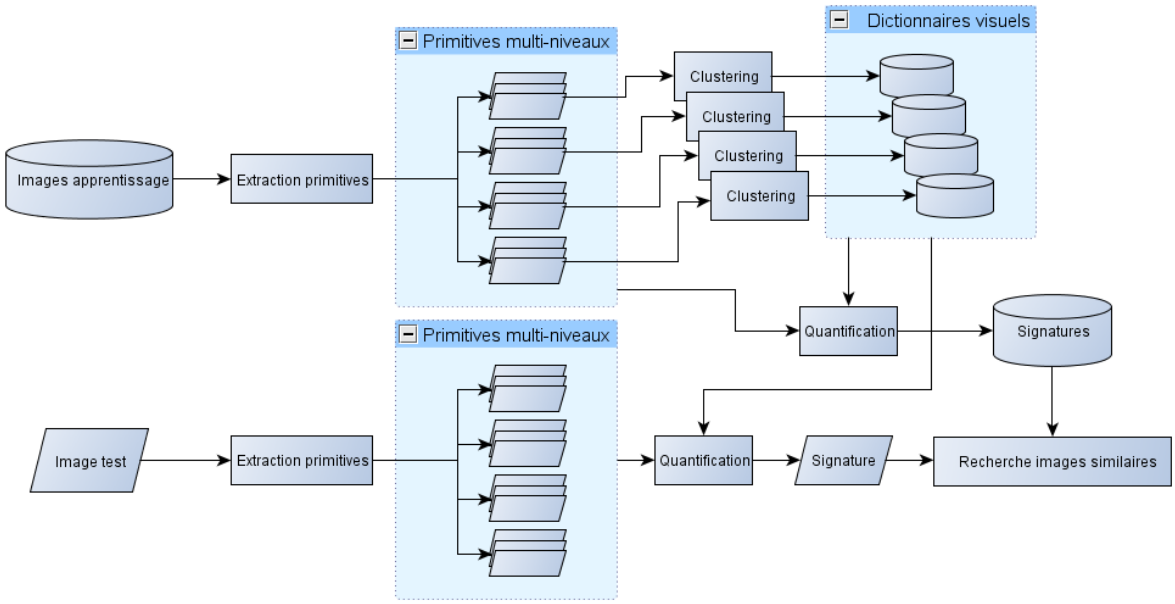


Fig. 1. Architecture générale de l'approche par sacs de graphes locaux multi-niveaux.

La démarche par sac de mots visuels (SdMV) [2] est une adaptation de l'approche par sac de mots (SdM) issue des techniques de recherche de texte. L'approche SdMV considère des descripteurs visuels locaux, de la même façon que l'approche SdM manipule des mots : un document (respectivement une image) est représenté par la distribution discrète des mots le composant (respectivement la distribution des mots visuels qui en sont extraits). Les descripteurs visuels sont quantifiés à l'aide d'un dictionnaire visuel obtenu par classification non supervisée et permettent d'associer à chaque image un histogramme des mots visuels qui en sont extraites. Le pouvoir sémantique d'un mot est beaucoup plus élevé que celui d'un mot visuel et également moins ambigu. Cependant, traiter la similarité entre images en tant que similarité des distributions de mots visuels fournit une information consolidée plus globale que la comparaison isolée de chaque primitive. Dans cette approche, l'approche SdMV ignore toute information spatiale sur la relation entre points d'intérêts détectés.

Afin de surmonter cette limitation du SdMV, plusieurs approches ont été développées au cours des dernières années. L'appariement par pyramide spatiale proposé dans [4] et utilisé dans [5] et [6] pour des applications à la reconnaissance d'objets et de scènes, calcule des histogrammes de mots visuels sur plusieurs zones spatiales de l'image au lieu de ne considérer qu'une zone unique. Cependant, cette approche n'est pas invariante aux transformations affines, perdant ainsi l'une des caractéristiques les plus importantes des primitives locales invariantes. Dans [7], une approche dite de phrases visuelles est proposée, qui regroupe les mots visuels par rapport à leur position spatiale dans l'image, sous la forme de séquences de mots. Chaque phrase visuelle est représentée par un histogramme contenant la distribution des mots visuels de la phrase. Dans ces travaux, l'idée commune est de construire des regroupements de primitives fondés sur l'organisation spatiale (soit de façon arbitraire, soit de façon adaptative), mais qui sont au final caractérisées par une distribution de mots sans information spatiale. L'approche que nous proposons consiste au contraire à intégrer l'information de la topologie spatiale explicitement au sein des primitives considérées et donc avant la quantification en mots visuels.

Dans cet article, les primitives considérées sont des graphes de Delaunay locaux. La motivation de construire de tels objets est que la triangulation de Delaunay est invariante vis-à-vis des transformations affines du plan (rotation, translation, échelle). Couplée à des points d'intérêt invariants tels que SIFT ou SURF, cette propriété d'invariance est ainsi conservée théoriquement pour des graphes construits à partir de telles primitives. Comme cette invariance n'est pas parfaite sur des données réelles, l'information structurelle de la topologie du graphe est combinée avec la robustesse apportée par l'approche SdMV.

Nous considérons ainsi un graphe de Delaunay local comme étant lui-même représenté par un mot visuel, et le plongeons dans le cadre de l'approche SdMV en construisant des dictionnaires visuels obtenus par classification non supervisée d'ensembles de tels graphes. On obtient ainsi des distributions de mots-graphes visuels sur lesquelles peuvent être appliquées les techniques de recherche classiques.

En faisant varier le nombre de nœuds des graphes locaux, nous pouvons de plus définir une approche multi-niveaux, où chaque niveau intègre plus de nœuds et donc fournit une attache spatiale plus forte au sein de la primitive visuelle. Nous considérons ainsi une approche emboitée

hiérarchique, où chaque graphe local est obtenu en rajoutant des nœuds à un graphe local de niveau inférieur. Cette méthode permet ainsi de couvrir à la fois le cas des graphes triviaux réduits à un seul nœud, correspondant à des points SURF isolés, ainsi que des graphes plus larges contenant une dizaine de nœuds. La méthode proposée est résumée sur la Figure **1**.

Cet article est organisé ainsi ; à la section 2 nous discutons le procédé permettant d'extraire les graphes et introduisons leur construction emboîtée. A la section 3, nous introduisons la mesure de dissimilarité employée pour comparer les graphes et la méthode pour obtenir les dictionnaires visuels par classification non-supervisée des graphes. Ces derniers sont présentés à la section 4. Les expérimentations utilisant ces nouveaux descripteurs visuels sont présentées à la section 5. La section 6 termine par les conclusions et perspectives de ce travail.

2 Construction des graphes locaux

Considérons un graphe G = ***(X, E)*** défini à partir d'un ensemble ***X*** de nœuds correspondant à des descripteurs visuels $x_{k,k=1,..,K}$ extraits de l'image, et connectés par un ensemble d'arêtes $E=\{e_{kl}\}_{,k=1,..,K,l=1,..,K}$, avec $e_{kl}=(x_k,x_l)$. Nous appelons un tel graphe un graphe local. Il est construit à partir de descripteurs visuels de type points d'intérêt voisins dans le plan-image. Par conséquent nous proposons de prendre en compte l'information spatiale du voisinage sous la forme de la structure du graphe. Pour définir de tels graphes deux questions doivent être abordées : i) le choix des descripteurs de base ***X*** et ii) la conception de la connectivité ***E.***

Pour définir l'ensemble ***X*** sur lequel des graphes sont établis nous utilisons un ensemble de points d'intérêt « germes ». Autour de chacun d'eux, d'autres points d'intérêt sont sélectionnés, afin de définir le support du graphe local. Les germes sélectionnées doivent être les plus stables, afin d'être détectées de façon répétable dans les différentes instances du même objet. Pour cela, nous utilisons des points SURF, fondés sur la détection de maxima de réponse du déterminant approché de la matrice Hessienne locale [3]. Ce critère fournit un indice de saillance, les plus hautes réponses correspondant aux points les plus saillants, supposés les plus répétables. Les points germes sont donc sélectionnés parmi les points SURF de plus haute réponse, en sélectionnant les N_{Seeds} points les plus saillants. Ceci définit l'ensemble des germes ***S*** :

$$S = \{s_1, \dots, s_{N_{seeds}}\} \quad (1)$$

Etant donné ***S***, notre but est d'ajouter l'information structurelle partielle de l'objet tout en maintenant le pouvoir distinctif des points d'intérêt SURF. Nous définirons donc les graphes à partir des germes et des points voisins.

Les ***k*** points SURF les plus proches d'un germe s_i définissent le voisinage P_i:

$$P_i = \{p_1, \dots, p_k\} \quad (2)$$

L'ensemble de tous les sommets utilisés pour établir le graphe G_i est noté X^{Gi}, contenant le germe et ses voisins :

$$X^{Gi} = \{x_1^{Gi}, \dots, x_k^{Gi}\} = P_i \cup \{s_i\} \quad (3)$$

Pour les arêtes, nous utilisons la triangulation de Delaunay invariante aux transformations affines du plan image préservant les angles: translation, rotation et changement d'échelle. En outre, le choix de la triangulation de Delaunay est également intéressant pour ses bonnes propriétés dans le suivi temporel des structures [8], ce qui devrait permettre d'étendre le travail à la vidéo.

La triangulation de Delaunay est calculée sur les points de X^{Gi}, établissant des triangles selon la contrainte de Delaunay. Une arête $e_{ij} = (x_i^{G_i}, x_j^{G_i})$ est définie entre deux sommets du graphe G_i si l'arête appartient à un triangle de la triangulation de Delaunay.

Au sein de l'approche multi-niveaux, chaque niveau ajoute des sommets par rapport au niveau précédent, et possède donc son propre ensemble de voisins autour de chaque germe s_i . La triangulation de Delaunay est lancée séparément à chaque niveau. Un niveau contiendra toujours les sommets des niveaux inférieurs, par conséquent nous appelons cette démarche « emboîtée » et l'illustrons dans la Figure **3**.

(a) Points SURF (b) Graphes à 3 plus proches voisins (c) Graphes à 6 plus proches voisins (d) Graphes à 9 plus proches voisins

Fig. 2. Points SURF et graphes-primitives détectés sur une image de la classe « fabricsoftenerbox » du corpus SIVAL.

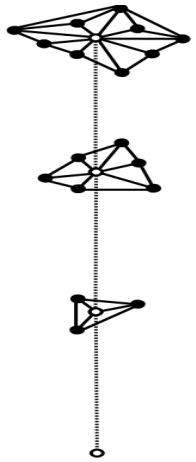


Fig. 3. Illustration de la construction emboitée des graphes locaux à 4 niveaux. De bas en haut: point SURF utilisé comme germe, puis graphes locaux à 3 voisins, à 6 voisins, puis à 9 voisins. Le germe est représenté par le nœud blanc, les voisins par les nœuds noirs.

Pour éviter un grand nombre de niveaux, le nombre de nœuds ajoutés à chaque niveau doit induire une modification importante de l'information structurelle. Pour établir une triangulation, au moins deux points doivent être rajoutés à un germe. Ajouter un nœud de plus peut engendrer trois triangles au lieu d'un seul, ayant pour conséquence une configuration locale plus complète. Ainsi, nous définissons quatre niveaux. Le premier niveau contient seulement des graphes singletons correspondant aux germes, et chaque niveau contient 3 nœuds de plus que le niveau précédent, jusqu'au niveau le plus haut, qui contient des graphes construits sur le germe et ses 9 voisins les plus proches, comme illustré à la Figure **2**.

3 Comparaison des graphes locaux

Afin d'intégrer ces graphes locaux dans l'approche par sacs de mots visuels, nous utilisons un dictionnaire fondé sur une méthode de regroupement non supervisé. Cette section décrit la mesure de dissimilarité permettant de comparer les graphes. Elle sera exploitée à la section suivante pour la construction du dictionnaire.

Les graphes étant attribués, les nœuds peuvent être comparés sur la base de leur seule apparence. Bien qu'il soit possible de ne prendre en compte que les similarités entre nœuds ou bien la topologie du graphe séparément, la combinaison des deux fournit une comparaison entre graphes plus complète. Nous nous fonderons pour cela sur les noyaux dépendants du contexte (*Context Dependant Kernel*, *CDK*) présentés dans [9]. La définition du CDK dépend de deux matrices : la matrice ***D*** contient les distances visuelles entre nœuds, la matrice ***T*** contient la topologie des graphes considérés.

Considérons deux graphes locaux ***A*** et ***B*** possédant respectivement ***m*** et ***n*** nœuds, et notons ***C*** leur union:

$$C = A \oplus B$$
$$\text{avec} \begin{cases} x_i^C = x_i^A & \text{for} \quad i \in [1..m] = I_A \\ x_i^C = x_{i-m}^B & \text{for} \quad i \in [m+1..m+n] = I_B \end{cases} \quad (4)$$

La matrice de correspondances ***D*** est carrée, de taille *(m+n)×(m+n)* et contient les distances euclidiennes entre les vecteurs de description SURF entre nœuds de ***C***:

$$\mathrm{D} = (\mathrm{d_{ij}})_{\mathrm{ij}}$$
$$\text{où } \mathrm{d_{ij}} = \left\| \mathrm{x_i^C} - \mathrm{x_j^C} \right\|_2 \quad (5)$$

La matrice de topologie ***T*** est carrée, de taille *(m+n)×(m+n)* et définit la matrice d'adjacence/connectivité entre nœuds de ***C***. Dans notre travail, nous avons utilisé une connectivité binaire : $\boldsymbol{T_{ij}}$ vaut 1 si les noeuds $\boldsymbol{x_i^C}$ et $\boldsymbol{x_j^C}$ sont connectés, 0 sinon. Comme les connections ne se trouvent qu'au sein d'un même graphe ***A*** ou ***B***, la matrice ***T*** est diagonale par blocs. Nous pouvons définir deux sous-matrices $\boldsymbol{T_{AA}}$ et $\boldsymbol{T_{BB}}$ correspondant à la topologie au sein de ***A*** et ***B*** respectivement, alors que les deux blocs hors diagonale $\boldsymbol{T_{AB}}$ et $\boldsymbol{T_{BA}}$ sont nuls : les graphes ***A*** et ***B*** ne sont pas connectés topologiquement.

$$T = (T_{ij})_{ij}$$
$$\text{où } T_{ij} = \begin{cases} 1 \text{ si l'arête } (x_i^C, x_j^C) \text{ appartient à } \boldsymbol{A} \text{ ou } \boldsymbol{B} \\ 0 \text{ sinon} \end{cases} \quad (6)$$

Le noyau CDK, noté par la suite ***K*** est calculé par un procédé itératif consistant à propager la similarité visuelle entre nœuds à l'aide de la topologie des graphes.

$$K^{(0)} = \frac{\exp\left(-\frac{D}{\beta}\right)}{\left\| \exp\left(-\frac{D}{\beta}\right) \right\|_1} \quad , \quad K^{(t)} = \frac{G(K^{(t-1)})}{\left\| G(K^{(t-1)}) \right\|_1} \quad (7)$$
$$G(K) = \exp\left(-\frac{D}{\beta} + \frac{\alpha}{\beta} T K^{(t-1)} T\right)$$

Dans les équations précédentes, *exp* représente l'exponentiation de chaque coefficient et $\|\boldsymbol{M}\|_1 = \Sigma_{ij} |\boldsymbol{M}_{ij}|$ représente la norme matricielle L_1. Les deux paramètres $\boldsymbol{\beta}$ et $\boldsymbol{\alpha}$ peuvent être vus respectivement comme une pondération de la distance visuelle et de la propagation topologique. La diffusion lors du processus itératif doit être appliquée à l'union des deux graphes pour saisir de façon unifiée les similarités visuelles entres points SURF et les liens topologiques. De façon analogue à la décomposition en blocs de ***T***, nous pouvons décomposer ***K*** en 4 blocs. La sous-matrice $\boldsymbol{K_{AB}^{(t)}}$ représente la force des liens inter-graphes entre les nœuds de ***A*** et ***B*** une fois que la topologie a été prise en compte. Nous pouvons les utiliser pour définir une mesure de similarité et de dissimilarité combinant la dissimilarité visuelle et la cohérence topologique sous la forme:

$$s(A,B) = \sum_{\{i \in I_A, j \in I_B\}} K_{ij}^{(t)} \in [0,1]$$
$$\rho(A,B) = s(A,A) + s(B,B) - 2s(A,B) \in [0,1] \quad (8)$$

Cette mesure de dissimilarité sera calculée indépendamment pour chaque niveau de graphes.

Cependant, au premier niveau, aucune topologie n'est définie, la distance euclidienne entre descripteurs SURF sera donc utilisée directement.

4 Construction du dictionnaire de graphes

L'approche usuelle de l'état de l'art pour calculer un dictionnaire est d'appliquer la classification non supervisée par K-moyennes sur un échantillonnage suffisamment représentatif des descripteurs vectoriels issus des objets considérés [2] avec un nombre assez élevé de classes (souvent plusieurs milliers). Un mot est alors associé au centre de chaque classe, ce qui permet de quantifier un nouveau descripteur en l'associant au centre de classe le plus proche.
Cette approche n'est pas adaptée dans notre cas. En effet, l'algorithme des K-moyennes repose sur la mise à jour itérative des centres de classes, en interpolant les données, et le calcul rapide des affectations de chaque échantillon à une classe repose sur des structures d'indexation. Ces deux opérations nécessitent une structure d'espace vectoriel, qui n'est pas disponible dans le cas des graphes locaux. Pour ces raisons, nous présentons maintenant la méthode de classification utilisée, reposant sur une approche agglomérative hiérarchique en deux passes.

Méthode de regroupement non-supervisé

L'approche en deux passes a été proposée dans [10] pour la quantification de bases de données de très grandes dimensions, afin de réduire les coûts calculatoires. Nous reprenons le principe de la façon suivante. La première passe applique un regroupement non-supervisé des graphes locaux par agglomération hiérarchique sur chaque catégorie de la base séparément, afin de diminuer la quantité de données à traiter. La seconde passe ne traite qu'un représentant par classe générée par la première passe. Dans notre cas, la moyenne de graphes n'étant pas définie, nous utilisons le représentant médian pour chaque classe, défini par

$$median = \operatorname{argmin}_{g \in V} \sum_{i=1}^{m} d(g_i, g) \qquad (9)$$

où V représente l'ensemble des m graphes dans la classe, g_i un membre de la classe, g le graphe choisi comme graphe médian candidat, et $d(\cdot,\cdot)$ la mesure de dissimilarité entre graphes définie précédemment.
Chaque niveau est traité indépendamment des autres. Pour la première passe, les dissimilarités entre tous les graphes d'un même niveau et extraits dans toutes les images d'un même objet sont calculées, puis donnent lieu au regroupement agglomératif hiérarchique. Pour la deuxième passe, les dissimilarités entre les graphes médians issus de la première passe sont calculés en prenant en compte tous les objets simultanément. Nous obtenons ainsi les dictionnaires visuels associés à chaque niveau, correspondant à des graphes de 1, 4, 7 et 10 nœuds (un germe associé à ces 3, 6 ou 9 plus proches voisins au niveau correspondant).

Signatures visuelles

La représentation usuelle d'une image dans le paradigme « sac de mots visuels » (SdMV) consiste à associer à chaque descripteur local extrait de l'image le mot visuel le plus proche par le dictionnaire visuel, puis à calculer l'histogramme des mots visuels ainsi obtenus. Nous utilisons cette représentation en associant toujours un mot à un descripteur, puis en normalisant l'histogramme de façon à obtenir une distribution sommant à 1. C'est cette distribution qui est finalement utilisée comme la signature visuelle de l'image.
Une fois les signatures calculées, la distance entre deux images est définie comme la distance entre leurs signatures. Dans les expériences préliminaires, la distance de Hamming, la distance Euclidienne et la distance L_1 ont été testées. La distance L1 fournissant les meilleurs résultats, les expérimentations finales présentées ici utilisent cette métrique.

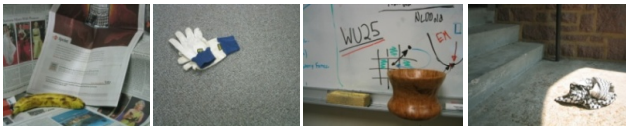
5 Expérimentations

Bases d'images utilisées

Les expérimentations ont été effectuées à partir de deux bases d'images disponibles publiquement.
La première base, SIVAL (Spatially Independent, Variable Area, and Lighting) [11] contient 25 objets, chacun présent dans 60 images prises dans 10 environnements différents, et sous différentes postures, pour un total de 1500 images. Cette base est assez difficile, les objets étant photographiés dans des conditions lumineuses et de points de vue variés. Elle a été choisie afin de représenter la perspective de ce travail à plus long terme au niveau de la reconnaissance d'objets de la vie quotidienne pouvant apparaître à différents endroits d'une maison.
La deuxième base, Caltech-101 [12], est composée de 101 catégories d'objets, couvrant notamment plusieurs types d'animaux, de plantes ou d'objets variés. Ces deux bases sont illustrées sur la Figure **4** par quelques exemples d'images typiques.

Protocole d'évaluation

Le protocole d'évaluation sépare un ensemble



(a) base SIVAL

(b) base Caltech-101

Fig. 4. Exemples d'images issus des corpus utilisés.

d'apprentissage et un ensemble de test par sélection aléatoire. Sur chaque base, 30 images de chaque catégorie sont choisies pour l'apprentissage, à la fois pour construire le dictionnaire visuel et pour la tâche de recherche. Les tailles des catégories de Caltech-101 sont très variables, allant de plusieurs centaines d'image à seulement quelques images au-delà de 30. Les images de test sont choisies aléatoirement dans le reste des images non utilisées pour l'apprentissage, jusqu'à 50 maximum par catégorie.

L'objectif de ce travail étant axé sur la reconnaissance, et pas encore sur la détection, nous utilisons l'information de la boîte englobante de chaque objet. Les points d'intérêt SURF avec un descripteur à 64 dimensions sont extraits au sein des boîtes englobantes. Le nombre de points germes pour la construction des graphes locaux est fixé à 300 par objet. Le deuxième niveau est constitué des graphes contenant une source et ses 3 plus proches voisins ; le troisième niveau utilise 6 plus proches voisins, et le quatrième utilise 9 plus proches voisins.

Pour les paramètres de noyau dépendant du contexte, α est fixé à 0.0001, β à 0.1 (afin de s'assurer que **K** est défini positif) et 2 itérations sont appliquées, ce qui est justifié par la convergence rapide du processus [9].

Pour la construction du dictionnaire, la première passe calcule 500 classes pour chaque objet. La taille du dictionnaire final varie entre 50 et 5000 classes. A chaque niveau est associé un dictionnaire calculé séparément. Les signatures obtenues à partir des graphes locaux sont notées « GW3NN », « GW6NN » et « GW9NN ».

Nous comparons notre approche avec l'approche par sac de mots standard. Pour ce faire, l'ensemble des points SURF extraits est utilisé pour construire le dictionnaire, à l'aide de l'algorithme des K-moyennes appliqué aux descripteurs associés. Chaque mot visuel correspond à un centre de classe. Cette approche est notée « ClassicalBoW » par la suite.

Les graphes locaux ne sont pas bâtis sur l'ensemble des points SURF disponibles, étant centrés autour des points germes. Afin de fournir une base de comparaison supplémentaire, nous calculons également les signatures sac de mots à partir des sous-ensembles de points SURF qui ont effectivement été utilisés pour former les graphes locaux, en ignorant dans ce cas la structure topologique. Ces configurations sont notées « SURF3NN », « SURF6NN » et « SURF9NN » par la suite, correspondant aux niveaux à 3, 6 et 9 voisins. Dans ce cas, le dictionnaire est construit avec la même procédure en deux passes que pour les graphes locaux.

Pour chaque image requête et chaque image référence, la signature est calculée à partir des points SURF isolés, ainsi que pour chaque niveau de la représentation par graphes locaux. Nous avons étudié différentes combinaisons d'approches par fusion précoce, en concaténant les signatures obtenues. Cette concaténation est faite entre la signature SURF*x*NN et les niveaux correspondants des signatures GW*y*NN, donnant lieu aux approches « SURF3GW3 », « SURF6GW3GW6 » et « SURF9ALLGW = SURF9GW3GW6GW9». La signature « ClassicalBoW » utilisée comme base de comparaison est construite sur l'ensemble des points SURF dans les boîtes englobantes des objets. Enfin, la distance L_1 entre histogrammes est utilisée pour comparer les signatures obtenues.

La performance est évaluée par la mesure Mean Average Precision (MAP) : pour chaque image requête, l'ensemble des images de référence sont triées par ordre croissant de dissimilarité ; la précision moyenne est alors calculée en considérant cet ordre. Le critère MAP est la moyenne de ces valeurs pour toutes les images d'un objet de test. La performance globale sur la base est calculée comme la

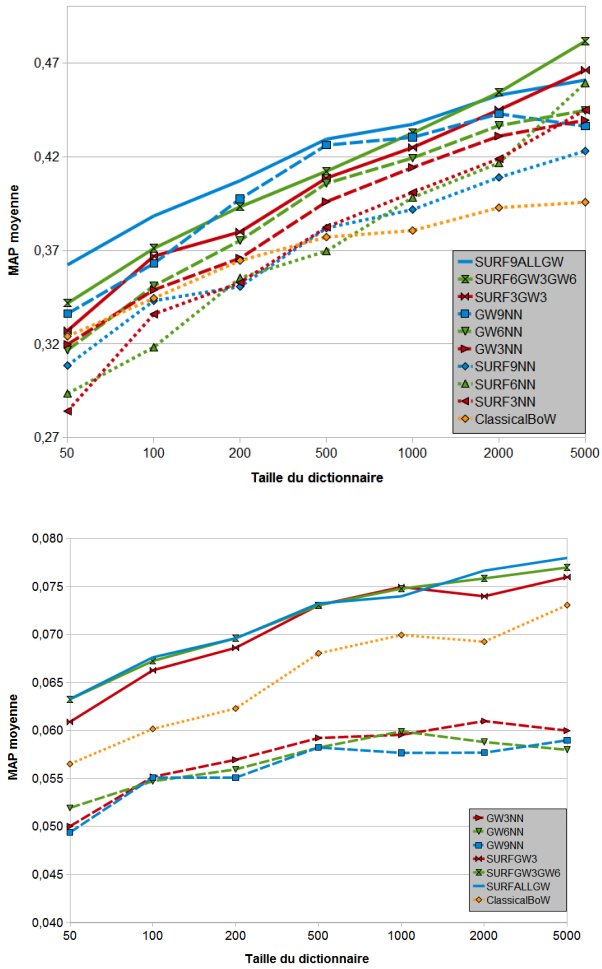


Fig. 5. Moyenne du critère *MAP* sur l'ensemble de la base SIVAL (en haut) et Caltech-101 (en bas). Les noms des méthodes sont définis dans le texte. Les signatures fondées sur différentes caractéristiques sont représentées graphiquement, par des pointillés pour les points SURF isolés, tirets pour les graphes locaux à un seul niveau, courbe pleine pour les graphes locaux multi-niveaux.

moyenne des critères MAP pour chaque objet, leur donnant ainsi le même poids et évitant donc le biais induit par les objets sous ou sur-représentés.

Résultats

Nous considérons en premier lieu la différence entre les approches, représentée à la Figure **5**. Sur la base SIVAL, les descripteurs SURF isolés (SURF*x*NN) ont les moins bonnes performances que l'utilisation séparée des niveaux de graphes locaux (GW*x*NN).

L'approche de classification non-supervisée que nous utilisons pour la construction du dictionnaire est, comparée à l'approche SdMV usuelle, moins performante pour des petits dictionnaires, mais fournit de meilleurs résultats quand la taille de ceux-ci augmente dans le cas de SIVAL. Dans ce cas, chaque niveau de graphe fournit de meilleurs résultats pour les mots graphes, que pour les mots à base de points SURF isolés. Dans tous les cas, la combinaison multi-niveaux (SURF*x*GW*y*…) fournit les meilleures performances.

Dans le cas de Caltech-101, l'approche en deux passes montre ici clairement une performance plus faible que l'utilisation d'un algorithme de regroupement pouvant prendre en compte l'ensemble des données.

Les performances moyennes cachent cependant quelques différences de comportement, que nous illustrons ici sur deux exemples. Sur la base SIVAL, pour l'objet « banana », les points SURF isolés donnent de meilleurs résultats que l'approche basée graphes locaux (voir Figure **7**). Sur la base Caltech-101, pour la catégorie « Faces », l'approche par graphes locaux est au contraire la meilleure (voir Figure **6**). Ces comportements complémentaires conduisent naturellement à utiliser une combinaison des signatures obtenues à chaque niveau, qui fournit des résultats aussi bons ou meilleurs que chacun pris séparément.

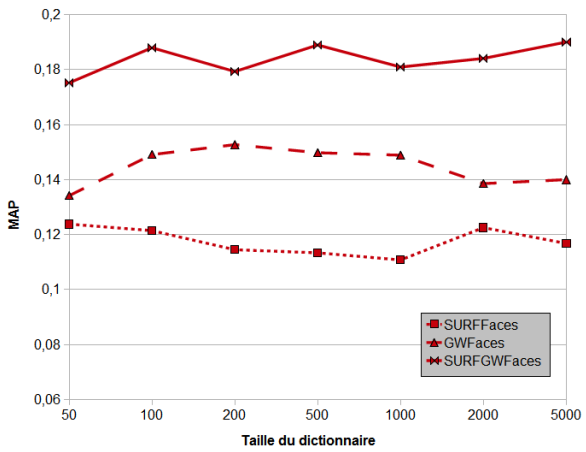


Fig. 6. Critère MAP pour la classe "Faces" de la base Caltech-101: l'approche basée points (pointillés) est moins bonne que l'approche basée graphes locaux (tirets). L'approche multi-niveaux (courbe pleine) fournit des résultats meilleurs que chacun des deux séparément.

Les performances de l'approche par combinaison des signatures sont bien meilleures que l'approche SdMV standard, voir Figure **5**. L'amélioration maximale par rapport à l'approche SdMV standard pour une taille de dictionnaire donnée est de 21% sur la base SIVAL et de 12,4% sur la base Caltech. L'amélioration moyenne pour les trois combinaisons étudiées et pour toutes les tailles de dictionnaires est de 14,3% sur SIVAL et 8,7% sur Caltech. L'approche par décomposition en quad-tree introduite dans [13] donne une amélioration de 19,2% par rapport à une approche SdMV standard et la méthode d'appariement par pyramide spatiale proposé dans [4] donne une amélioration de 8,3%. Ces approches introduisent l'information spatiale en calculant des distributions locales de mots visuels et peuvent donc être utilisées conjointement avec la méthode proposée.

6 Conclusion et perspectives

Dans cet article, nous avons présenté de nouvelles primitives fondées sur la construction de graphes locaux dont les nœuds sont des points d'intérêt SURF et qui expriment les relations spatiales entre ces points. La démarche multi-niveaux exploite une croissance de voisinage progressive afin de capturer l'information discriminante au niveau pertinent pour chaque type d'objets. Cette approche améliore nettement les résultats de reconnaissance, alors que l'utilisation de chaque niveau séparément apporte des améliorations plus légères. De plus, cette démarche introduit l'information spatiale au sein des graphes locaux et est donc compatible et complémentaire avec d'autres améliorations récentes de l'approche par sacs de mots visuels visant à prendre la géométrie en considération, comme les méthodes à base de pyramide spatiale [4].

En perspective à ce travail se trouve l'application de la

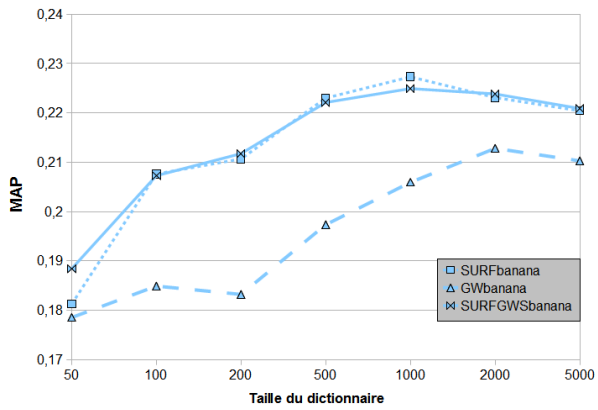


Fig. 7. Critère MAP pour l'objet "banana" de la base SIVAL: l'approche basée points (pointillés) est meilleure que l'approche basée graphes locaux (tirets). Les meilleures performances sont capturées par l'approche multi-niveaux (courbe pleine).

méthode à la reconnaissance des objets dans des vidéos. La démarche pourrait être améliorée en raffinant les phases de construction et de comparaison des graphes locaux. Par exemple, la sélection des germes pourrait être faite par une méthode adaptative et la matrice de topologie pourrait être définie avec un indice de connectivité continu. Afin d'être efficace en traitant un grand nombre d'images, c'est-à-dire dans des vidéos, une procédure de plongement ou de hachage de ces primitives pourrait être appliquée pour obtenir une représentation numérique utilisable dans les structures d'indexation qui accéléreraient le procédé de reconnaissance.

7 Remerciements

Ce travail est partiellement supporté par un financement de l'ANR (Agence Nationale de la Recherche) avec la référence ANR-09-BLAN-0165-02, dans le cadre du projet IMMED.

8 Références

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60(2), pp. 91-110, 2004.

[2] J. Sivic et A. Zisserman, "Video google: a text retrieval approach to object matching in videos," In *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV'2003)*, vol. 2, pp. 1470-1477, 2003.

[3] H. Bay, A. Ess, T. Tuytelaars et L. V. Gool, "Surf:Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110(3), pp. 346-359, 2008.

[4] K. Grauman et T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," In *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV'*2005), vol. 2, pp. 1458-1465, 2005.

[5] S. Lazebnik, C. Schmid, et J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,". In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'2006)*, vol. 2, pp. 2169-2178, 2006.

[6] A. Bosch, A. Zisserman, et X. Munoz, "Representing shape with a spatial pyramid kernel,". In *Proceedings of the 6th ACM international Conference on Image and Video Retrieval (CIVR'07)*, pp. 401-408, New York, NY, USA, 2007.

[7] R. Albatal, P. Mulhem, Y. Chiaramella, "Visual Phrases for automatic images annotation," *IEEE International Workshop on Content-Based Multimedia Indexing (CBMI'2010)*, pp. 1-6, Grenoble, France, 2010.

[8] A. Mahboubi, J. Benois-Pineau et D. Barba, "Joint tracking of polygonal and triangulated meshes of objects in moving sequences with time varying content," In *Proceedings of the IEEE International Conference on Image Processing (ICIP'2001)*, vol. 2, pp. 403-406, 2001.

[9] H. Sahbi, J.-Y. Audibert, J. Rabarisoa et R. Keriven, "Robust matching and recognition using context-dependent kernels," In *Proceedings of the 25th International Conference on Machine Learning (ICML'2008)*, pp. 856-863, 2008.

[10] P. H. Gosselin, M. Cord, et S. Philipp-Foliguet, "Combining visual dictionary, kernel-based similarity and learning strategy for image category retrieval," *Computer Vision and Image Understanding*, vol. 110(3), pp. 403-417, 2008.

[11] SIVAL Data set: http://accio.cse.wustl.edu/sg-accio/SIVAL.html

[12] L. Fei-Fei, R. Fergus et P. Perona. "One-Shot learning of object categories". IEEE Trans. *Pattern Recognition and Machine Intelligence*, vol. 28(4). pp. 594-611, 2006.

[13] V. Ramanathan, S. Mishra et P. Mitra, "Quadtree decomposition based extended vector space model for image retrieval". *IEEE Workshop on Applications of Computer Vision (WACV'2011)*, pp. 139-144, 2011.