

---

# Modélisation de l'extraction des descripteurs visuels - Intégration de relations topologiques

**Rami Albatal, Philippe Mulhem, Yves Chiaramella**

*Laboratoire d'informatique de Grenoble (LIG)  
Equipe MRIM - Bâtiment B  
Domaine Universitaire  
385 rue de la Bibliothèque  
38400 Saint Martin d'Hères  
{Rami.Albatal}{Philippe.Mulhem}{Yves.Chiaramella}@imag.fr*

---

*RÉSUMÉ. Malgré son rôle majeur dans l'annotation automatique, le processus d'extraction des descripteurs visuels n'est pas encore explicitement modélisé, et la contribution de chacune de ces étapes sur la qualité de l'annotation n'est pas suffisamment étudiée. Dans cet article, nous proposons un modèle (appelé phrasage) pour l'extraction des descripteurs visuels. Afin de construire des descripteurs plus riches, nous définissons, à partir de ce modèle, la prise en compte de relations topologiques entre régions d'intérêt via une nouvelle technique de regroupement. Des expérimentations sur le corpus VOC2009 montrent que notre approche améliore significativement les résultats de l'annotation.*

*ABSTRACT. Despite its major role in the automatic annotation, the visual descriptors extraction process is not explicitly modeled and the contribution of each of its steps on the quality of automatic annotation is not sufficiently studied. In this article, we first propose a model for visual descriptors extraction. Using this model, we propose the inclusion of relationships by applying a topological grouping technique of regions of interest; leading to the definition of richer descriptors. Experiments on VOC2009 corpus show that our approach significantly improves the automatic annotation results.*

*MOTS-CLÉS : Annotation automatique d'image, régions d'intérêt, regroupement topologique, sac de mots visuels.*

*KEYWORDS: Automatic image annotation, regions of interest, topological grouping, bag of visual words.*

---

## 1. Introduction et problématique

Rechercher des images selon leur contenu sémantique implique actuellement qu'elles soient associées à des annotations textuelles. Cette tâche se heurte au problème du manque ou de l'absence des annotations décrivant le contenu sémantique des images disponibles dans de nombreuses collections (y compris les images sur le web).

L'annotation automatique est une solution pour faire face au problème de description sémantique des images. Un tel processus automatique, une fois bien achevé, permet de réduire le coût et le temps d'annotation des grandes collections d'images riches en contenu sémantique. L'annotation automatique consiste à analyser le contenu visuel (couleurs, textures, formes) des images (ou des objets visibles dans les images) afin de le transformer en informations symboliques/textuelles. La difficulté majeure de cette analyse est la dépendance du contenu visuel des objets (et par conséquent leurs images) à de nombreux facteurs tels que l'instance considérée, les conditions de prise de vue et le contexte d'occurrence des objets. Ces facteurs amènent à des variations visuelles qui compliquent l'analyse du contenu visuel et, par conséquent, toute annotation automatique basée sur cette analyse. La figure 1 montre des variations visuelles de la classe d'objets *avion*. Nous remarquons dans les images que :

- les avions n'ont pas tous les mêmes couleurs.
- les avions n'ont pas la même échelle : il y a des avions grands (proches), d'autres qui sont petits (loin) ;
- les avions n'ont pas les mêmes orientations : les objets sont 3D, et les prises de vues reflètent cet état de fait ;
- les avions n'ont pas tous la même position dans les images prises ;
- la luminosité des images diffère : certaines images sont plus claires que d'autres.



**Figure 1.** Images exemples de la classe d'objets avion.

Afin de minimiser les effets négatifs de ces facteurs, les méthodes récentes d'annotation automatique essayent de trouver des zones dans les images qui contiennent des informations visuelles robustes contre de telles variations. En particulier, des techniques d'extraction et de description des régions d'intérêt sont appliquées avec succès pour détecter ces zones. Si une région d'intérêt doit, si possible, être robuste à plusieurs variations visuelles locales, tels que le changement d'échelle et d'éclairage, ainsi que la rotation et la translation, elle n'est pas capable individuellement de décrire et distinguer des objets visuels ou des images (Zheng *et al.*, 2008). Des méthodes fournissant un très bon compromis entre la complexité de calcul et la qualité des résultats sont des méthodes basées sur le modèle de sac de mots visuels introduit par (Sivic *et al.*, 2003). Ces méthodes regroupent les régions d'intérêt de l'image (ou des parties de l'image), et construisent un descripteur de chaque groupe créé (sous forme d'histogramme), puis effectuent un apprentissage supervisé ou non supervisé sur l'ensemble des descripteurs. Malgré les bonnes performances des méthodes à base de sacs de mots visuels, il n'existe pas d'argument clair sur le choix de regroupement appliqué sur les régions, outre l'intuition que "le regroupement de plusieurs régions d'intérêt cumule leurs informations visuelles, ce qui conduit à la construction de descripteurs capables de décrire et distinguer le contenu sémantique des images". De nombreux travaux dans ce contexte ont montré d'une façon heuristique la validité de cette intuition, mais ils ont tous au moins l'un des inconvénients suivant :

- 1) Pas d'explication explicite sur le choix des groupes de régions d'intérêt constituant les sacs de mots visuels ;
- 2) Pas de prise en compte de relations topologiques entre les régions d'intérêt constituant les sacs de mots visuels.

Afin de combler ces lacunes, nous proposons dans cet article un modèle, appelé "*modèle de phrasage*", pour l'extraction de descripteurs visuels de bas niveau. Nous appelons les descripteurs générés à partir de ce modèle des "*Phrases Visuelles*". Ces phrases, dans un mode simplifié sont des sacs de mots visuels classiques, mais elles ont pour objectif de représenter des descripteurs plus sophistiqués. Le but du phrasage est de standardiser la présentation du processus d'extraction de descripteurs en le décomposant en trois étapes successives. Le regroupement des régions d'images constitue la deuxième étape du phrasage. Identifier explicitement ces aspects permet de mieux l'explicitier, le contrôler, le justifier et le comparer à d'autres méthodes de regroupement afin de comprendre leurs impacts sur l'annotation automatique. En nous basant sur notre modèle de phrasage, nous proposons un regroupement tenant compte d'une relation topologique entre les régions d'intérêt. L'utilisation d'un tel regroupement dans une méthode d'annotation automatique améliore significativement les résultats par rapport à une méthode d'annotation basée sur un regroupement classique en sac de mots visuels.

Le plan de cet article est structuré comme suit : la section 2 présente des travaux de l'état de l'art basées sur les régions d'intérêt, particulièrement les méthodes inspirées du modèle de sac de mots visuels, ainsi que des méthodes proposant de tenir compte des relations topologiques et spatiales entre les régions d'intérêt dans le domaine de la

recherche d'images par le contenu. La section 3 présente le modèle de phrasage. Dans la section 4, nous instancions notre modèle de phrasage afin de générer deux types de Phrases Visuelles : la première est une représentation classique en sac de mots visuels, la seconde prend en compte une relation topologique entre régions d'intérêt. Dans la section 5 les deux instances sont évaluées sur la collection VOC2009 <sup>1</sup>. Enfin, nous concluons en section 6.

## 2. État de l'art

### 2.1. Régions d'intérêt et modèles de Sac de Mots Visuels

Bres et ses collègues (Bres *et al.*, 1999) indiquent que l'idée sous-jacente des régions d'intérêt est que lorsque quelqu'un regarde une image il suffit de regarder ces points pour identifier les objets existants ; même si on n'a pas assez de temps pour totalement visualiser l'image, on identifie des caractéristiques visuelles importantes de l'image grâce à ces points. David Lowe (Lowe, 1999) a par ailleurs cité des recherches en neurosciences qui ont montré que la reconnaissance des objets chez les primates fait usage des caractéristiques d'éléments de complexité intermédiaire qui sont largement invariants aux changements d'échelle, de localisation et l'éclairage (Tanaka, 1997) (Perrett *et al.*, 1998). Ces travaux ont conduit à la proposition de plusieurs détecteurs de régions d'intérêt. Les détecteurs sont capables de localiser des régions à partir desquelles des descripteurs visuels robustes aux variations visuelles sont construits. Une description des régions d'intérêt qui a montré son efficacité dans plusieurs domaines est la description par mot visuel. Cette description passe par trois étapes :

1) extraire pour chaque région d'intérêt un descripteur basique, tel que SIFT (Lowe, 1999), SURF (Bay *et al.*, 2006), GLOH (Winder *et al.*, 2007), DAISY (Tola *et al.*, 2008), etc. Ces descripteurs sont en général des vecteurs d'une dimensionnalité élevée ;

2) quantifier l'espace du descripteur basique en appliquant une technique de clustering (par exemple les K-moyennes) ; chaque centroïde créé par le clustering est dénoté par un identifiant qui représente un « mot » visuel, l'ensemble des mots visuels est appelé vocabulaire visuel (codebook) ;

3) décrire chaque région d'intérêt avec l'identifiant du plus proche centroïde de son descripteur basique. On obtient alors une description beaucoup plus compacte que celle par descripteur basique.

Cette description a été appliquée avec succès dans la recherche d'image par le contenu, ainsi que l'annotation et la classification d'image. Par contre, comme nous l'avons indiqué dans l'introduction, (Zheng *et al.*, 2008) montre qu'une région d'intérêt individuelle n'est pas capable de décrire et distinguer les différents objets ou les images. Nous notons ici que les mots visuels sont beaucoup plus ambigus que les mots

---

1. <http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2009/>



textuels, (Larlus, 2008) a montré qu'*Il est impossible de créer des mots qui sont toujours observés sur la même partie d'un objet et jamais ailleurs*. C'est la raison pour laquelle les approches d'annotation essayent de retrouver des descriptions plus discriminantes que des simples mots visuels. La tendance actuelle est de prendre en compte simultanément plusieurs mots visuels décrivant plusieurs régions d'intérêt. C'est surtout le cas du modèle de sac de mots visuels introduit par (Sivic *et al.*, 2003). Dans cette représentation classique du contenu visuel, on perd toute information relative à l'organisation topologique des régions d'intérêt dans l'image, sachant que cette information peut être importante pour décrire et différencier des objets ou des images. Afin de surmonter ce problème, une extension de ce modèle appelée "pyramide spatiale"<sup>2</sup> est proposée par (Lazebnik *et al.*, 2006) : il s'agit de décomposer l'image en  $2^l \times 2^l$  zones rectangulaires à différentes échelles ( $l = 0, 1, 2$ ) ; ensuite un sac de mots visuels est construit pour chaque zone ; enfin un seul histogramme est construit par la concaténation de tous les sacs de mots visuels de l'image. En cas de  $l = 0$ , la pyramide spatiale se réduit à une représentation classique en un sac de mots visuels. La figure 2 montre un exemple de l'extraction d'une représentation à base d'une pyramide spatiale à trois niveaux, le niveau 0 correspond à la représentation classique. Malgré le succès des méthodes à base de pyramides spatiales, les décompositions appliquées sont prédéfinies, et ne s'adaptent donc pas aux variations visuelles, même si des résultats expérimentaux obtenus sont bons.

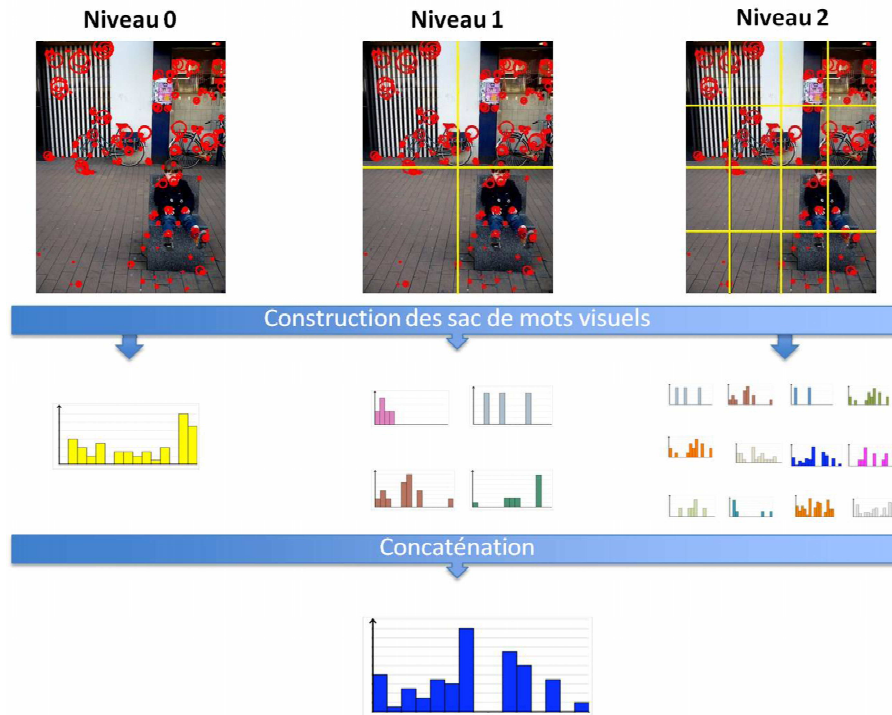
D'autres travaux se proposent de dépasser les limites de l'étape 3) décrite plus haut (assignation d'un cluster par région d'intérêt) : certains proposent d'assigner une région (ou toutes les régions d'une image) à plusieurs clusters (Tuytelaars *et al.*, 2007) ou de la représenter par un mélange de gaussiennes (Perronnin *et al.*, 2007). La philosophie de ces approches restent cependant dans un cadre où toutes les régions sont considérées en même temps pour décrire les images, ce qui est similaire au sacs de mots vus plus haut. Ces approches ne sont pas considérées dans la suite car nous choisissons de nous concentrer sur des approches peu complexes.

## 2.2. Exploitation des relations topologiques entre les régions d'intérêt

Dans le domaine de recherche d'image par le contenu, certaines méthodes ont proposé de grouper les régions d'intérêt d'une image suivant des critères topologiques et/ou statistiques. Dans (Zheng *et al.*, 2006) et (Zheng *et al.*, 2008), les auteurs établissent une analogie entre la recherche d'image (recherche d'objet visuel) et la recherche des documents texte. L'idée principale est de construire des "phrases visuelles"<sup>3</sup>, chaque phrase visuelle est un couple de régions d'intérêt à la fois adjacentes et fréquentes. L'adjacence ici signifie une proximité spatiale relative à la taille de chaque région et à la distance entre les deux régions. En codant chaque paire avec des mots visuels on transforme l'image en document textuel, sur lequel on peut appliquer les

2. Spatial Pyramid Matching (SPM).

3. Dans la suite de l'article, nous utilisons "phrase visuelle" (en minuscules) pour les approches existantes, et "Phrase Visuelle" pour dénoter notre méthode.



**Figure 2.** Extraction d'une représentation du contenu visuel d'une image à base d'une pyramide spatiale de trois niveaux.

techniques de recherche d'information textuelle (tf, idf, similarité par cosinus, etc.). Les auteurs de (Yuan *et al.*, 2007) proposent une autre technique de regroupement des régions d'intérêt basée sur les  $k$  plus proches voisins. Pour chaque région d'intérêt, une "phrase visuelle" est formée en prenant les 4 régions d'intérêt les plus proches d'après leurs centres dans l'image. Puis, en appliquant des techniques de fouille de données, on arrive à détecter des patrons de cooccurrence entre les mots visuels des phrases. Cette méthode est appliquée sur un ensemble d'image de visages et a permis de différencier différentes parties de visages (yeux, nez, bouche). Nous signalons les limitations suivantes dans ces approches :

- la cardinalité des phrases visuelles : les phrases visuelles dans ces approches contiennent un nombre de mots visuels défini a priori, sans réelle justification ;
- les phrases visuelles ne sont pas disjointes : dans une image, il peut y avoir des phrases visuelles qui partagent des régions. Ce fait génère un grand nombre de phrases visuelles par image (jusqu'à plusieurs milliers) ce qui complique et ralentit beaucoup un apprentissage ultérieur éventuel.

Dans la méthode proposée par (Tirilly *et al.*, 2008), l'axe principal de la localisation des régions d'intérêt est extrait à travers une analyse en composantes principales (ACP). Puis, les régions sont projetées sur cet axe principal pour formuler une phrase visuelle dans laquelle l'ordre des mots est pris en compte. La projection permet de préserver des informations visuelles liées à l'organisation spatiale des régions d'intérêt ce qui peut améliorer la reconnaissance des objets. Le problème de cette méthode est qu'elle est peu adaptée au cas où il y a plusieurs objets dans l'image, ainsi qu'aux images possédant des fonds complexes.

Dans notre proposition, nous définissons un modèle pour la génération et la description des groupes de régions d'intérêt (et tout type de régions). Appelé "modèle de phrasage", ce modèle offre la possibilité de définir et de comparer différents schémas de regroupement. Nous instancions ce modèle afin de générer des groupes (appelés "Phrases Visuelles") évitant les limitations de l'état de l'art en particulier les approches basées sur des décompositions prédéfinies. Notre objectif est de proposer un regroupement qui, d'une part, génère d'une façon dynamique (non prédéfinie) des groupes de régions d'intérêt en tenant compte des relations topologiques et qui, d'autre part, peut être appliqué pour l'annotation automatique d'images. Nous pensons qu'un tel groupement peut donner des résultats de meilleure qualité que ceux obtenus par un modèle classique de sac de mots visuels. Avant de définir et d'évaluer ce regroupement, nous procédons d'abord à la description du modèle d'extraction de descripteurs visuels.

### 3. Modèle de phrasage et Phrases Visuelles

En tentant de rester général, un processus d'extraction de descripteurs visuels peut être décomposé en trois étapes :

- 1) détermination, par une étape de segmentation, des régions contenant les informations visuelles jugées utiles pour la tâche souhaitée ;
- 2) regroupement des régions déterminées selon un ou plusieurs critères ;
- 3) description du contenu visuel des groupes créés.

Nous formalisons ces trois étapes, ainsi que le résultat de ces traitements, dans un modèle de *Phrasage*<sup>4</sup> décrit ci-dessous :

$$Phrasage = \langle F_{seg}, F_{gr}^c, F_{desc}, PH \rangle \quad [1]$$

avec :

---

4. Nous reprenons la définition de l'image de (Martinet, 2004) correspondant à notre vue de l'image : Une image  $I$  est un ensemble de pixels connexes organisés dans une matrice rectangulaire.

\*  $F_{seg}$  : une fonction de segmentation d'image en régions :

$$F_{seg} : IM \longrightarrow ER^{IM} = \bigcup_{i=1}^{n_I} \mathcal{P}'(I_i)$$

$$\forall I \in IM : F_{seg}(I) = R_{seg}^I \subseteq \mathcal{P}'(I) \quad [2]$$

avec :

- $IM$  l'ensemble de toutes les images d'une collection donnée.  $IM = \{I\}$ .
- $ER^{IM}$  l'ensemble de toutes les régions des images de  $IM$  (organisées image par image).
- $\mathcal{P}'(I_i)$  l'ensemble des parties non vides d'une image  $I_i$ .
- $R_{seg}^I$  l'ensemble des régions de l'image  $I$  obtenues par une segmentation  $seg$ . (ex. détection de régions d'intérêt, décomposition en pyramide spatiale, segmentation basé sur les couleurs, etc.).

\*  $F_{gr}^c$  : une fonction de regroupement des régions d'image, basée sur un critère  $c$ , qui engendre les groupes de régions constituant les Phrases :

$$F_{gr}^c : ER^{IM} \longrightarrow \mathcal{P}'(ER^{IM})$$

$$\forall R_{seg}^I \in ER : F_{gr}^c(R_{seg}^I) = G_c^I \subseteq \mathcal{P}'(R_{seg}^I) \quad [3]$$

La fonction de regroupement  $F_{gr}^c$  prend en entrée l'ensemble des régions d'une image  $R_{seg}^I$ , déterminé par la fonction de segmentation  $F_{seg}$ , et renvoie en sortie un ensemble de groupes des régions notées  $G_c^I$  satisfaisant le critère  $c$ .

Le critère de regroupement  $c$  définit la condition qu'une fonction de groupement doit vérifier lors de la création des Phrases Visuelles. Cette condition peut inclure des contraintes spatiales comme la distance entre les régions, leurs tailles, leurs emplacements ou leurs positionnements relatifs ; elle peut également se baser sur les valeurs des descripteurs des régions. Ce critère de regroupement est défini par :

$$c : \mathcal{P}'(R_{seg}^I) \rightarrow Boolean \quad [4]$$

Avec  $\mathcal{P}'(R_{seg}^I)$  l'ensemble des parties non-vides de  $R_{seg}^I$ .

\*  $F_{desc}$  : une fonction de description des groupes engendrés par  $F_{gr}^c$ . Elle prend en entrée un groupe de régions et renvoie un descripteur :

$$F_{desc} : \mathcal{P}'(ER) \longrightarrow DESC$$

$$\forall G_c^I \in ER : F_{desc}(G_c^I) = D_c^I \in DESC \quad [5]$$

Avec  $DESC$  le domaine des valeurs possibles des descripteurs.

La fonction de description peut être simple (par ex. la moyenne des couleurs des régions du groupe), ou complexe (par ex. générer un modèle de langue basé sur les mots visuels décrivant les régions du groupe).

\*  $PH$  : le résultat obtenu par l'application successive de  $F_{seg}$ , de  $F_{gr}^c$  et de  $F_{desc}$ . Ce résultat est un ensemble de groupes appelés *Phrases Visuelles*. Les Phrases Visuelles représentent les descripteurs visuels des images dans notre modélisation. La notion de Phrase Visuelle est définie comme suit :

*Une Phrase Visuelle est un ensemble de régions dans une image, regroupées suivant un critère prédéfini, et muni d'un descripteur.* Notons  $ph^i = \langle G, D \rangle$  une Phrase Visuelle appartenant à une image  $I$ , où  $G$  est un groupe de régions de cette image, et  $D$  est le descripteur de la Phrase. Nous pouvons maintenant définir  $PH$  comme suit :

$$PH = \{ \langle G, F_{desc}(G) \rangle : G \in G^I \wedge I \in IM \} \quad [6]$$

Après avoir détaillé notre modèle de phrasage, nous l'utilisons pour définir des méthodes d'extraction de descripteurs visuels.

#### 4. Instances

Dans cette section, nous évaluons l'effet de la prise en compte des relations topologiques entre les régions d'intérêt sur les résultats de l'annotation automatique. Afin d'effectuer cette évaluation, nous utilisons notre modèle de phrasage pour définir et comparer deux méthodes d'extraction de descripteurs :

- 1) Une méthode classique de sac de mots visuels, regroupant les régions d'intérêt de l'image sans tenir compte d'aucune relation topologique (notée  $Phrasage_{image}$ );
- 2) Une méthode basée sur un regroupement topologique des régions d'intérêt (notée  $Phrasage_{topo}$ ).

Les deux phrasages partagent les mêmes fonctions de segmentation et de description :

- a) Fonction de segmentation en régions d'intérêt, notée  $F_{seg-ri}$  ;
- b) Fonction de description en histogramme normalisé de fréquence des mots visuels, noté  $F_{desc-hist}$ .

Ces deux phrasages diffèrent donc au niveau de la fonction de regroupement. Suivant notre modèle présenté en 3, nous décrivons les deux instances considérées :

$$Phrasage_{image} = \langle F_{seg-ri}, F_{gr-image}, F_{desc-hist}, PH_{image} \rangle \quad [7]$$

$$Phrasage_{topo} = \langle F_{seg-ri}, F_{gr-connx}, F_{desc-hist}, PH_{topo} \rangle \quad [8]$$

La modélisation des deux instances ci-dessus permet de montrer explicitement que la différence entre les deux méthodes d'extraction des descripteurs visuels se situe au niveau du regroupement des régions. L'application du modèle de phrasage

proposé dans cet article permet donc de faciliter et de clarifier la comparaison entre les méthodes d'extraction des descripteurs visuels. En explicitant et en formalisant la présentation les différentes étapes, les méthodes peuvent être comparées suivant une référence formelle. Ceci évite des présentations peu structurées compliquant la compréhension et la comparaison entre les paramètres appliqués lors de l'extraction des descripteurs visuels.

#### 4.1. Regroupement classique

La première regroupe toutes les régions d'intérêt de l'image dans une seule Phrase Visuelle, sans tenir compte des informations ou des relations topologiques entre les régions regroupées. Cette fonction notée  $F_{gr-image}$  est définie comme suit :

$$\begin{aligned} F_{image} : ER &\longrightarrow \mathcal{P}'(ER) \\ \forall R_{seg-ri}^I \in ER : F_{image}(R_{seg-ri}^I) &= R_{seg-ri}^I \end{aligned} \quad [9]$$

La figure 3 montre le résultat de l'application de  $F_{gr-image}$  sur une image (toutes les régions d'intérêt, circulaires, ont la même couleur pour montrer qu'elles appartiennent à la même Phrase Visuelle).



**Figure 3.** Exemple du résultat de l'application d'un regroupement classique en sac de mots visuels.

#### 4.2. Regroupement en phrases visuelles à base de relations topologiques

Le regroupement de la deuxième méthode tient compte d'une relation topologique entre régions d'intérêt. Inspirés de la théorie des graphes, nous créons des composants faiblement connexes afin de créer des groupes de région d'intérêt ayant des pixels en commun. Chaque région d'intérêt dans une image représente un nœud dans un graphe ; deux nœuds sont reliés si et seulement si ils correspondent à deux régions d'intérêt connexes<sup>5</sup>. Pour un groupe de régions connexes noté  $G_{connx}$ , toute région d'intérêt  $r_i \in G_{connx}$ ,  $r_i$  a les deux propriétés suivantes :

(\*) si  $G_{connx}$  n'est pas singleton, il existe au moins une autre région  $r_y \in G_{connx}$  qui satisfait avec  $r_i$  le critère  $c$  ;

(\*\*)  $r_i$  ne satisfait  $c$  avec aucune région en dehors de  $G_{connx}$ .

Nous exprimons ces deux propriétés dans la formule 10.

$$\begin{aligned}
 &\forall r_i \in G_{connx} : \\
 &(*) \text{ if } (|G_{connx}|) > 1 : \exists r_y \in G_{connx} : r_y \neq r_i \wedge c(r_i, r_y) \\
 &(**) \text{ if } (|G_{connx}|) \geq 1 : \forall r_u \notin G_{connx} : \neg c(r_i, r_u)
 \end{aligned} \tag{10}$$

Nous choisissons cette fonction de regroupement, notée  $F_{connx}$  pour les raisons suivantes :

- Les Phrases Visuelles générées sont robustes aux rotations et aux translations : ces changements n'affectent pas les tailles des régions d'intérêt ni leurs positionnements relatifs, la contrainte topologique est donc préservée.

- Les Phrases Visuelles générées sont partiellement robustes aux changements d'échelle : quand l'échelle devient plus grande, les régions s'agrandissent proportionnellement à l'échelle, préservant la contrainte topologique ; cependant, cette robustesse est partielle, car un changement d'échelle peut faire apparaître ou disparaître des régions d'intérêt, et affecter ainsi la contrainte de connectivité.

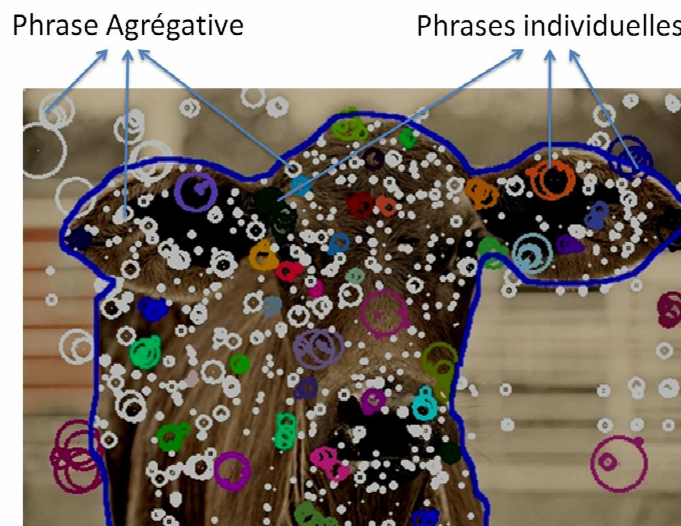
- Ce regroupement ne dépend pas du point de départ choisi, ce qui évite beaucoup de problèmes concernant le choix des régions de départ, et garantit un comportement d'extraction homogène dans toutes les images.

- Les groupes créés par ce regroupement sont disjoints, ce qui aide à limiter le nombre de Phrases générées et à éliminer la redondance qui peut ralentir l'apprentissage.

**Cas des Phrases courtes :** le regroupement proposé n'évite pas la création de Phrases ne contenant qu'une seule région. Comme nous l'avons mentionné dans l'introduction, une seule région d'intérêt n'est pas capable individuellement ni de caractériser ni

5. Deux régions sont considérées connexes si et seulement si elles partagent au moins un pixel en commun.

de distinguer des objets visuels ou des images. Nous généralisons cette idée en considérant que les Phrases Visuelles comprenant un nombre de régions d'intérêt inférieur à un seuil donné ne sont pas capables individuellement de caractériser ou de distinguer des objets visuels ou des images. Nous regroupons donc ces Phrases "courtes" en une seule "Phrase Agrégative". L'idée sous-jacente à ce regroupement est d'accumuler les informations visuelles des Phrases Visuelles "courtes" pour créer une "Phrase Agrégative" plus descriptive et discriminante. Pour appliquer cette idée, nous introduisons un seuil de cardinalité  $S_{card}$  à la fonction de regroupement  $F_{gr-connx}$ , pour chaque classe d'objets ou catégorie d'images. La figure 4 montre un exemple d'une image après l'utilisation de  $F_{gr-connx}$  avec un seuil de cardinalité égal à 5, les régions de la Phrase Agrégative sont de la couleur la plus claire, les autres Phrases Visuelles sont considérées individuellement et de couleur plus foncée.



**Figure 4.** Exemple d'application la fonction de regroupement  $F_{gr-connx}$  avec un seuil de cardinalité égal à 5.

## 5. Expérimentations

Les expérimentations sont menées sur la collection VOC2009 contenant 14 743 images organisées en 20 classes d'objets, et divisée en deux ensembles de même taille :

- 1) Ensemble d'apprentissage : contenant des images associées avec des étiquettes précisant les objets qui présents dans chaque image.



2) Ensemble de test : contenant des images non annotées, sur lesquelles les approches sont testées.

La mesure d'évaluation de VOC2009 est la précision moyenne de l'annotation par classe d'objets, et la moyenne sur toutes les classes *MAP* (Mean Average Precision).

Nous choisissons d'appliquer les deux instances de notre modèle de phrasage dans une méthode d'annotation automatique à base d'apprentissage automatique utilisant une machine à vecteurs de supports (SVM) (Cortes *et al.*, 1995). Grâce à leurs bons résultats dans plusieurs campagnes d'évaluation, l'apprentissage par SVM est actuellement le plus populaire dans le domaine d'annotation automatique d'image. Notre méthode d'annotation utilise une SVM d'un noyau gaussien *RBF* en mode "un contre tous". Le processus d'apprentissage consiste à retrouver des séparateurs entre les exemples positifs et négatifs d'une classe d'objets donné, ce processus diffère selon l'instance de phrasage considérée :

1) Pour la première instance, l'apprentissage est effectué sur les Phrases Visuelles des images (une Phrase par image), il génère un modèle de reconnaissance pour chaque classe d'objets. Le modèle estime un score d'annotation indiquant la relation entre une classe d'objets et une Phrase Visuelle (correspondant à une image), plus le score est élevé plus la relation est considérée forte. Pour annoter une nouvelle image, la Phrase Visuelle correspondant au sac de mots visuels de cette image est d'abord extraite, ensuite un score d'annotation pour chaque classe d'objets est estimé par le modèle d'annotation correspondant.

2) Dans la deuxième instance, nous appliquons deux apprentissages distincts, un pour les Phrases Agrégatives (une par image), et un autre pour les Phrases Visuelles considérés individuellement (plusieurs par image). On obtient ainsi deux modèles de reconnaissance correspondant par classe d'objets. Pour annoter une nouvelle image, nous en extrayons les deux types de Phrases Visuelles, puis nous invoquons les deux modèles de reconnaissance correspondant aux deux types (pour une classe d'objets donnée). Enfin, nous fusionnons les deux scores obtenus par chaque modèle. La fusion appliquée est une fusion linéaire pondérée normalisés, cette fusion est de la forme suivante :

$$\alpha_{obj} * score_{Agrégative} + (1 - \alpha_{obj}) * score_{Individuelles} \quad [11]$$

Avec  $\alpha_{obj} \in [0, 1]$  le poids attribué aux scores des Phrases Agrégatives d'une classe d'objets *obj*.

Les paramètres des deux méthodes (les paramètres des SVM, les seuils de cardinalité  $S_{card}$  pour chaque classe d'objets et les poids des Phrases Agrégatives  $\alpha_{obj}$  pour chaque classe d'objets) sont optimisés par validation croisée sur l'ensemble d'apprentissage. Après l'optimisation des seuils  $S_{card}$  nous avons pu vérifier notre hypothèse sur la dépendance de ce seuils et la classe d'objets à identifier.

Les résultats obtenus par chaque instance ainsi que les valeurs du seuil de cardinalité pour chaque classe d'objets sont présentés dans le tableau 1. Ce tableau montre également les différences relatives entre ces deux approches (pourcentage d'amé-

| Objet        | $Phrasage_{image}$ | $Phrasage_{topo}$ | $S_{Card}$ |
|--------------|--------------------|-------------------|------------|
|              | AP                 | AP                |            |
| aeroplane    | 0,6516             | 0,691 (+6.0%)     | 2          |
| bicycle      | 0,3453             | 0,3734 (+8.1%)    | 3          |
| bird         | 0,3415             | 0,3427 (+0.3%)    | 20         |
| boat         | 0,3847             | 0,4081 (+6.1%)    | 3          |
| bottle       | 0,1777             | 0,1868 (+5.1%)    | 5          |
| bus          | 0,4279             | 0,4407 (+3.0%)    | 20         |
| car          | 0,3158             | 0,356 (+12.7%)    | 12         |
| cat          | 0,3834             | 0,3983 (+3.9%)    | 6          |
| chair        | 0,4255             | 0,4267 (+0.3%)    | 12         |
| cow          | 0,1932             | 0,2266 (+17.3%)   | 8          |
| dining table | 0,3120             | 0,3272 (+4.9%)    | 7          |
| dog          | 0,2514             | 0,2637 (+4.9%)    | 12         |
| horse        | 0,3507             | 0,365 (+4.1%)     | 15         |
| motorbike    | 0,3121             | 0,3331 (+6.7%)    | 20         |
| person       | 0,713              | 0,7186 (+0.8%)    | 20         |
| potted plant | 0,1709             | 0,1809 (+5.9%)    | 20         |
| sheep        | 0,2486             | 0,2523 (+1.5%)    | 4          |
| sofa         | 0,1787             | 0,2476 (+38.6%)   | 2          |
| train        | 0,4639             | 0,4677 (+0.8%)    | 10         |
| tv/monitor   | 0,2992             | 0,3738 (+24.9%)   | 3          |
| MAP          | 0,3474             | 0,3690 (+6.2%)    |            |

**Tableau 1.** Précisions moyennes des méthodes évaluées sur la collection VOC2009.

lioration entre parenthèses) Nous remarquons que la deuxième instance, fondée sur un regroupement topologique, donne une meilleure précision moyenne sur toutes les classes d'objets, avec des différences relatives allant de +0.3% (pour *bird* et *chair*) à +38.6% pour la classe *sofa*. Ces résultats montrent sans ambiguïté que la prise en compte des relations topologiques entre les régions d'intérêt amène à une annotation automatique de meilleure qualité. La valeur *MAP* de cette instance est supérieure de +6,2% à celle de la méthode classique de sac de mots visuels (0,3690 contre 0,3474 respectivement). Afin de déterminer si cet écart est significatif, nous avons effectué un test bilatéral non-paramétrique de Wilcoxon (Wilcoxon, 1945) sur les précisions moyennes des deux instances avec un seuil de signification de 5%. Ce test confirme que la différence entre les deux résultats est statistiquement significative.

## 6. Conclusion et perspectives

Dans cet article nous avons proposé un modèle d'extraction des descripteurs visuels appelé modèle de phrasage. Le modèle proposé décompose le processus d'ex-

traction des descripteurs en trois étapes : la segmentation en régions, le regroupement des régions et la description des regroupements. Les descripteurs construits sont appelés Phrases Visuelles. Le modèle de phrasage permet de définir formellement chaque étape, ce qui standardise et facilite l'étude de chacune d'entre elles.

Nous nous sommes concentrés dans cet article sur l'étape de regroupement des régions, afin d'évaluer l'effet de la prise en compte des relations topologiques entre régions d'intérêt dans le contexte de l'annotation automatique d'image. Nous avons instancié notre modèle en deux méthodes : une première représentant une description classique de l'image en sac de mots visuels, cette instance est basée sur un regroupement de toutes les régions d'intérêt de l'image dans une seule Phrase Visuelle sans prise en compte de relation topologique entre les régions. Dans la deuxième instance nous avons utilisé une fonction de regroupement topologique basé sur l'intersection entre les régions d'intérêt. Nous avons évalué les deux instances en dans le cadre d'une annotation automatique basée sur un apprentissage supervisé par SVM. Les évaluations effectuées sur la collection VOC2009 montrent que notre méthode proposée se comporte mieux que la méthode classique en sac de mots visuels en donnant de meilleurs résultats, avec une différence statistiquement significative.

Les résultats obtenus nous encouragent à faire une analyse plus approfondie sur les Phrases Visuelles topologiques afin de déterminer s'il existe des Phrases correspondant à certaines sémantiques (objets ou parties d'objets) : des résultats préliminaires, non décrits dans cet article, montrent que pour certaines classes d'objets (par exemple les classes aeroplane et TV/monitor) il existe des Phrases qui ont une géométrie commune et qui correspondent à des parties précises des objets. Ces éléments permettent de raffiner l'annotation automatique en estimant la localisation de certaines parties d'objets.

## 7. Bibliographie

- Bay H., Tuytelaars T., Van Gool L., « SURF : Spcedded Up Robust Features », *Computer Vision ECCV 2006*, vol. 3951 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, chapter 32, p. 404-417, 2006.
- Bres S., Jolion J.-M., « Detection of Interest Points for Image Indexation », *In 3rd Int. Conf. on Visual Inf. Systems, Visual 99*, Springer, p. 427-434, 1999.
- Cortes C., Vapnik V., « Support-vector networks », *Machine Learning*, vol. 20, n° 3, p. 273-297, September, 1995.
- Larlus D., Création et utilisation de vocabulaires visuels pour la catégorisation d'images et la segmentation de classes d'objets, PhD thesis, INPG, nov, 2008.
- Lazebnik S., Schmid C., Ponce J., « Beyond Bags of Features : Spatial Pyramid Matching for Recognizing Natural Scene Categories », *CVPR '06 : Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, IEEE Computer Society, p. 2169-2178, October, 2006.

Albatal, Mulhem, Chiaramella

- Lowe D. G., « Object Recognition from Local Scale-Invariant Features », *ICCV '99 : Proceedings of the International Conference on Computer Vision-Volume 2*, IEEE Computer Society, Washington, DC, USA, p. 1150, 1999.
- Martinet J., Un modèle vectoriel relationnel de recherche d'information adapté aux images, PhD in computer science, Université Joseph Fourier, Grenoble, 2004.
- Perrett D. I., Oram M. W., « Visual recognition based on temporal cortex cells : viewer-centred processing of pattern configuration. », *Z Naturforsch C*, vol. 53, n° 7-8, p. 518-41, 1998.
- Perronnin F., Dance C. R., « Fisher Kernels on Visual Vocabularies for Image Categorization », *CVPR*, 2007.
- Sivic J., Zisserman A., « Video Google : a text retrieval approach to object matching in videos », *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, p. 1470-1477 vol.2, April, 2003.
- Tanaka K., « Mechanisms of visual object recognition : monkey and human studies. », *Curr Opin Neurobiol*, vol. 7, n° 4, p. 523-529, August, 1997.
- Tirilly P., Claveau V., Gros P., « Language modeling for bag-of-visual words image categorization », *CIVR '08 : Proceedings of the 2008 international conference on Content-based image and video retrieval*, ACM, New York, NY, USA, p. 249-258, 2008.
- Tola E., Lepetit V., Fua P., « A fast local descriptor for dense matching », *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, p. 1-8, 2008.
- Tuytelaars T., Schmid C., « Vector Quantizing Feature Space with a Regular Lattice », *ICCV*, p. 1-8, 2007.
- Wilcoxon F., « Individual Comparisons by Ranking Methods », *Biometrics Bulletin*, vol. 1, n° 6, p. 80-83, 1945.
- Winder S. A., Brown M., « Learning Local Image Descriptors », *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, p. 1-8, 2007.
- Yuan J., Wu Y., Yang M., « Discovery of Collocation Patterns : from Visual Words to Visual Phrases », *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, p. 1-8, June, 2007.
- Zheng Q.-F., Gao W., « Constructing visual phrases for effective and efficient object-based image retrieval », *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 5, n° 1, p. 1-19, 2008.
- Zheng Q.-F., Wang W.-Q., Gao W., « Effective and efficient object-based image retrieval using visual phrases », *MULTIMEDIA '06 : Proceedings of the 14th annual ACM international conference on Multimedia*, ACM, New York, NY, USA, p. 77-80, 2006.