Phrases Visuelles pour l'annotation automatique d'images

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. *L'annotation automatique d'images photographiques est un problème complexe. En effet, les caractéristiques visuelles des objets d'une classe varient selon l'instance considérée et les conditions de prise de vue. Nous proposons dans cet article une caractérisation visuelle des parties d'objets appelées "Phrases Visuelles", robuste à ces variations. Une Phrase Visuelle est un ensemble de régions d'intérêts construit suivant des critères prédéfinis; un critère proposé et étudié ici est de nature topologique. Basé sur notre définition et caractérisation de Phrases Visuelles, nous proposons une méthode d'annotation d'images. Une expérimentation sur le corpus VOC2009 est présentée, et nous montrons que fusionner notre approche avec une approche standard à base de sac de mots visuels sur les images complètes fournit de meilleurs résultats qu'une annotation obtenue par approche standard seule.*

ABSTRACT. *Photographic images annotation is a complex problem. Indeed, the visual characteristics of objects of a class vary with the considered instance and the shooting conditions. In this paper we proposed a visual characterization of object parts, called "Visual Phrase", robust to these variations. A Visual Phrase is a set of regions of interest built according to pre-difined criteria; a topological criterium was studied in this paper. An automatic annotation method is proposed based on our definition and characterization of Visual Phrases. An experiment on VOC2009 corpus is presented, and we show that the fusion of our method with a standard bag of visual words approach on full images provides better results than those obtained via the standard approach.*

MOTS-CLÉS : *Phrase Visuelle, Région d'intérêt, Sac de mot visuel, Annotation d'image*

KEYWORDS: *Visual Phrase, Region of interest, Bag of visual word, Image annotation*

1. Introduction et problématique

Les Systèmes de Recherche d'Image par le Contenu (CBIR) ont fait ces dernières années de gros progrès pour rechercher des images visuellement proches d'une certaine image requête, ou bien pour retrouver un objet spécifique dans une image. Cependant, ces systèmes sont toujours peu performants en ce qui concerne la recherche sémantique d'images par requête textuelle. Une des raisons à cela vient de la façon dont les images sont décrites dans les systèmes informatiques.

Pour effectuer une recherche sémantique il faut être en mesure de transformer le contenu visuel des images (couleurs, textures, formes) en informations sémantiques ; cette transformation est appelée annotation automatique d'image. L'annotation automatique nécessite l'analyse des caractéristiques visuelles des objets afin de bien les décrire, les différencier et de bien les identifier dans les images.

L'annotation peut être réalisée de deux manières : en indiquant la présence d'une instance d'une classe d'objet (la voiture x est dans l'image I), ou bien juste la présence d'une instance quelconque de la classe (une voiture est dans l'image I). Dans cet article nous nous intéressons au deuxième cas d'annotation.

Un des problèmes majeurs de l'analyse des caractéristiques visuelles des classes d'objets est que les différentes instances des objets d'une classe n'ont pas toujours des caractéristiques visuelles identiques ; ces caractéristiques varient selon plusieurs facteurs :

1) l'instance considérée : une instance peut avoir des caractéristiques visuelles différentes des autres instances. Exemple : une voiture rouge et une voiture verte sont deux instances de la classe "Voiture" mais avec des caractéristiques de couleur différentes.

2) les conditions de prise de vue de l'image : comme la luminosité, l'axe de prise de vue, la distance entre le point de vue et l'objet photographié. Ces conditions peuvent avoir des effets sur l'échelle, la rotation, et l'emplacement de l'objet dans l'image.

3) le contexte d'occurrence de l'objet dans l'image : la présence de fonds complexes et d'autres objets dans l'image peut cacher ou *polluer* des caractéristiques visuelles de l'objet. Exemple : une voiture peut être occultée par un arbre.

Ces facteurs résultent de variations visuelles qui compliquent beaucoup l'analyse des caractéristiques visuelles des objets et, conséquemment, l'annotation basée sur cette analyse. Cette complexité est appelée *le fossé sémantique* introduit par (Smeulders *et al.*, 2000) : "*Le fossé sémantique est le manque de concordance entre l'information que l'on peut extraire des données visuelles et l'interprétation que les mêmes données ont pour un utilisateur dans une situation donnée*" [1].

Les approches récentes qui traitent le problème du fossé sémantique essayent de trouver des zones dans les images qui contiennent des informations visuelles robustes

1. *The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation.*

aux variations visuelles afin de minimiser cet effet négatif. En particulier, l'extraction et la description des régions d'intérêt sont appliquées avec succès pour détecter ces zones. Une méthode basée sur les régions d'intérêt et qui donne des bons résultats d'annotation, est la méthode de sac de mots visuels ; cette méthode, standard, regroupe toutes les régions d'une image dans une structure d'histogramme et effectue des processus d'apprentissage et de classification sur cette structure. Une région d'intérêt individuelle est stable face à plusieurs variations visuelles, mais elle ne possède pas d'informations visuelles suffisantes pour décrire et différencier des classes objets (Zheng *et al.,* 2008a). La méthode proposée dans cet article essaye de regrouper des régions d'intérêt dans l'image afin de former des parties qui sont, spatialement, plus grandes que les régions d'intérêt, et qui possèdent les mêmes propriétés invariantes. Ces regroupements sont capables de décrire et de différencier les classes d'objets. Nous appelons ces parties les "*Phrases Visuelles*".

En section 2, nous proposons un état de l'art sur la caractérisation visuelle des images basée sur les régions d'intérêt et le modèle de sac de mots visuels, et nous présentons des méthodes d'annotation automatique qui inspirent notre travail. Nous définissons en section 3 notre notion de "*Phrase Visuelle*", et nous montrons comment nous représentons les Phrases Visuelles Dans les images. Une approche d'annotation automatique d'images basée sur les "*Phrases Visuelles*" est également proposée dans cette section. En section 4 nous instancions notre approche d'annotation automatique et nous appliquons cette approche dans le cadre d'une expérimentation sur la collection VOC2009[2]. Dans cette partie, nous montrons que la fusion entre notre approche et l'approche standard décrite ci-dessus donne toujours des résultats supérieurs aux résultats de chaque approche indépendamment. Enfin, nos conclusions et perspectives sur ce travail sont présentées en section 5.

2. Etat de l'art

2.1. *Extraction et description des régions d'intérêt*

L'extraction et la description des régions d'intérêt est une technique de plus en plus utilisée avec succès dans l'annotation d'image, ainsi que dans plusieurs domaines de la vision par ordinateur comme la reconnaissance des formes (Mikolajczyk *et al.,* 2003), suivi (Zhou *et al.,* 2009), reconstruction (Ni *et al.,* 2008), calibrage (Yun *et al.,* 2006). Les bons résultats obtenus sont dus à la robustesse aux variations visuelles venant des techniques de détection et de description des régions d'intérêt.

Il n'existe pas de définition unique et claire du terme "régions d'intérêt" ou "point d'intérêt". Dans (Bres *et al.,* 1999), les auteurs indiquent que la majorité de la littérature suppose qu'une région/point d'intérêt est équivalente à un coin dans l'image, ou, plus généralement, une région caractérisée par une valeur intéressante du gradient de luminosité dans plusieurs directions.

2. http ://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2009/

En général, une région d'intérêt possède les caractéristiques suivantes :

– Elle a une définition mathématique formelle,

– Elle a une position précise dans l'image,

– Elle est riche en informations visuelles locales,

– Elle est stable face à des variations locales et globales de l'image, i.e., elle conserve les mêmes informations visuelles en cas de variation.

(Grand-Brochier *et al.,* 2009) indique que l'idée sous-jacente des régions d'intérêt est liée au fait que lorsque quelqu'un regarde une image, il suffit de regarder ces points : 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. (Lowe, 1999) présente 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, location, et d'éclairage (Perret *et al.,* 1998) (Tanaka, 1997). Les détecteurs des régions/points d'intérêt essaient donc de détecter des régions qui simulent les éléments de complexité intermédiaire utilisés par le système de vision humaine. Dans ce cas, l'image est transformée en un ensemble de petites régions.

Normalement, les régions sont localisées via un détecteur, puis un descripteur utilise les pixels de chaque région pour la décrire. (Grand-Brochier *et al.,* 2009) a catalogué différentes techniques de détection et de description de régions d'intérêt par leur robustesse aux rotations et aux changements d'échelle. La robustesse provient en fait soit du détecteur, soit du descripteur, ou bien d'un couplage des deux. Les auteurs ont montré trois techniques robustes contre toutes ces variations : DOG (différence de gaussiennes) (Lowe, 1999), LOG (Tabbone *et al.,* 1993), Harris-Affine (Mikolajczyk *et al.,* 2005) et SURF (Bay *et al.,* 2006). Parmi ces techniques, la différence de gaussiennes et Harris-Affine sont les plus utilisées actuellement.

2.2. *Modèle de sac de mots visuels*

En raison de son efficacité et de la qualité des résultats qu'il permet d'obtenir, le modèle par sac de mots visuels est devenu populaire durant ces dernières années (Jiang *et al.,* 2007) (Lazebnik *et al.,* 2006) (Fei-Fei *et al.,* 2005) (Sivic *et al.,* 2003) (Zhang *et al.,* 2006) (Zhao *et al.,* 2006). Cette synthèse du contenu visuel est présentée premièrement par (Sivic *et al.,* 2003) dans le cas de la recherche de vidéo et par (Csurka *et al.,* 2004) dans le domaine de la catégorisation des images. Cette approche a été initialement développée pour la catégorisation de textes, domaine dans lequel elle s'est avérée très performante (Joachims, 1998) : chaque document est représenté par un histogramme basé sur la fréquence d'apparition de chaque mot du vocabulaire et les histogrammes subissent diverses normalisations. Puis, différentes techniques d'apprentissage peuvent être appliquées sur ces histogrammes.

Pour appliquer le modèle de sac de mots dans le domaine visuel, il faut créer des "mots" visuels. Cette création est effectuée par une quantification des valeurs des descriptions visuelles. En général, la quantification utilise un algorithme de *clustering* (par exemple : K-Moyennes (Macqueen, 1967)) qui prend comme entrée un échantillon des descriptions et crée, à partir de cet échantillon, des ensembles (clusters). Ensuite, le centroïde de chaque cluster représente un "mot" visuel. L'ensemble des mots visuels est appelé vocabulaire visuel. Après l'attribution des mots visuels aux régions, les images sont vues comme des sacs de mots visuels, et on peut construire des histogrammes dont la dimension est égale à la taille de vocabulaire visuel (nombre de clusters), chaque bin contient la fréquence du mot visuel correspondant dans l'image ou dans une partie de l'image.

Notons cependant que, selon (Larlus, 2008), les mots visuels sont beaucoup plus ambigus que les mots textuels : « Il est impossible de créer des mots qui soient toujours observés sur la même partie d'un objet et jamais ailleurs ». Dans la représentation en sac de mots visuels, on perd les informations liées à l'organisation topologique des régions dans l'image, ces information peuvent être importantes pour décrire et différencier les objets. Nous décrivons dans la partie suivante des approches qui ont proposé des regroupements des régions d'intérêt en prenant en comptes leur relations topologiques.

2.3. *Regroupement des régions d'intérêt*

Deux méthodes ont proposé des techniques de regroupement pour créer des groupes de régions d'intérêt qui ont certaines caractéristiques spatiales. Au lieu de caractériser (visuellement) la totalité de l'image, la caractérisation est effectuée au niveau des groupes de régions, les groupes créés servant à décrire le contenu visuel en intégrant des informations spatiales de ces régions.

(Zheng *et al.,* 2006) (Zheng *et al.,* 2008a) 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"[3], chaque phrase visuelle est un couple de régions d'intérêt à la fois adjacentes et fréquentes. En codant chaque paire avec des mots visuels on transforme l'image en document textuel, sur lequel on peut appliquer les techniques de recherche d'information textuelle (tf, idf, similarité par cosinus, etc.). L'approche proposée est, donc, de type (requête image → réponses image).

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

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

phrases. Cette méthode est appliquée sur un ensemble d'image de visages et à permis de différencier différentes parties de visages (yeux, nez, bouche).

Nous remarquons les limitations suivantes dans ces approches :

– La longueur 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 l'analyse en composantes principales (ACP). Puis, les régions sont projetées sur l'axe principal, pour formuler une phrase visuelle où 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 donnons une définition formelle et générale de la notion de "Phrase Visuelle" ; et nous l'instancions d'une façon qui évite les limitations des autres approches et qui peut être appliquée dans des approches d'annotation automatique d'images réelles.

3. Phrase Visuelle

3.1. *Motivations*

Comme nous l'avons indiqué dans l'introduction, une région d'intérêt individuelle n'est pas suffisante pour caractériser ou différencier visuellement les classes d'objets. Il a été montré dans (Zheng *et al.,* 2008b) que la jointure entre plusieurs régions d'intérêt caractérisées (mots visuels) peut décrire et différencier des classes d'objets. La question qui se pose ici est : "*Quelles sont les régions d'intérêt dans une image qu'il faut regrouper pour décrire et différencier les classes d'objets ?*".

Le modèle classique de sac de mots visuels regroupe toutes les régions d'intérêt dans une image, et ne prend en compte aucune relation spatiale entre les régions. Le modèle proposé dans la suite crée des sous-ensembles de régions d'intérêt, appelés "*Phrases Visuelles*", et les représente comme des sacs de mots visuels contraints par les relations spatiales entres les régions d'intérêt.

3.2. *Description*

Une Phrase Visuelle est un ensemble de régions d'intérêt regroupées suivant un critère prédéfini.

Une région d'intérêt r_p^I est une quadruplet $(x_p, y_p, ra_p, ca_p) \in M$, avec : x_p et y_p les cordonnées du centre de la région dans l'image I, ra_p le rayon de la région, ca_p la caractéristique visuelle de la région (un vecteur à n dimensions qui prennent ces valeurs du domaine CA) et M le domaine des valeurs possibles des régions d'intérêt $M = N \times N \times \Re \times CA$.

Notons R^I l'ensemble des régions d'intérêt dans une image I : $R^I = \{r_p^I\} \in \mathcal{P}'(M)$, sachant que r_p^I une région d'intérêt dans l'image I et $\mathcal{P}'(M)$ est l'ensemble des parties de M ôté de l'ensemble vide $\emptyset$.

L'ensemble des Phrases Visuelles dans une image I est noté $PH^I = \{ph_u^I\} \subset \mathcal{P}'(M)$, sachant que $ph_u^I \in \mathcal{P}'(M)$ est une Phrase Visuelle dans l'image I. L'ensemble des Phrases Visuelles dans une image est une sélection d'ensembles de régions d'intérêt. Pour faciliter la notation, notons : $S(M) = \mathcal{P}'(\mathcal{P}'(M))$ l'ensemble des parties de $\mathcal{P}'(M)$ ôté de l'ensemble vide $\emptyset$.

φ_c est une fonction de regroupement de régions d'intérêt dans les images qui a c comme critère. Cette fonction prend comme entrée l'ensemble des régions d'intérêt d'une image et renvoie en sortie un ensemble de Phrases Visuelles qui satisfont le critère c :

$$\begin{aligned} \varphi_c \ &: \mathcal{P}'(M) \rightarrow S(M) \\ R^I \quad &\mapsto \varphi_c(R^I) = PH^I \end{aligned}$$

Il reste à définir la fonction et le critère de regroupement adaptés à un cadre donné. Deux approches, (Yuan *et al.,* 2007) et (Zheng *et al.,* 2008b), ont utilisé une fonction de regroupement de K plus proches voisins (KNN) avec un voisinage utilisant un critère topologique (la distance euclidienne entre les régions d'intérêt). Comme nous l'avons décrit plus haut, deux conséquences directes de ce type de regroupement sont :

1) un nombre élevé de phrases visuelles : il se peut que le nombre de phrases visuelles soit égal au nombre de régions d'intérêt ;

2) dans une image les phrases visuelles crées ne sont pas disjointes et partagent des régions d'intérêt communes, cela amène à des redondances d'informations visuelles dans les phrases visuelles, ce qui peut compliquer des étapes d'analyse suivantes.

Un regroupement de types KNN rend l'apprentissage plus difficile et lent à cause du grand nombre des phrases visuelles et de la redondance d'informations ; de plus, il est difficile de fixer et justifier le choix de valeur de K (K=5 dans l'approche de (Yuan *et al.,* 2007), K=1 ou 4 ou 8 ou 12 dans l'approche de (Zheng *et al.,* 2008b)). Ces deux approches sont obligées de passer par des étapes supplémentaires pour réduire

le nombre des phrases visuelles extraites, et leurs buts ne sont pas d'effectuer une classification ou une annotation de d'images.

Notre but dans cet article étant de classifier les images selon des classes d'objets, nous cherchons pour cela à faire émerger des parties d'objets qui discriminent ces classes d'objets. La fonction que nous proposons ici est une fonction de clustering de type "Single Link" avec un critère de proximité topologique, noté c. Les Phrases Visuelles extraites qui contiennent au moins deux régions d'intérêt sont définies comme suit :

$$\begin{array}{ll} \forall ph_y^I \in PH^I \ : & \\ ph_y^I = \{r_p^I \in M^I \ : & \\ \forall r_p^I \in ph_y^I, \exists r_p^I \neq r_y^I \in ph_y^I \ : c(r_p^I, r_n^I) & (*) \\ \forall r_x^I \notin ph_y^I, \forall r_p^I \in ph_y^I \ : \neg c(r_y^I, r_x^I)\} & (**) \end{array}$$

Une Phrase Visuelle ph_y^I extraite via une fonction "Single link" est un ensemble de régions d'intérêt ; pour chaque région r_p^I dans ph_y^I : (*) il existe, au moins, une autre région de la même Phrase ph_y^I avec laquelle la condition de proximité c est satisfaite ; et (**) il n'existe aucune région qui n'appartient pas à ph_y^I et qui satisfait la condition de proximité en même temps.

Nous choisissons ce type de clustering pour les raisons suivantes :

– Il ne dépend pas du point de départ, ce qui évite beaucoup de problème concernant le choix des régions de départ, et garantit un comportement d'extraction homogène dans toutes les images.

– Les groupes créés par ce clustering sont disjoints, ce qui évite d'avoir des régions d'intérêt communes à plusieurs Phrases Visuelles. Cela aide à limiter le nombre de Phrases Visuelles extraites et à éliminer la redondance qui peut compliquer l'apprentissage.

Le critère topologique que nous utilisons a été proposé par (Zheng *et al.*, 2006), et est basé sur la distance euclidienne entre les régions d'intérêt :

$$c(r_p^I, r_n^I) \equiv \sqrt{(x_p - x_n)^2 + (y_p - y_n)^2} \leq ra_p + ra_n$$

Ce critère exprime que deux régions d'intérêt sont considérées comme proches si et seulement si la distance euclidienne entre leurs centres est inférieure ou égale à la somme de leurs rayons.

3.3. *Caractérisation visuelle des Phrases Visuelles*

Après l'extraction des Phrases Visuelles il faut les représenter sous un format permettant d'effectuer un apprentissage supervisé. Nous proposons d'adopter la représentation en sac de mot visuel qui a montré son efficacité dans le domaine de l'annotation automatique d'image (voir la section 2.2). Dans cette représentation, chaque Phrase

Visuelle est un histogramme de N dimensions, avec N le nombre des mots dans un vocabulaire visuel choisi.

Une étape précédente de la caractérisation des Phrases Visuelles est la construction d'un vocabulaire visuel. En général, cette étape est réalisé par un algorithme de clustering qui prend comme entrée une échantillon de vecteurs de caractérisation visuelle initiale (cf. l'état de l'art) des régions d'intérêt, et renvoie les centroïdes des clusters construits. Chaque centroïde est considéré comme un mot visuel, et est annoté par un identifiant. Ensuite, chaque région d'intérêt sera caractérisée par l'identifiant du plus proche centroïde. En fin, Pour construire l'histogramme de sac de mots visuel qui représente une Phrase Visuelle, il suffit de compter combien de fois chaque mot visuel est utilisé pour caractériser une région d'intérêt de la Phrase Visuelle. Une normalisation des valeurs des dimensions peut être effectuée comme étape finale.

3.4. *Propriété des Phrases Visuelles*

Nous montrons ici les propriétés des Phrases Visuelles extraites à travers une fonction de regroupement "Single Link" avec le critère de proximité topologique défini dans la section 3.2 :

1) les Phrases Visuelles sont invariantes aux changement d'échelle : le changement d'échelle ne modifie pas le critère de proximité topologique. Quand l'échelle devient plus grande, les distances entre les régions et les rayons des régions s'élargissent proportionnellement ;

2) les Phrases Visuelles sont invariantes aux rotations : la rotation n'affecte ni la distance entre les régions ni les rayons des régions, et le critère de proximité est donc préservé ;

3) les Phrases Visuelles sont invariantes aux translations : comme la rotation, la translation ne change pas le critère de proximité ;

4) si les régions d'intérêt sont invariantes aux changements de luminosité alors les Phrases Visuelles sont invariantes aux changement de luminosité : la robustesse aux changements de luminosité est une propriété qui ne dépend pas de la Phrase Visuelle, mais de chaque région d'intérêt individuelle dans la Phrase Visuelle. Si chaque région est invariante aux changements de luminosité, la Phrase Visuelle va hériter de cette propriété.

Ces propriétés invariantes nous encouragent à aller plus loin pour déterminer si les Phrases Visuelles ont des propriétés invariantes non seulement au niveau visuel mais aussi au niveau sémantique, lors d'une tâche d'annotation. Pour cela nous proposons, dans la suite, l'approche d'annotation automatique basée sur les Phrases Visuelles.

3.5. *Annotation à base de Phrases Visuelles*

Cette approche est organisée en trois étapes :

1) Etape de préparation des Phrases Visuelles pour l'apprentissage : cette étape comprend trois sous-étapes (l'extraction et caractérisation des régions d'intérêt dans l'image, l'extraction et la caractérisation des Phrases Visuelles et l'attribution des labels aux Phrases Visuelles extraites) ;

2) Etape d'apprentissage : nous appliquons un algorithme d'apprentissage discriminatif supervisé. L'entrée de l'algorithme est l'ensemble des vecteurs labélisés de caractérisation des Phrases Visuelle. L'algorithme génère un modèle d'annotation par classe, et ce modèle est capable d'attribuer à chaque Phrase Visuelle un score numérique qui indique si la Phrase Visuelle représente une partie d'objet de la classe considérée ou non ;

3) Etape d'annotation automatique : pour annoter automatiquement une image avec un label d'une classe d'objet on s'appuie sur les scores des Phrases Visuelles de l'image obtenus par le modèle d'annotation de cette classe. Il est nécessaire alors de calculer un score global d'image en utilisant les scores des Phrases Visuelles.

4. Expérimentations

Pour expérimenter notre approche nous avons choisi la collection VOC 2009 qui contient 14,743 images organisées en 20 classes d'objet, divisée en deux ensembles équivalents en taille :

1) Ensemble d'apprentissage : il contient des images annotées avec des boites rectangulaire englobantes qui précisent l'emplacement des objets dans l'image ;

2) Ensemble de test : il contient des images non annotées, sur les quelles les approches peuvent être testées.

Le mesure d'évaluation est la précision moyenne de l'annotation par classe d'objet, et la moyenne sur toutes les classes "MAP - Mean Average Précision".

4.1. *Instanciation de l'approche à base des Phrases Visuelles*

L'approche que nous proposons dans cet article est une instance de l'approche proposée dans 3.5. Les paramètres utilisés sont les suivants :

– Détection des régions d'intérêt via le détecteur "Harris Laplace" (Harris *et al.*, 1988) ;

– Description des régions d'intérêt via le descripteur "rgSift" (Van de Sande *et al.*, 2008) ;

– Création d'un vocabulaire visuel de 4000 mots visuels en appliquant un algorithme de clustering "K-means" sur un échantillon de 1 million de vecteurs de description rgSift sélectionnés aléatoirement ;

– Re-description des régions d'intérêt avec les mots visuels au lieu des vecteurs de description rgSift (en choisissant les centroïdes les plus proches des vecteurs de

description) ;

– Extraction des Phrases Visuelles avec un algorithme de clustering de type "single link" avec le critère de proximité géométrique défini dans la section 3.2 ;

– Description des Phrases Visuelles en sac de mots visuels ;

– Labélisation des Phrases Visuelles de l'ensemble de l'apprentissage : les Phrases Visuelles héritent les labels des boites englobantes qui les incluent[4] ;

– Apprentissage discriminatif supervisé utilisant des Machines à vecteurs de support un SVM (Vapnik, 1995) avec un noyau RBF. Les Phrases Visuelle du longueur 1 (qui contiennent une seule région d'intérêt) ne sont pas prises en compte puisqu'elles sont ambiguës par nature (Zheng *et al*., 2008a). Il en est de même avec les Phrases Visuelles de longueur 2 trop nombreuses, et qui ralentissent beaucoup le processus d'apprentissage. L'apprentissage est effectué en mode "un contre tous", et nous obtenons un modèle d'annotation des Phrases Visuelles pour chaque classe d'objet. Chaque modèle estime la probabilité qu'une Phrase Visuelle soit une partie d'un objet de la classe ;

– Pour annoter une image de l'ensemble de test, et pour une classe considérée, les Phrases Visuelles de l'image sont extraites et caractérisées visuellement. Puis, le modèle d'annotation des Phrases Visuelles de la classe attribue à chaque Phrase Visuelle un score de probabilité. Enfin, un score global pour l'image est calculé. Ce score combine par une somme une information provenant de l'image dans sa globalité (la moyenne des scores de toutes les Phrases Visuelles dans l'image) avec une information locale (le score maximal pour les Phrases Visuelles de l'image).

4.2. *Résultats*

Après avoir annoté toutes les images de l'ensemble de test VOC2009, nous avons obtenu les résultats présentés en figure 1. La précision moyenne sur toutes les classes d'objets est de 0,2624. Nous remarquons que la précision d'annotation varie selon la classe d'objets. Deux classes sont mieux détectées que les autres : "Aeroplane" 62% et "Person" 64,6%. La régularité des formes des objets de la classe "Aeroplane" aide à bien la caractériser, et donc bien la détecter. La classe "Person" a une probabilité d'occurrence très forte (plus d'un tiers des images de la collection contiennent une personne), ce qui explique la précision élevée d'annotation. D'autre part, quatre classes ne sont pas bien détectées, nous les catégorisons sous deux types :

1) des objets apparaissant sur des fonds complexes : "Dining Table" 10,6% et "Sofa" 11,3%, où le fond contient beaucoup d'autres objets ;

2) des objets de petites tailles : "Bottle" 12,2% et "Potted plant" 10,1%, où le fond occupe la majorité de l'image, il n'y a donc pas beaucoup d'information visuelle à extraire de l'objet.

4. Pour limiter la perte d'information, nous labélisons aussi les Phrases Visuelles qui ont au moins 90% de leurs points dans une boite englobante.

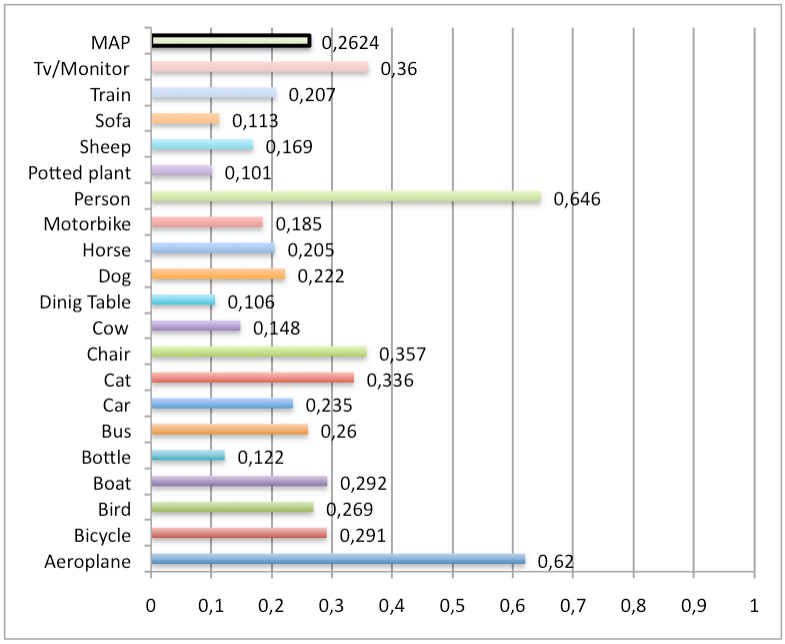

Figure 1. *Précisions moyennes de l'annotation à base de Phrases Visuelles sur la collection VOC2009.*

Du fait que nous n'analysons pas les fonds des objets dans l'apprentissage de l'approche à base de Phrases Visuelles, nous perdons probablement des informations visuelles extérieures à l'objet lui même mais qui peuvent être utiles pour son identification du fait d'une concordance significative pouvant exister entre les caractéristiques visuelles de la classe et celles du fond.

Nous présentons dans la figure 2 les résultats obtenu par l'approche standard de sac de mots visuels décrite en 2.2. La précision moyenne sur toutes les classes d'objets est de 0,3847. Nous remarquons que les résultats sont supérieurs aux ceux de l'approche basée sur les Phrases Visuelles, ceci du à la prise en compte des informations visuelles provenant à la fois des objets et de leurs fonds. Notamment, pour les deux classes "Dining Table" et "Train" les résultats sont supérieurs respectivement de +197% et +155%.

L'approche à base de Phrases Visuelles est centrée sur les objets : l'apprentissage est effectué sur les Phrases Visuelles à l'intérieur des boîtes englobantes des objets, tandis que l'approche standard de sac de mots visuels analyse l'image dans sa globalité (l'apprentissage est effectué sur des images complètes). Nous nous sommes alors posé la question de savoir si la fusion de notre analyse centrée sur l'objet avec l'approche standard pouvait améliorer les performances. Nous avons effectué une fusion

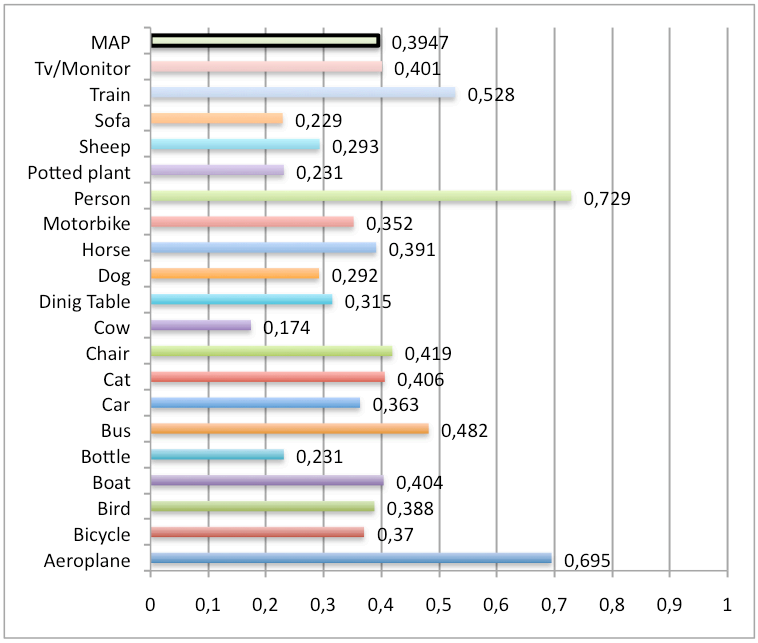

Figure 2. *Précisions moyennes de l'annotation d'une approche standard de sac de mots visuels sur la collection VOC2009.*

linéaire pondérée des scores d'annotations normalisés renvoyés par chaque approche, cette fusion est de la forme suivante :
$\alpha_{obj} * score_{Phrase} + (1 - \alpha_{obj}) * score_{standard}$ avec $\alpha_{obj} \in [0, 1]$ par pas de 0.05, pour chaque objet obj.
Les résultats ainsi obtenus sont supérieurs à ceux des deux approches prises séparément, comme le montre la figure 3. La précision moyenne est de 0,4195. Nous remarquons que la précision d'annotation pour plusieurs objets est nettement améliorée après la fusion par rapport à l'approche standard : "Motorbike" +22%, "Train" +15%, "Bus" +13%, "Chair" +11% et "Boat" +11%. Les deux meilleures classes restent toujours en tête sans grande amélioration sur la précision, "Aeroplane" +3% et "Person" +2%. En moyenne, la fusion améliore les résultats de l'approche standard de 9% et les résultats de l'approche basée sur les Phrases Visuelles de 59,87% . Ces améliorations des performances montrent que notre analyse des objets basée sur les Phrases Visuelles apporte un contribution très positive à l'approche standard.

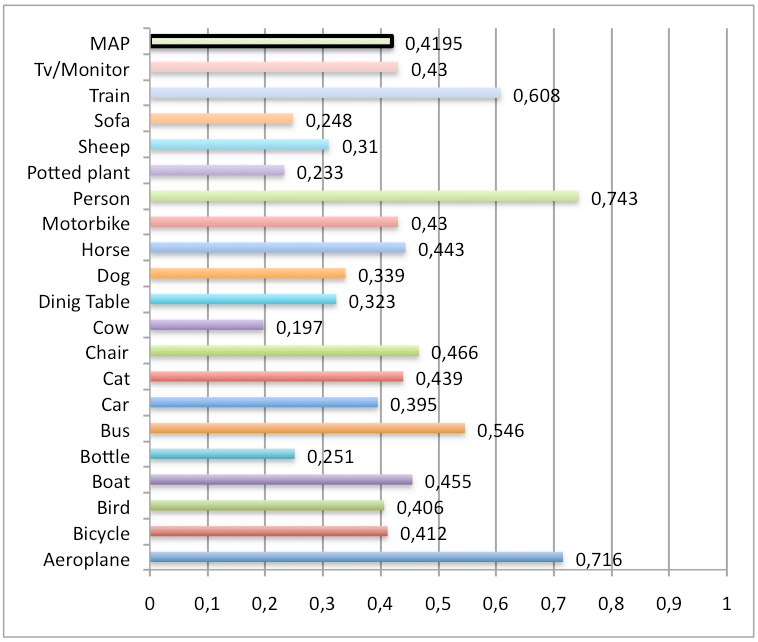

Figure 3. *Précisions moyennes de l'annotation de la fusion entre l'approche basée sur les Phrases Visuelles et l'approche standard de sac de mots visuels, sur la collection VOC2009.*

5. Conclusion et perspectives

Nous avons proposé dans cet article une méthode d'extraction des parties d'objets, appelées Phrases Visuelles robustes aux variations visuelles (changements d'échelle, rotation, translation et changement de luminosité), qui décrivent et différencient les parties d'objets. Ces Phrases Visuelles sont extraites par regroupement de régions d'intérêt suivant un critère topologique prédéfini. Nous avons proposé une approche d'annotation automatique d'images basée sur les Phrases Visuelles et nous l'avons expérimentée sur la collection VOC2009. Les résultats obtenus montrent qu'une fusion de notre approche avec une approche standard à base de sac de mots visuels apporte une amélioration nette aux résultats d'annotation de l'approche standard.

L'approche basée sur les Phrases Visuelles est centrée sur les objets et ne prend pas en compte leur fond. Pour améliorer les résultats d'une approche à base des Phrases Visuelles, nous estimons que la prise en compte des fonds d'objet dans le processus d'apprentissage et d'annotation est nécessaire. Des résultats préliminaires encourageants ayant déjà été obtenu, nous allons à présent élaborer des approches intégrant les caractéristiques des objets et des fonds.

6. Bibliographie

Bay H., Tuytelaars T., Gool V., L., « SURF : Speeded Up Robust Features », *9th European Conference on Computer Vision*, Graz Austria, p. 404-417, May, 2006.

Bres S., michel Jolion J., « Detection of Interest Points for Image Indexation », *In 3rd Int. Conf. on Visual Inf. Systems, Visual 99*, Springer, p. 427-434, 1999.

Csurka G., Dance C. R., Fan L., Willamowski J., Bray C., « Visual categorization with bags of keypoints », *In Workshop on Statistical Learning in Computer Vision, ECCV*, p. 1-22, 2004.

Fei-Fei L., Perona P., « A Bayesian Hierarchical Model for Learning Natural Scene Categories », *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 2, IEEE Computer Society, Washington, DC, USA, p. 524-531, 2005.

Grand-Brochier M., Tilmant C., Dhome M., « Descripteur local d'image invariant aux transformations affines », *ORASIS'09 - Congrès des jeunes chercheurs en vision par ordinateur*, Trégastel, France France, 2009.

Harris C., Stephens M., « A Combined Corner and Edge Detection », *Proceedings of The Fourth Alvey Vision Conference*, p. 147-151, 1988.

Jiang Y. G., Ngo C. W., Yang J., « Towards optimal bag-of-features for object categorization and semantic video retrieval », *CIVR '07 : Proceedings of the 6th ACM international conference on Image and video retrieval*, ACM, New York, NY, USA, p. 494-501, 2007.

Joachims T., « Text categorization with support vector machines : learning with many relevant features », *in* , C. Nédellec, , C. Rouveirol (eds), *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, Springer Verlag, Heidelberg, DE, p. 137-142, 1998.

Larlus D., Création et utilisation de vocabulaires visuels pour la catégorisation d'images et la segmentation de classes d objets, phd, 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*, IEEE Computer Society, Washington, DC, USA, p. 2169-2178, 2006.

Lowe D. G., « Object Recognition from Local Scale-Invariant Features », *ICCV*, p. 1150-1157, 1999.

Macqueen J. B., « Some Methods for classification and analysis of multivariate observations », *Procedings of the Fifth Berkeley Symposium on Math, Statistics, and Probability*, vol. 1, University of California Press, p. 281-297, 1967.

Mikolajczyk K., Tuytelaars T., Schmid C., Zisserman A., Matas J., Schaffalitzky F., Kadir T., Van Gool L., « A Comparison of Affine Region Detectors », *Int. J. Comput. Vision*, vol. 65, n° 1-2, p. 43-72, 2005.

Mikolajczyk K., Zisserman A., Schmid C., « Shape recognition with edge-based features », *Proceedings of the British Machine Vision Conference*, vol. 2, p. 779-788, 2003.

Ni D., Qu Y., Yang X., Chui Y.-P., Wong T.-T., Ho S. S. M., Heng P.-A., « Volumetric Ultrasound Panorama Based on 3D SIFT », *MICCAI (2)*, p. 52-60, 2008.

Perret D. I., Orman M. W., « Visual recognition based on temporal cortex cells : viewer-centered processing of pattern configuration », *Zeitschrift für Naturforschung*, vol. 53c, n° C, p. 518-541, 1998.

Sivic J., Zisserman A., « Video Google : A Text Retrieval Approach to Object Matching in Videos », *ICCV '03 : Proceedings of the Ninth IEEE International Conference on Computer Vision*, IEEE Computer Society, Washington, DC, USA, p. 1470-1477, 2003.

Smeulders A. W. M., Worring M., Santini S., Gupta A., Jain R., « Content-Based Image Retrieval at the End of the Early Years », *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, n° 12, p. 1349-1380, 2000.

Tabbone S., Lorraine C. C., « Corner Detection Using Laplacian of Gaussian Operator », *In SCIA 93*, p. 1055-1059, 1993.

Tanaka K., « Mechanisms of visual object recognition ; monkey and human studies », *CURRENT OPINION IN NEUROBIOLOGY*, vol. 7, n° 4, p. 523-529, 1997.

Tirilly P., Claveau V., Gros P., « Language modeling for bag-of-visual words image categorization », *CIVR*, p. 249-258, 2008.

Van de Sande K. E. A., Gevers T., Snoek C. G. M., « A Comparison of Color Features for Visual Concept Classification », *ACM International Conference on Image and Video Retrieval*, p. 141-150, 2008.

Vapnik V. N., *The nature of statistical learning theory*, Springer-Verlag New York, Inc., New York, NY, USA, 1995.

Yuan J., Wu Y., Yang M., « Discovery of Collocation Patterns : from Visual Words to Visual Phrases », *CVPR*, p. 1-8, 2007.

Yun J.-H., Park R.-H., « Self-Calibration with Two Views Using the Scale-Invariant Feature Transform », *ISVC (1)*, p. 589-598, 2006.

Zhang J., Marszalek M., Lazebnik S., Schmid C., « Local Features and Kernels for Classification of Texture and Object Categories : A Comprehensive Study », p. 13, 2006.

Zhao W., Jiang Y.-G., Ngo C.-W., « Keyframe Retrieval by Keypoints : Can Point-to-Point Matching Help ? », *CIVR*, p. 72-81, 2006.

Zheng Q.-F., Gao W., « Constructing visual phrases for effective and efficient object-based image retrieval », *TOMCCAP*, 2008a.

Zheng Q.-F., Wang W.-Q., Gao W., « Effective and efficient object-based image retrieval using visual phrases », *ACM Multimedia*, p. 77-80, 2006.

Zheng Y.-T., Zhao M., Neo S.-Y., Chua T.-S., Tian Q., « Visual Synset : Towards a higher-level visual representation », *Proc. of Conf. on Computer Vision and Pattern Recognition*, Anchorage, Alaska, U.S., 2008b.

Zhou H., Yuan Y., Shi C., « Object tracking using SIFT features and mean shift », *Computer Vision and Image Understanding*, vol. 113, n° 3, p. 345-352, March, 2009.