Segmentation de catégories d'objets par combinaison d'un modèle d'apparence et d'un champs de Markov

Diane Larlus Eric Nowak Frédéric Jurie

Projet LEAR (INPG, INRIA, CNRS)

prenom.nom@inrialpes.fr

Résumé

Nous nous intéressons ici à la segmentation de catégories d'objets dans des images. Si les modèles d'apparence par *sac-de-mots* sont ceux qui donnent à ce jour les meilleures performances en termes de classification d'image et de localisation d'objets, ils ne permettent pas de localiser précisément les frontières des objets. Cela vient du fait que les objets ne sont considérés que comme des collections non-structurées d'informations éparses. Parallèlement, les modèles basés sur des champs de Markov (MRF) utilisés pour la segmentation d'images se basent essentiellement sur les frontières et permettent une régularisation spatiale, mais utilisent difficilement des contraintes globales liées aux objets, ce qui est indispensable lorsqu'on travaille avec des catégories d'objets dont l'apparence peut varier significativement dune instance à l'autre. La principale contribution de cet article est la combinaison élégante de ces deux approches. Notre approche comporte un mécanisme basé sur la détection d'objets par sac-de-mots produisant une segmentation grossière des images, et simultanément, un second mécanisme, lui basé sur un MRF, produit des segmentations propres. Ce second mécanisme est guidé à la fois par des indices locaux de l'image (couleur, texture et arrêtes) et par des dépendances à plus large échelle, données par le modèle sac-de-mots qui renforcent la consistance entre les labels. Des expériences sur les bases Pascal VOC 2006 et Graz-02 montrent des résultats impressionnants dans le contexte difficile de la segmentation de catégories d'objets en présence de fonds encombrés et de larges changements de points de vue.

Mots Clef

Reconnaissance d'objets, segmentation

Abstract

Object models based on bag-of-words representations achieve state-of-the-art performance for image classification and object localization. However, as they consider objects as lose collection of information they fail to accurately locate object boundaries and thus produce inaccurate object segmentation. On the other hand, Markov Random Field based models used for image segmentation focus on object boundaries but can hardly use global object constraints, which is required when dealing with object categories whose appearance may vary significantly. The key contribution of this paper is to combine elegantly the advantages of these two approaches. First, a blob-based mechanism allows to detect objects using visual words occurrences, and produces rough image segmentation. Second, a MRF component produces clean cuts, guided by local image cues (color, texture and edge cues) and by long-distance dependency given by the blob model, which enforces label consistency. Gibbs sampling is used to infer the model. Experiments on Pascal VOC 2006 and Graz-02 datasets show impressive results in the difficult context of object categories segmentation in presence of cluttered backgrounds and large view point changes.

FIG. 1 – 4 exemples de segmentations obtenues par notre méthode. Les instances de catégories d'objets sont automatiquement localisées dans l'image, produisant des masques de segmentations qui peuvent être utilisés pour extraire automatiquement les objets.

Keywords

Object recognition, segmentation

1 Introduction

Cet article s'intéresse au problème de la création de segmentations précises et propres de classes d'objets dans les images, sans aucune information a priori sur l'orientation, la position et l'échelle des objets dans les images.

La segmentation d'image a été largement étudiée dans un grand nombre de travaux pendant ces dernières années. Beaucoup d'approches ont été proposées, combinant différentes propriétés d'images, comme la couleur, la texture,

FIG. 2 – Images représentatives de 2 catégories différentes de la base Pascal VOC2006 (chats et personnes). La segmentation de ces objets, sans connaitre leur position et malgré un changement d'échelle et une pose arbitraire, représente une tache difficile.

les contours ou encore le mouvement ... (voir [3]), de façon non supervisée. Cependant, obtenir une segmentation précise en utilisant uniquement des processus ascendants (*bottom-up*) est difficile : la segmentation d'une image est en effet intimement liée à sa compréhension, ce qui en fait un problème complexe.

Grâce aux récentes avancées en représentation d'images, en détection d'objets et en technique d'apprentissage, il est maintenant possible d'entrainer des algorithmes capables de reconnaître, localiser et segmenter les objets simultanément.

La figure 1 donne une illustration du type de problème qui nous intéresse ici ainsi que des exemples de résultats obtenus avec notre algorithme. A partir d'images encombrées contenant des objets d'intérêt, la méthode est capable de localiser ces objets et de produire automatiquement des masques de segmentation qui peuvent être utilisés par la suite pour extraire l'objet.

Le problème considéré ici est le problème de la segmentation d'objets appartenant des catégories connues (*figure-ground segmentation*), en supposant que les catégories sont définies par un ensemble d'images d'apprentissage[1] utilisées pour apprendre des modèle d'apparence d'objets. Dans ces conditions, la segmentation d'objets est intimement liée à la reconnaissance et à la détection d'objets.

Cet article s'intéresse à des images difficiles, en condition réelles, où les objets peuvent avoir des apparences très différentes, et apparaissent dans l'image à n'importe quelle taille ou position.

Le reste de cet article est organisé comme suit : nous verrons d'abord l'état de l'art ainsi qu'une présentation rapide de la méthode proposée. Après une courte description des bases d'images considérées, nous présenterons notre modèle ainsi que les méthodes utilisées pour son estimation. Enfin, nous étudierons les résultats expérimentaux, et présenterons les conclusions de cette étude.

[1] Il est a noté que la segmentation d'images et d'objets sont deux problèmes différents. Dans le cadre de la segmentation d'images, tout doit être segmenté, alors que dans la segmentation d'objet, seul les objets d'intérêt sont à segmenter.

1.1 Etat de l'art

Des segmentations d'objets de grande qualité ont été obtenues récemment par différents auteurs [10, 7] dans un contexte où la position des objets est supposée interactivement définie. L'idée clef est que le premier plan et le fond sont décrits à l'aide de distribution de couleurs, estimées itérativement lors d'une minimisation d'énergie et appliquée à un découpage de graphe. L'image est considérée comme un graphe sur les valeurs de couleurs, modélisé à l'aide d'un champ de Markov (*Markov Random Field* ou *MRF*). L'énergie totale du champ de labels objet-fond dépend de

– la similarité entre les pixels voisins qui ont des labels différents,
– la probabilité des couleurs de pixels connaissant les modèles de couleurs globaux objet-fond.

Les MRF et leurs variantes (CRF [13], DRF[12]) ont une longue histoire liée à la segmentation d'image. Un des avantages majeur des MRF est la *régularisation*. Les labels de deux pixels voisins sont corrélés et quand l'évidence locale d'un label est faible, les labels du voisinage peuvent être d'une grande aide. Shotton *et al* [13] ont utilisé un CRF amélioré basé sur un ordre spatial entre les parties d'objets pour gérer les occultations.

Cependant, les segmentations obtenues avec un MRF sans aucun modèle de forme produisent rarement des segmentations réalistes, c'est pourquoi plusieurs auteurs ont tenté de fusionner ces deux concepts. Citons par exemple Kumar *et al* [4] qui proposent une méthodologie pour combiner un CRF et un modèle *pictoriel* de structure.

Liebe and Schiele [5] utilisent des images segmentées à la main pour apprendre des masques de segmentation correspondant aux mots du vocabulaire visuel. Ensuite, un modèle implicite de forme permet de localiser les objets et de segmenter l'image en combinant des masques de segmentation locaux correspondant aux entrées du vocabulaire visuel. D'autres articles majeurs [1, 6, 12, 16] proposent différentes manières d'utiliser un modèle de forme pour la tache de segmentation. Cependant, les hypothèses géométriques simples qui sont faites sur ces modèles ne

permettent pas d'appréhender les objets d'apparence complexe ou faiblement structurés.

Enfin, il a été montré récemment [2] que les modèles considérant les images comme des ensembles de mots visuels (modèle très populaire pour la classification d'image) peuvent être appliqués avec succès à la localisation de classes d'objets dans les images. Ce type de méthode est particulièrement adapté à nos besoins puisqu'il permet de gérer de très fortes variations d'apparence.

Ce modèle peut être également combiné à un processus de Dirichlet, permettant de produire des clusters de localisation spatiale [14]. Malheureusement, la forme des objets est très mal définie par ces modèles.

Pour finir, seulement un petit nombre de ces méthodes sont capables de produire des segmentations précises d'objets ayant une grande variété d'apparences. Cela laisse une large place à leur amélioration, spécialement quand le fond est trop riche ou trop encombré pour être modélisé.

1.2 Présentation succincte de l'approche

La contribution principale de cet article est un modèle pour la segmentation d'objet, qui tire parti de deux composants complémentaires.

- un modèle aux propriétés de MRF pour sa capacité à produire des champs de labels localement cohérents ainsi qu'une segmentation qui s'adapte aux frontières bas niveau de l'image
- un modèle de type *sac-de-mots* qui permet la reconnaissance et la localisation des objets malgré de fortes variations de point de vue et qui assure une cohérence globale des informations visuelles.

Les frontières d'objets sont définies localement, mais les structures globales (comme les classes d'objet), qui sont primordiales dans la sémantique de l'image, assurent la cohérence de ces informations locales.

2 Description du modèle

Comme présenté dans l'introduction, la force de notre modèle repose sur la combinaison de deux composants différents mais complémentaires : un modèle génératif de "blobs" utilisant des mots visuels et permettant une bonne localisation de l'objet (mais grossière) et une structure en champ de Markov (MRF) qui permet d'avoir des champs de labels cohérents et qui suivent les contours de l'objet.

L'équation 1 permet d'avoir un bon aperçu de la combinaison des différents composants, plus de détails seront fournis plus tard. La segmentation consiste en l'affectation de patchs à des blobs (groupes de patchs). La distribution conjointe de toutes les affectations de patchs b est donnée par : $p(b) \propto p_{mrf}^{-\gamma}(b) p_{dp}(b)$ où p_{mrf} est la probabilité donnée par le MRF, p_{dp} celle donnée par le processus de Dirichlet et γ permet d'équilibrer le poids entre ces deux probabilités. Si on ajoute l'information sur les patchs $\mathcal{P}_i$, l'équation suivante (voir l'illustration Fig. 3) est obtenue :

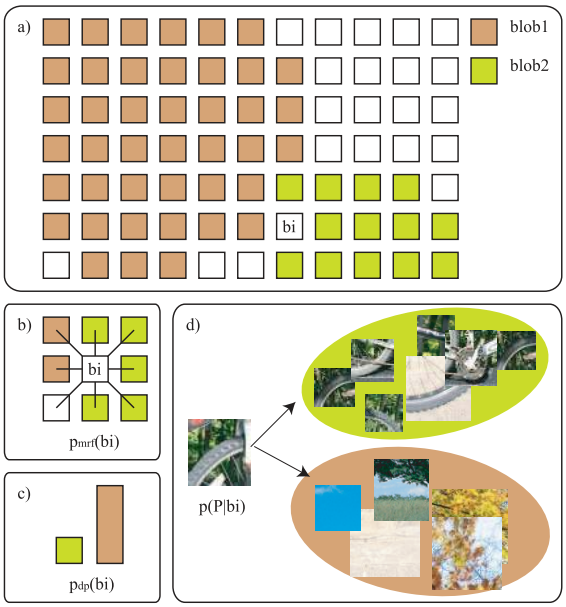

FIG. 3 – a) Le modèle calcule les meilleures affectations de patchs à chaque blob. La probabilité de chaque affectation b_i (connaissant le patch) est le produit de 3 probabilités provenant de b) le champ de Markov (MRF) c) le processus de Dirichlet d) le modèle d'objet. Voir le texte pour de plus amples explications.

$$p(b|\mathcal{P}_i) \propto \underbrace{p_{mrf}^{-\gamma}(b)}_{MRF} \quad \underbrace{p_{dp}(b)}_{Dirich.Proc.} \quad \underbrace{p(\mathcal{P}_i|b)}_{ObjectModel} \tag{1}$$

Notre modèle est complètement spécifié par la probabilité conditionnelle de chacun de ces paramètres, ce qui permet d'échantillonner la valeur de ces paramètres suivant leur loi jointe grâce à un échantillonneur de Gibbs. Cette section décrit les deux composants : le modèle basé sur des blobs et la structure du champ de Markov, puis il détaille l'estimation des paramètres.

2.1 Un modèle génératif de blobs

Cette partie spécifie un modèle génératif adapté pour une segmentation objet/fond grossière. Notre modèle s'inspire de [14] qui utilise des informations de structures spatiales explicites : considérons qu'une image est constituée de blobs et que chaque blob génère une partie des patchs selon son propre modèle. Intuitivement, si une image contient 3 objets (une voiture, un piéton et un vélo) nous pouvons obtenir 3 blobs, chacun couvrant une région de l'image. Chaque blob est ensuite responsable de la génération des pixels de l'image qui sont dans sa région, par exemple en générant un ensemble de patchs dont l'apparence correspond à une catégorie d'objet (des patchs de voiture pour le blob de voiture, etc.) Ceci renforce la cohérence spatiale des patchs générés au sein de la région du blob.
La génération d'un patch nécessite de a) sélectionner un blob et b) générer un patch selon le modèle de patch *spéci-*

fique à ce blob. Le reste de cette section détaille la probabilité de sélectionner un blob et de générer un patch connaissant le blob.
La génération des blobs est supposée suivre un processus de Dirichlet. Le processus de Dirichlet possède une propriété d'auto-renforcement : plus une valeur a été échantillonnée par le passé, plus sa probabilité d'être générée une nouvelle fois augmente.
Nous considérons le processus de Dirichlet comme un modèle de mixtures avec K composantes [2] pour rendre l'échantillonnage plus aisé [9]. Cela signifie en pratique que pour chaque nouveau patch généré, il peut soit appartenir a un blob B_k déjà généré avec une probabilité $\frac{N_k+\alpha/K}{n-1+\alpha}$ où N_k est sa population, soit il peut créer une nouvelle région avec une probabilité $\frac{\alpha}{n-1+\alpha}$, avec α le paramètre de concentration du processus de Dirichlet et n le nombre de patchs.
Chaque blob $B_{k,1\leq k\leq K}$ est caractérisé par un ensemble de variables aléatoires : $\Theta_k = \{\mu_k, \Sigma_k, C_k, l_k, N_k\}$.
μ_k, Σ_k représentent la moyenne et la matrice de covariance décrivant la forme géométrique des blobs, l_k est le label du blob (la catégorie d'objet), C_k est le modèle de mixtures de gaussienne représentant les couleurs de chaque blob et N_k et le nombre de patchs générés par le blob.
Chaque patch $\mathcal{P}_i$ est caractérisé par les descripteurs $(w_i^{sift}, w_i^{color}, rgb_i, X_i)$ mais aussi par deux autres variables aléatoires b_i et c_i. b_i est l'index du blob qui a généré le patch ($1 \leq b_i \leq K$) et c_i la composante de la mixture de couleurs à laquelle le patch est affecté (ceci serai détaillé plus tard).
Définissons alors la probabilité de générer un patch, sachant qu'il est généré par le blob B_k de paramètres Θ_k : $p(\mathcal{P}|\Theta_k)$. Cette probabilité est composée de 4 parties différentes puisque le modèle suppose que la position, la couleur et l'apparence des patchs sont indépendantes, sachant le blob qui l'a généré.

$$\begin{aligned} p(\mathcal{P}|\Theta_k) &= p(w^{sift}, w^{color}, rgb, X|\Theta_k) \\ &= p(w^{sift}|\Theta_k)p(w^{color}|\Theta_k)p(rgb|\Theta_k)p(X|\Theta_k) \end{aligned} \quad (2)$$

La position du patch X est choisie selon une distribution normale de paramètres μ_k et Σ_k pour les blobs d'objets. La distribution est uniforme pour les blobs de fond. Et pour les blobs d'objet

$$p(X|\Theta_k, l_k \neq \text{fond}) = \mathcal{N}(X, \mu_k, \Sigma_k) \quad (3)$$

Nous supposons que les blobs de fond ont une distribution de couleur uniforme et que les blobs d'objet ont un modèle de couleur modélisé par une mixture de gaussiennes (GMM), comme suggéré par [10]. Nous utilisons 5 composantes dans nos expériences. Chaque patch du blob est généré par un composant unique de GMM, et ceci est représenté par la variable c_i introduite précédemment. Le fait de supposer que les patchs sont générés par un seul composant rend le calcul des GMM plus facile, mais une affectation partagée aurait aussi été possible.
Finalement, la probabilité des mots visuels SIFT et couleurs ne dépend que du label de classe, c'est-à-dire de : $p(\omega^{sift}|\Theta_k)$=$p(\omega^{sift}|l_k)$ et $p(\omega^{color}|\Theta_k)$=$p(\omega^{color}|l_k)$.
Ces distributions représentent les informations connues sur l'apparence des objets et sont responsables des capacités de reconnaissance de notre modèle. Elles sont apprises à partir d'images d'entraînement annotées dans lesquelles les mots visuels sont extraits. Les distributions sont alors estimées par un processus de comptage.

[2] K peut aller vers l'infini bien qu'en pratique le nombre fini de patchs impose que K soit fini

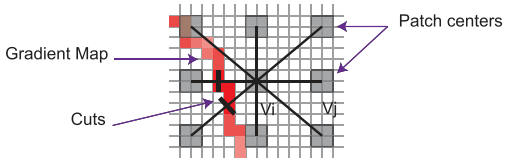

FIG. 4 – La structure du MRF régularise le champ de labels objet/fond et aligne le découpage sur les contours naturels de l'image.

2.2 Une structure de champs de Markov pour l'affectation aux blobs

L'affectation des patchs aux blobs *objet* ou *fond* détermine la segmentation de l'image. Cette segmentation est améliorée par notre deuxième composant qui régularise les affectations de patchs voisins et qui aligne le découpage avec les contrastes naturels de l'image. Ce champ est défini par rapport à une grille (8-connectivité) qui correspond au centre des patchs.
Ce composant définit une énergie de Gibbs qui est utilisée pour calculer la probabilité conditionnelle d'affectation des patchs. Cette énergie à un terme d'ajustement au modèle basé sur la représentation en blobs présentée précédemment ainsi qu'un terme basé sur des contraintes de voisinage pour la régularisation spatiale et l'adaptation aux contrastes de l'image.
L'énergie totale E du champ total est donnée comme la somme des énergies E_i définie pour chaque patch $\mathcal{P}_i$.

$$E_i = U_i + \gamma \sum_{j\in\mathcal{N}(i)} V_{i,j} \quad (4)$$

où $\mathcal{N}(i)$ représente les voisins de $\mathcal{P}_i$, γ pondère la proportion des deux termes et

$$U_i = -\log p(b_i|\mathcal{P}_i, N_{1:K}, \Theta_{b_i}) \quad (5)$$

est un potentiel qui mesure la cohérence entre le patch et le modèle de blob. $p(b_i|\mathcal{P}_i, N_{1:K}, \Theta_{b_i})$ est la probabilité d'affectation aux blobs sachant le patch et les paramètre des blobs. Elle dérive du modèle présenté dans le paragraphe précédent et fait le lien entre les deux composants du modèle. Les détails de cette dérivation seront donnés section 2.3.

$V_{i,j}$ est défini par

$$V_{i,j} = [l_{b_i} \neq l_{b_j}] \exp(-\beta \Phi(Xi, Xj, \mathcal{G})), \qquad (6)$$

où $[.]$ est la fonction indicatrice. $V_{i,j}$ est un potentiel qui force la cohérence locale des labels objet-fond à partir de contraintes de similarité entre les labels de patchs voisins, et encourage également le découpage le long des gradients de l'image via la fonction Φ. $\Phi(Xi, Xj, \mathcal{G})$ est la valeur maximale du gradient entre X_i et X_j (qui sont les positions respectives des patchs $\mathcal{P}_i$ et $\mathcal{P}_j$) et β est une constante calculée de la même façon que dans [10] (voir Fig.4 pour une illustration).
Ainsi, $V_{i,j}$ est nul si les patchs ont des labels similaire et sinon il pénalise davantage les patchs qui ont des labels différents sans que des contours ne les séparent. En effet, nous souhaitons que le modèle sépare entre l'objet et le fond principalement le long des contours de l'image.

2.3 Estimation du modèle

Le modèle étant défini, tous ces paramètres doivent être estimés pour chaque image de façon à un produire les blobs (l_i) et le champ d'affectations des patchs à ces blobs (b_i).
Un échantillonneur de Gibbs génère des valeurs pour les paramètres. Pour cela, chaque variable est successivement échantillonnée à partir de sa distribution conditionnelle à la valeur courante des autres variables.
Cette section définit la distribution conditionnelle sur chaque variable ainsi que la façon de l'échantillonner. L'ensemble de paramètre à estimer est le suivant :

$$\Theta = \{\mu_{1:K}, \Sigma_{1:K}, C_{1:K}, l_{1:k}, b_{1:n}, c_{1:n}\} \qquad (7)$$

Observations. Certaines des variables peuvent être directement calculées à partir des images et ne sont pas conditionnées par d'autres variables, en particulier la carte de gradient $\mathcal{G}$ et la description des n patchs : $(w_i^{sift}, w_i^{color}, rgb_i, X_i)$.
Échantillonnage des paramètres du blob. Le premier paramètre du blob est μ_k. Si on appelle X'_i la position du i^{me} patch du blob B_k, alors

$$\begin{array}{lll} p(\mu_k|\Theta \setminus \{\mu_k\}) & = & p(\mu_k|X_{1:n}, b_{1:n}) \\ & = & p(\mu_k|X'_{1:N_k}) \\ \mu_k \sim \mathcal{N}(\mu, & Mean(X'_{1:N_K}) & , \frac{1}{N_k} Cov(X'_{1:N_K})) \end{array} \qquad (8)$$

Le deuxième paramètre est Σ_k. De la même façon, si W_p désigne une distribution de Wishart,

$$\begin{array}{lll} p(\Sigma_k|\Theta \setminus \{\Sigma_k\}) & = & p(\Sigma_k|\mu_k, X'_{1:N_K}) \\ \Sigma_k & \sim & W_p(Cov(X'_{1:N_K}), N_k - 1) \end{array} \qquad (9)$$

Le troisième paramètre du blob est la mixture gaussienne pour la couleur, qui est estimée simplement par application de l'algorithme EM-stochastique, chaque mixture étant elle-même composée de nc composants (5 dans notre cas).

$$C_k = \sum_{j=1}^{nc} \alpha_{j,k} \mathcal{N}(RBG, \mu_{j,k}^{RGB}, \Sigma_{j,k}^{RGB}) \qquad (10)$$

Le dernier paramètre du blob est le label de classe l_k, échantillonné par :

$$\begin{array}{lll} p(l_k|\Theta \setminus \{l_k\}) & = & p(l_k|\mathcal{P}'_{1:N_k}, \Theta_k) \\ & \propto & \prod_{i=1}^{N_k} p(w_k'^{sift}|l_k) p(w_k'^{color}|l_k) \end{array} \qquad (11)$$

par hypothèse d'indépendance des patchs, sachant le blob qui les a générés.
Échantillonnage des paramètres du patch. c_i est le composant de la mixture couleur affectée à un patch, caractérisé par $RBG = rbg_i$ et $b_i = k$, il est calculé par

$$c_i = \arg\max_j \alpha_{j,k} \mathcal{N}(rgb, \mu_{j,k}^{RGB}, \Sigma_{j,k}^{RGB}) \qquad (12)$$

Et pour finir, l'estimation conditionnelle des appartenances au blob b_i. Le théorème d'Hammersley-Clifford garantie la validité des expressions suivantes.

$$p(b_i|b_{j\neq i}, \Theta \setminus \{b_{1:n}\}) = \frac{1}{Z_i} \exp - \left(U_i + \gamma \sum_{j \in \mathcal{N}(i)} V_{i,j} \right) \qquad (13)$$

où Z_i est la fonction de partition, $\mathcal{N}(i)$ est l'ensemble de patchs dans le voisinage de $\mathcal{P}_i$. V est défini dans l'équation équation 6, et U est détaillée ci-dessous.

$$\begin{array}{lll} U_i(\Theta \setminus \{b_k, k \neq i, k \notin \mathcal{N}(i)\}) & = & U_i(b_i, \mathcal{P}_i, N_{1:K}, \Theta_{b_i}) \\ & = & -\log p(b_i|\mathcal{P}_i, N_{1:K}, \Theta_{b_i}) \\ p(b_i|\mathcal{P}_i, N_{1:K}, \Theta_{b_i}) & \propto & p(\mathcal{P}_i|b_i, \Theta_{b_i}) / p(b_i|N_{1:K}) \\ & \propto & p(\mathcal{P}_i|\Theta_{b_i}) \frac{N_{b_i} + \alpha/K}{n - 1 + \alpha} \end{array} \qquad (14)$$

3 Expérimentations

3.1 Détails de l'implémentation

L'ensemble d'apprentissage est constitué de 150 images par catégories pour la base Graz et de 50 images par catégories pour la base Pascal. Il est bon de rappeler que seules les images de Graz disposent de masques de segmentations précis, alors que les objets de la base Pascal sont annotés seulement par une région d'intérêt rectangulaire, et ainsi les modèles d'objets contiennent des informations sur le fond.
Les patchs sont extraits à échelle fixe. Le paramètre de chevauchement des patchs est tel que l'image est divisée en 3000 patchs et chaque pixel est inclus dans au moins 25 patchs.
Le vocabulaire visuel SIFT [8] contient 5000 éléments et le vocabulaire couleur [15] contient 100 éléments. Ils sont calculés dans les deux cas par un algorithme k-means appliqués aux descripteurs des images d'apprentissage.

Modalités	SC	SP	CP	SCP	SCPR
Image 1	80.6%	70.8%	22.9%	89.6%	89.0%
Image 2	79.4%	81.5%	82.7%	90.3%	88.7%
Image 3	83.7%	83.0%	79.3%	83.1%	92.2%
Image 4	82.1%	86.4%	78.9%	87.3%	88.0%

TAB. 1 – EER (Equal Error Rate) de la courbe précision rappel sur les images de Graz pour différentes modalités de représentions des patchs : combinaison des mots visuels SIFTS (S), des mots visuels couleur (C), de la couleur RGB (R) et de la position (P)

Le paramètre γ qui spécifie le compromis entre l'adéquation au modèle de blob est les contraintes de cohérence spatiale est fixé à 10, et le paramètre de Dirichlet α est fixé à 0,5.
Les masques de segmentations présentés dans cette section sont des masques à l'échelle du pixel, obtenus par interpolation des labels de segmentations, à l'échelle des patchs. Les labels des pixels sont obtenus par un modèle de mixture où les poids sont proportionnels à la distance entre la position du pixel et le centre du patch.

3.2 Évaluation des performances

Le choix de la mesure de performance est important puisque c'est lui qui révèle le comportement de la méthode. Nous avons décidé d'utiliser les courbes de précision-rappel et non des courbes ROC ou du taux de bonne classification des pixels.
Le taux de bonne classification des pixels (nombre de pixels d'objet ou de fond correctement classifiés) est biaisé par la taille des objets dans l'image : quand tous les pixels prédits sont de type fond, la précision tend vers 100% quand la taille des objets décroit.
Les courbes ROC ne sont pas non plus adaptées à nos besoin, car les performances décroissent dramatiquement lorsque les frontières d'objets prédites sont à l'intérieur de l'objet plutôt qu'à l'extérieur, en raison du rapport déséquilibré entre les ensembles de pixels d'objet et ceux du fond. La courbe précision-rappel ne souffre pas de ce biais, puisqu'elle évalue vraiment la précision de la segmentation : quand la frontière prédite est à l'intérieur de l'objet, le rappel décroît, autant que la précision baisse lorsque la frontière est à l'extérieur de l'objet.
Ainsi les différents composants de notre système seront évalués par des courbes précision rappel.

3.3 Représentation multimodale de patchs

Les patchs ont une représentation multimodale. Ils sont en effet représentés par des mots visuels SIFT, couleur, une valeur de couleur RGB et une position. Cette partie étudie l'importance de ces différentes représentations. Les résultats sont résumés dans le tableau 1. Trois conclusions peuvent être tirées de ce tableau. Tout d'abord combiner les SIFT et la couleur est meilleur que d'utiliser le SIFT

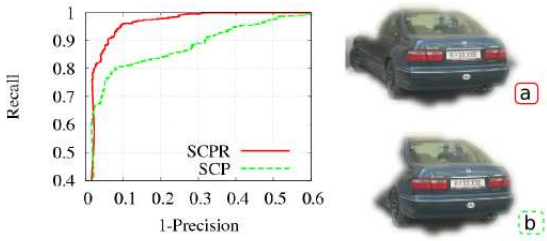

FIG. 5 – Notre modèle, avec et sans le modèle de couleur par blob. Gauche : courbe précision rappel. Droite : les images correspondantes, en haut (en bas) : avec (sans) la mixture de couleur

MRF	avec MRF	sans MRF	Gain
Image 1	89.6%	84.4%	5.2%
Image 2	90.3%	85.0%	5.3%
Image 3	83.1%	83.5%	-0.4%
Image 4	87.3%	82.6%	4.7%

TAB. 2 – Equal Error Rate de précision rappel pour le modèle complet avec ou sans le MRF.

ou la couleur seuls (SCP>SP and SCP>CP). Le gain varie entre 0 et 20%. Deuxièmement, utiliser la position est réellement utile (SCP>SC), car elle fournit des contraintes sur la cohérence spatiale et cela peut donner jusqu'à 10% d'amélioration. Troisièmement, la mixture de couleur des blobs améliore en général les résultats (SCPR>SCP), jusqu'à 10%. Cela peut être expliqué par le fait qu'une partie d'objet dont l'apparence n'est pas discriminante mais dont la couleur est cohérente avec l'ensemble de l'objet est intégrée au modèle d'objet. L'influence de la mixture de couleur est illustrée sur la figure 5.

3.4 Influence du MRF

Notre modèle génératif d'images combine un modèle de blob et un champ de Markov pour la régularisation de la segmentation. En effet le MRF pousse le modèle à suivre les frontières, ce qui améliore la précision de la segmentation et donc la valeur de l'Equal Error Rate. En général le gain est d'environ 5%. La segmentation de la troisième image est détériorée par le MRF parce que le contour de l'objet est faible par rapport à un contour proche qui est plus fort, et les patchs juste derrière la frontière n'ont pas d'apparence discriminante. Dans ces conditions, le contour fort agit comme un attracteur pour le MRF.

3.5 Résultat qualitatif

La figure 6 illustre nos motivations à intégrer un processus de Dirichlet dans le modèle de génération des blobs. Le modèle a estimé que cette image était mieux décrite avec deux blobs d'objet, et qu'une configuration avec un

FIG. 6 – Le processus de Dirichlet appartenant au modèle de l'image a produit deux blobs d'objet différents, configuration plus probable qu'un seul.

seul blob serait moins probable. En effet, deux blobs permettent d'avoir deux modèles de couleurs spécifiques ce qui est plus précis qu'un modèle commun.
La figure 7 montre des exemples d'images segmentées par notre algorithme sur les bases Graz et Pascal. Trois images et leurs masques de segmentation associés sont présentés pour chaque catégorie. Les personnes, les vélos et les voitures ont des segmentations beaucoup plus propres que les autres puisqu'ils ont été appris sur des masques de segmentations précis. La figure montre que nous avons pu segmenter différentes catégories apparaissant avec une grande variété de poses et d'apparences, et différentes formes globales. Même de tout petits objets sont correctement segmentés lorsque les masques d'apprentissage sont précis et que le fond n'est pas trop encombré.

3.6 Comparaison avec d'autres travaux

La plupart des travaux précédents se basent sur des modèles de forme qui améliorent la précision des segmentations [1, 4, 5, 6, 12, 16]. De telles méthodes ne peuvent être appliquées dans notre contexte en raison de la diversité des apparences des objets dans les images. Les méthodes décrites dans [11, 13] sont très similaires aux nôtres. Cependant, leurs résultats ne sont pas comparables puisque des catégories de fonds spécifiques sont apprises (herbe, eau, route, ciel ...) alors que notre méthode a été conçue dans le but de fonctionner avec un fond générique. De plus, leur évaluation multi-classe ne peut pas être utilisée avec notre classe de fond générique.

4 conclusion

Dans cet article, nous avons présenté une nouvelle méthode pour la segmentation de catégories d'objet. L'élément clef qui distingue cette méthode des précédentes est la combinaison, dans le même modèle, de deux composants complémentaires. Tout d'abord, un composant basé sur les blobs permet de détecter les objets en utilisant les occurrences de mots visuels. Il en résulte une segmentation approximative, séparant grossièrement les différents composants de l'image. Ensuite, un composant basé sur un MRF produit un découpage propre, guidé par les contours d'intensité et de textures de l'image. Un échantillonneur de Gibbs permet d'estimer efficacement les paramètres du modèle produit. Des expériences conduites sur les bases Pascal VOC 2006 et Graz montrent des résultats impressionnants dans un contexte difficile où les fonds sont encombrés et les objets présentent des points de vues très diverses.

Références

[1] E. Borenstein and J. Malik. Shape guided object segmentation. In *CVPR'06*, pages 969–976, 2006.

[2] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google"s image search. In *ICCV '05*, pages 1816–1823, 2005.

[3] R.M. Haralick and L.G. Shapiro. Image segmentation techniques. *Computer Vision, Graphics, and Image Processing*, 29 :100–132, 1985.

[4] M. P. Kumar, P. H. S. Torr, and A. Zisserman. OBJ CUT. In *CVPR'05*, 2005.

[5] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *BMVC'03*, 2003.

[6] A. Levin and Y. Weiss. Learning to combine bottom-up and top-down segmentation. In *ECCV06*, pages IV : 581–594, 2006.

[7] Y. Li, J. Sun, C.K. Tang, and H.Y. Shum. Lazy snapping. *ACM Trans. Graph.*, 23(3) :303–308, 2004.

[8] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2) :91–110, 2004.

[9] Radford M. Neal. Markov chain sampling methods for dirichlet process mixture models. Technical Report 9815, Dept. of Statistics, University of Toronto, Sep 1998.

[10] C. Rother, V. Kolmogorov, and A. Blake. "grabcut" : interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3) :309–314, 2004.

[11] F. Schroff, A. Criminisi, and A. Zisserman. Single-histogram class models for image segmentation. In *ICCVGIP06*, pages 82–93, 2006.

[12] J. Shotton, A. Blake, and R. Cipolla. Contour-based learning for object detection. In *ICCV '05*, pages I :503–510, 2005.

[13] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost : Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV06*, pages I : 1–15, 2006.

[14] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Describing visual scenes using transformed dirichlet processes. In *NIPS'05*, 2005.

[15] J. van de Weijer and C. Cordelia Schmid. Coloring local feature extraction. In *ECCV'06*, pages 334–348, 2006.

[16] J. Winn and N. Jojic. Locus : Learning object classes with unsupervised segmentation. In *ICCV '05*, pages 756–763, 2005.

FIG. 7 – Exemples d'images et leur masque de segmentation respectifs(la couleur représente le label de classe) pour les 10 catégories d'objet. Les voitures, les vélos et les personnes sont mieux segmentées que les 7 autres catégories puisque des masques de segmentation précis ont été utilisés pour l'apprentissage. (des masques des segmentation précis sont disponibles pour les 3 catégories de la base Graz)