Object class segmentation using boosted conditional random fields

Interprétation d'images par champs conditionnels aléatoires boostés

Xi Li
Hichem Sahbi

2010D015

Juillet 2010

Département Traitement du Signal et des Images
Groupe TII : Traitement et Interprétation des Images

Object Class Segmentation Using Boosted Conditional Random Fields

Interprétation d'Images par Champs Conditionnels Aléatoires Boostés

Xi Li XI.LI@TELECOM-PARISTECH.FR
CNRS LTCI UMR 5141
Telecom ParisTech, Paris, France

Hichem Sahbi HICHEM.SAHBI@TELECOM-PARISTECH.FR
CNRS LTCI UMR 5141
Telecom ParisTech, Paris, France

Abstract

Object class segmentation (OCS) is a key issue in semantic scene labeling and understanding. Its general principle consists of naming object entities into scenes according to their intrinsic visual features as well as their dependencies. In this paper, we propose a novel superpixel-based framework for object class segmentation using conditional random fields (CRFs). The framework proceeds in two steps: (i) superpixel label estimate; and (ii) CRF label propagation. Step (i) is achieved using multi-scale boosted classifiers over superpixels and makes it possible to find coarse estimates of initial labels. Fine labeling is afterward achieved in Step (ii), using an anisotropic contrast sensitive pairwise function designed in order to characterize the intrinsic interaction potentials between objects according to 4-neighborhoods. Finally, a higher-order criterion is applied to enforce region label consistency of OCS. Experimental results demonstrate the effectiveness of the proposed framework.

Résumé

La segmentation d'images en objets est une étape importante dans le processus d'étiquetage et d'annotation d'images. Son principe général consiste à nommer des entités d'objets dans les scènes en fonction de leurs propriétés visuelles ainsi que leurs dépendances. Dans cet article, on propose ne nouvelle approche de segmentation et d'annotation d'images basée sur les champs conditionnels aléatoires. L'approche procède en deux étapes: (i) extraction et étiquetage des superpixels, et (ii) propagation des étiquettes. L'étape (i) est effectuée en utilisant des classifieurs multi-échelles boostés, et permet d'obtenir une estimation grossière des étiquettes. L'étape (ii) permet de raffiner ces étiquettes en utilisant une fonction anisotrope sensible aux changements de contraste et permet ainsi de caractériser les interactions entre objets selon quatre directions préférées. Enfin, un critère d'ordre supérieur est appliqué afin de renforcer des étiquetages sur certaines régions plutôt que d'autres. Les résultats expérimentaux démontrent clairement les bonnes performances de l'approche proposée.

Keywords: Conditional random fields, scene annotation and understanding, image segmentation, boosting, statistical learning.

1. Introduction

Object class segmentation (OCS) is a fundamental and challenging problem in computer vision. It has many potential applications including object recognition and image retrieval.

Most of the OCS methods are considered as labelling problems working either on individual pixels or on constellations of spatially homogeneous ones, referred to as *superpixels*. Accordingly, state of the art methods may be categorized depending on their partitioning strategies; some of them perform label prediction directly at the pixel level (1) (as the finest partitioning), while others rely on superpixels (2; 3; 4; 5), and variants of them using grouping and intersection (6; 7).

The aforementioned two kinds of methods have their own advantages and weaknesses. In general, pixel-based OCSs are capable of accurately characterizing object boundaries. However, the limited amount of information (color), contained into pixels, is not enough in order to determine their corresponding object labels precisely. Furthermore, computational cost of pixel-based OCS is very expensive. In comparison, superpixel-based OCSs are able to capture rich shape and texture contextual informations, which are crucial for object label prediction. Moreover, they are computationally efficient, which is very important for practical applications. Nevertheless and resulting from their coarseness, superpixel-based methods are not suitable in order to delimit object boundaries with high precision. In order to address this issue, existing methods are based on tree-structured graphical models for multi-scale image segmentation, and make belief propagation reasoning between parent and child nodes (8).

Considering superpixel-based OCS approaches, two sub-categories of methods may be found. The first one is based on the use of irregular superpixel lattices (9), while the other considers instead regular ones(10). Based on spatial location (and also color, and texture distributions), methods using irregular lattices use unsupervised segmentations (11; 12) as initialization. However, segmentation performances (11; 12) are known to be highly dependent on the setting of the underlying parameters. In contrast, methods based on regular lattices consider image patches, generated from uniform image subdivisions. Consequently, they are very easy-to-set and computationally efficient with no need of a priori segmentation. Moreover, they are convenient in order to capture multi-scale object informations by adjusting dimensions of image patches.

In addition, another key issue of OCS is how to model pixel dependencies. Previous work on this aspect mainly uses Conditional Random Field (CRF) models in order to learn the conditional distribution over the class labeling. In general, the CRF model can be decomposed into unary and interaction potential functions. The former measures the likelihood of a pixel belonging to a particular class, while the latter encodes the dependency information which enforces label consistency among neighboring pixels, resulting into a shrinkage bias (9). The most widely used interaction potential function is formulated as a pairwise

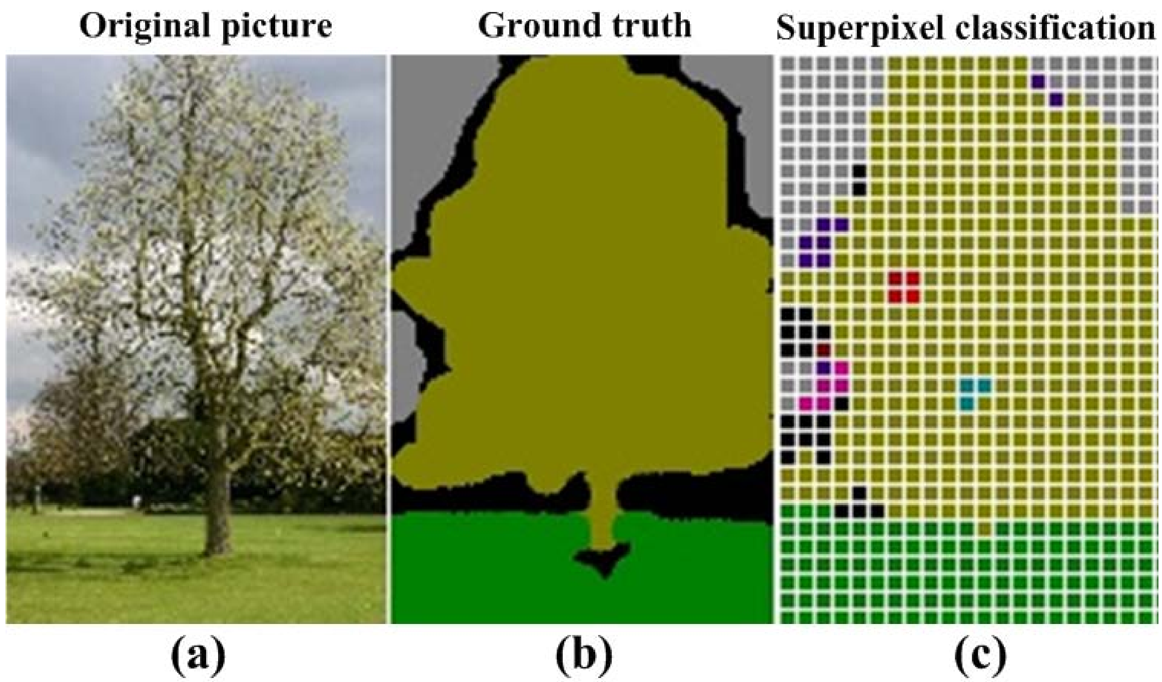

Figure 1: **Illustration of superpixel classification.**

one (1). But the pairwise CRF formulation suffers from its weakness in characterizing high-order dependencies between pixels. To address this problem, Kohli *et al.* (9) propose a robust P^N model, which is a novel family of higher-order potentials containing the P^n Potts model as well as its robust variants. However, the robust P^N model heavily relies on *a priori* mean shift-based object segmentation whose computational cost is known to be expensive, especially for high-resolution images. Besides, different parameter configurations, for the used kernels in the mean shift algorithm, may produce different segmentation results. Following the previous statements, it is impractical to directly use pixel-based object segmentations in order to construct higher-order interaction potential functions. Instead, simple and efficient superpixel-based segmentation turns to be more feasible.

In this paper, we propose a novel framework for object-driven image segmentation and labeling. The framework takes advantage of a higher-order conditional random field (CRF) in order to capture the spatial contextual informations of superpixels. The main contribution of the method is three-fold. First, an anisotropic contrast sensitive pairwise potential function is designed in order to characterize dependencies between superpixels in the 4-neighboring directions, resulting into a more accurate segmentation. Second, a higher-order region consistency criterion is introduced in order to produce smooth segmentations. Finally, further post-processing operations (including region merging) are also considered in order to refine segmentation results.

2. Conditional random field

Given N lattice points $\mathcal{V} = \{1, \ldots, N\}$, a discrete random field $\mathbf{X} = \{X_1, \ldots, X_N\}$ is defined as a set of random variables. Each lattice point $i \in \mathcal{V}$ corresponds to the random variable X_i which takes a value from the label set $\mathcal{L} = \{1, \ldots, K\}$. A clique $\mathbf{c}$ is defined as a collection of conditionally independent random variables, indexed by c as $\mathbf{X}_c$. Let $\mathbf{x} = \{x_1, \ldots, x_N\}$ denote a possible label assignment for $\mathbf{X}$ (i.e., $X_i = x_i \in \mathcal{L}$), and $\mathbf{y}=\{y_1, \ldots, y_N\}$ denote the observation data (superpixels). A conditional random field (CRF) is associated with the posterior distribution $\Pr(\mathbf{x}|\mathbf{y})$ (13), which can be expressed in the form of a *Gibbs* distribution: $\Pr(\mathbf{x}|\mathbf{y}) = \frac{1}{Z}\exp(-\sum_{c\in\mathcal{C}} \psi_c(\mathbf{x}_c))$, where Z is the partition function, $\mathcal{C}$ denotes the clique set, and $\psi_c(\mathbf{x}_c)$ stands for the potential function associated with the clique c, s.t. $\mathbf{x}_c = \{x_i | i \in c\}$. Consequently, we have the following Gibbs energy function:

$$E(\mathbf{x}) = -\log\ \Pr(\mathbf{x}|\mathbf{y}) = \sum_{c\in\mathcal{C}} \psi_c(\mathbf{x}_c) + \log Z. \tag{1}$$

The optimal solution $\mathbf{x}^*$ can be computed by: $\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathcal{L}^N} \Pr(\mathbf{x}|\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathcal{L}^N} E(\mathbf{x})$.

3. The segmentation framework

3.1 Superpixel classification

In our framework, we first generate multiple superpixels (also named patches) at different resolutions by uniform image subdivisions. Each patch corresponds to a particular scale denoted as $s\times s$ where $s \in \{5, 10, 15, 20, 25\}$ in practice. Let $\mathbf{y}^s = \{y_i^s\}$ be the observation set at scale s, and $\mathbf{x}^s = \{x_i^s\}$ denote the corresponding label set (each x_i^s belongs to $\mathcal{L}$). Adaboost classifiers are trained in order to evaluate the likelihood of a superpixel belonging to a given object class. For convenience, let $H_s(k|y_i^s)$ denote the output score of the learned k-th class Adaboost strong classifier making an additive combination of several decision stump-based weak learners. As a result, we have a tree-structured classification model which allows passing messages between parent and child nodes. The topology of the tree-structured classification model depends on the locations of superpixels. Namely, a child superpixel is connected to a parent one which contain it (i.e., maximal pixel overlap). For instance, given an observation y^5 and its label x^5 at the finest resolution, its corresponding parent nodes are denoted as y^{10}, y^{15}, y^{20}, and y^{25}. Namely, y^5 inherits y^{10}, y^{15}, y^{20}, and y^{25}. In this way, the posterior distribution over x^5 is defined as:

$$h(x^5) = \Pr(x^5|\mathcal{Y}) = \sum_s \left(\frac{\omega_s \exp\left(H_s(x^5|y^s)\right)}{\sum_k \exp\left(H_s(k|y^s)\right)} \right), \tag{2}$$

where $\mathcal{Y} = \{y^5, y^{10}, y^{15}, y^{20}, y^{25}\}$ and $\sum_s \omega_s = 1$ (s.t. $\sum_s \omega_s$=1 and $\omega_s \geq 0$). Consequently, the initial labels of superpixels may be determined by maximizing $h(x^5)$. Fig. 1 gives an intuitive illustration of superpixel classification. In what follows, $\mathbf{x}^5 = \{x_i^5\}$ and $\mathbf{y}^5 = \{y_i^5\}$ are referred to as $\mathbf{x} = \{x_i\}$ and $\mathbf{y} = \{y_i\}$ respectively.

3.2 Higher-order CRF model

A conditional random field (CRF) model with higher-order potential functions is applied to model the spatial contextual dependencies of superpixels. It consists of three terms: 1) a unary potential term; 2) a pairwise potential function; and 3) a higher-order criterion. The unary potential function (denoted as ψ_i) is defined as a linear combination of classification scores referred to in Eq. (2):

$$\psi_i(k) = \mathbf{u}_k^T \mathbf{h}_i, \tag{3}$$

where $\mathbf{h}_i = [h(1) \dots h(K)\ 1]^T$, $\mathbf{u}_k$ denotes a vector of parameters, and $\mathbf{u}_k^T$ is the transpose of $\mathbf{u}_k$. The pairwise potential function $\psi_{ij}(x_i, x_j)$ is composed of four contrast sensitive Potts models in the corresponding 4-neighboring directions:

$$\psi_{ij}(x_i, x_j) = \begin{cases} \mathbf{w}_n^T \mathbf{p}_{ij} & \text{if } x_i \neq x_j \text{ and } \xi_{ij} = n \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $\mathbf{p}_{ij} = [\mathcal{S}(y_i, y_j)\ 1]^T$ with $\mathcal{S}(\cdot,\cdot)$ being a function for evaluating the similarity between two feature observations (see Eq. (10)), ξ_{ij} is an indicator variable determining the neighboring relationship of superpixels i and j, $n \in \{0, 1, 2, 3\}$ corresponding to the {up, right, down, left} neighboring directions, and $\mathbf{w}_n$ is the model parameter vector in the neighboring direction n. Let g denote a group of superpixels. We define our higher-order potential function $\psi_g(\mathbf{x}_g)$ as:

$$\psi_g(\mathbf{x}_g) = \begin{cases} \gamma & \text{if } \frac{\mathbf{F}(x_i)}{|g|} > 0.9 \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $i \in g$, γ is a pre-defined constant, $\mathbf{F}(x_i)$ denotes the number of superpixels whose labels are x_i in g, and $|g|$ represents cardinality of g. Note that g belongs to a partition of superpixel groups generated by applying the unsupervised segmentation algorithm (14) to the superpixels of the input image (each superpixel is represented by its mean RGB color). Now, we define our CRF energy function $E(\mathbf{x})$ as:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \left(\mu(x_i, \mathbf{x}_{\mathcal{N}_i} | \mathbf{y}) + \log Z_i \right), \tag{6}$$

where $\mathcal{N}_i$ denotes the 4-neighboring clique of i, $\mu(x_i, \mathbf{x}_{\mathcal{N}_i} | \mathbf{y})$ integrates the aforementioned three terms of our CRF model:

$$\mu(x_i, \mathbf{x}_{\mathcal{N}_i} | \mathbf{y}) = \psi_i(x_i) + \sum_{j \in \mathcal{N}_i} \psi_{ij}(x_i, x_j) + \psi_g(\mathbf{x}_g), \tag{7}$$

and Z_i is a normalizing constant formulated as:

$$Z_i = \sum_{x_i \in \mathcal{L}} \sum_{j \in \mathcal{N}_i} \sum_{x_j \in \mathcal{L}} \exp\left(-\mu(x_i, \mathbf{x}_{\mathcal{N}_i} | \mathbf{y}) \right).$$

In our CRF model, the model parameters $\mathbf{u}_k$ and $\mathbf{w}_n$ need to be learned using training data. A simple gradient descent method can be used to minimize $E(\mathbf{x})$:

$$\frac{\partial E(\mathbf{x})}{\partial \mathbf{u}_k} = \sum_{i \in \mathcal{V}} \sum_{x_i = k} \left(1 - \frac{\sum_{j \in \mathcal{N}_i} \sum_{x_j \in \mathcal{L}} \exp(-\mu)}{Z_i} \right) \mathbf{h}_i, \tag{8}$$

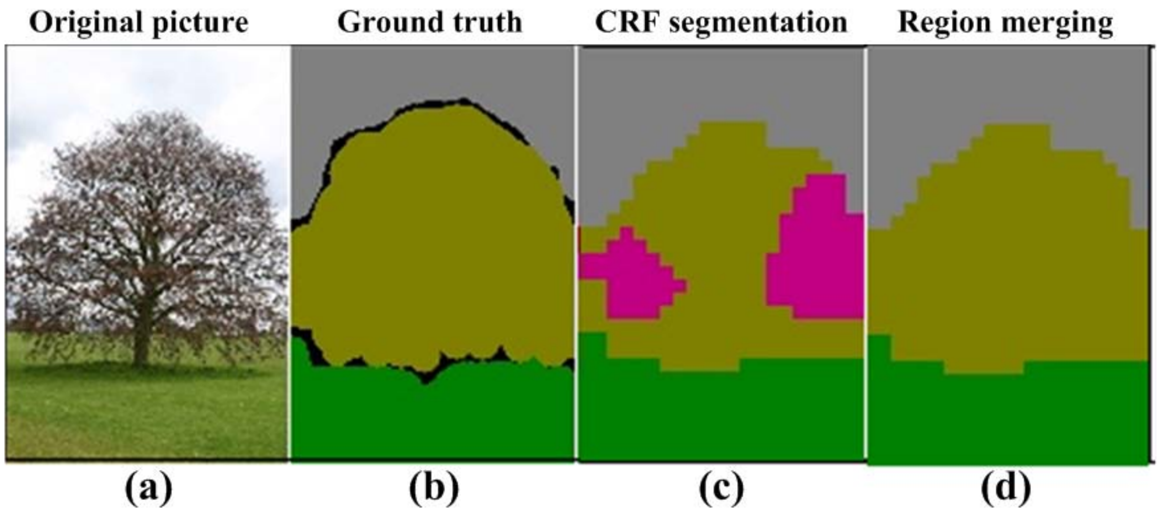

Figure 2: **Illustration of region merging for segmentation refining.**

$$\frac{\partial E(\mathbf{x})}{\partial \mathbf{w}_n} = \sum_{i\in\mathcal{V}} \sum_{\substack{j\in\mathcal{N}_i \\ \xi_{ij}=n}} \left(1 - \frac{\sum\limits_{x_i\in\mathcal{L}} \sum\limits_{x_j\in\mathcal{L}} \exp(-\mu)}{Z_i} \right) \mathbf{p}_{ij}, \tag{9}$$

where μ is the shorthand form of $\mu(x_i, \mathbf{x}_{\mathcal{N}_i}|\mathbf{y})$.

As for maximum a posterior (MAP) inference in our CRF model, an algorithm called Iterated Conditional Modes (ICM) (15) is used to iteratively maximize the local conditional posterior given an initial label configuration (obtained by superpixel classification in Sec. 3.1). If the segmentation performance is still dissatisfactory, we can resort to region merging for segmentation refining. For example, considering a region $\mathbb{R}$ containing M superpixels (i.e., $\mathbb{R} = \{y_i\}_{i=1}^{M}$), there is another region $\mathbb{R}^{'}$ containing Q superpixels (i.e., $\mathbb{R}^{'} = \{y_j^{'}\}_{j=1}^{Q}$). $\mathbb{R}$ is merged into $\mathbb{R}^{'}$ if the following three conditions are satisfied: 1) $\mathbb{R}$ is adjacent to $\mathbb{R}^{'}$; 2) the cardinality of $\mathbb{R}$ is smaller than that of $\mathbb{R}^{'}$; and 3) $\frac{1}{M}\sum_{i=1}^{M}\left(\max\limits_{1\leq j\leq Q} \mathcal{S}(y_i, y_j^{'})\right) > \tau$ where τ is a threshold, and $\mathcal{S}(\cdot,\cdot)$ is a similarity function of any two superpixels (referred to in Eq. (10)). Fig. 2 gives an intuitive illustration of region merging in order to refine segmentation.

4. Experiments

Data set and representation. we evaluate the performance of our method using the Microsoft Research Cambridge (MSRC) database[1], including 240 images of 9 object classes. We consider uniform image subdivisions at multiple scales, resulting in a regular grid of $s \times s$ superpixels per scale ($s \in \{5, 10, 15, 20, 25\}$). Given a scale s, for each superpixel, we extract basic image features, including color, texture, and shape. More specifically, the color feature $f_s^{\mathtt{C}} \in \mathcal{R}^{39}$ is the concatenation of a mean RGB vector and a 36-bin histogram which measures

1. http://research.microsoft.com/en-us/projects/objectclassrecognition/

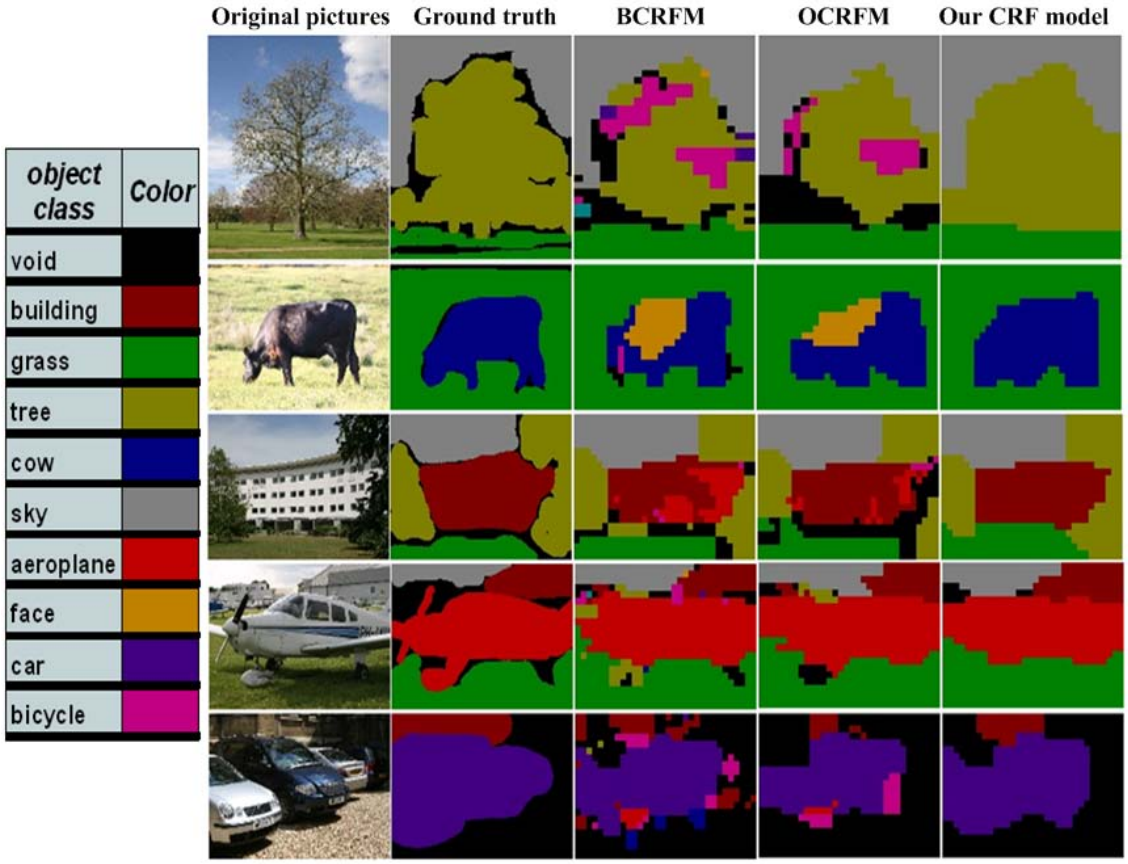

Figure 3: **Segmentation performances of three different methods.**

the distribution of the hue component of HSV colors in the superpixel. The texture feature $f_s^{\mathtt{T}} \in \mathcal{R}^{40}$ is the mean of the Gabor filter response at 8 orientations and 5 scales, estimated on all the locations of the superpixel. The shape feature $f_s^{\mathtt{S}} \in \mathcal{R}^{24}$ is the mean of the 2D log-polar shape contextual histograms (at 8 orientations and 3 scales), taken from different locations in the superpixel.

We denote the global superpixel feature as $f_s = \big(f_s^{\mathtt{C}}\ f_s^{\mathtt{T}}\ f_s^{\mathtt{S}}\big) \in \mathcal{R}^{103}$ and we define the similarity function $\mathcal{S}(f_s, f_s^{'})$ as:

$$\exp\big(- \alpha \| f_s^{\mathtt{C}} - f_s^{\mathtt{C}'} \|^2 - \beta \mathrm{HI}^2(f_s^{\mathtt{S}}, f_s^{\mathtt{S}'}) - \lambda \| f_s^{\mathtt{T}} - f_s^{\mathtt{T}'} \|^2 \big), \tag{10}$$

here $\mathrm{HI}(\cdot,\cdot)$ denotes histogram intersection and (α, β, λ) are respectively set to $(0.01, 0.5, 2.0)$. In Eq. (2), ω_s is set to 0.2. The threshold γ (referred to in Eq. (5)) is set to a very small value ($\gamma = -10^5$ in practice) while the merging criterion threshold τ is set to 0.6.

Comparison. we compare our CRF model with two baseline ones: 1) an orientation-driven CRF model (OCRFM); and 2) a basic CRF model (BCRFM). OCRFM's energy function contains the unary and the pairwise potential functions (i.e., Eqs. (3)-(4)). BCRFM, commonly used for object segmentation, is characterized by its energy function also defined on unary and pairwise superpixel cliques. Its unary potential function has the same form as OCRFM. But its pairwise potential function does not consider directional information between neighboring superpixels while OCRFM does.

Performance on MSRC. MSRC dataset is randomly split into 70% for training and 30% for testing. Fig. 3 shows segmentation (and annotation) results for different methods. Intuitively, Fig. 3 shows that our method performs a more accurate segmentation. Fig. 4 reports false negative (FN) and false positive (FP) rates for different methods, and for each class. Quantitatively, the average FN rates of BCRFM, OCRFM, and our CRF model are 0.1187, 0.1149, and 0.0797, respectively, while the average FP rates of BCRFM, OCRFM, and our CRF model are 0.2995, 0.2485, and 0.2507, respectively. As is known to us, balancing FN and FP is an important criterion for segmentation performance evaluations. Thus, we introduce a quantitative evaluation criterion: $\mathscr{F} = \mathrm{FN}/(1-\mathrm{FP})$. The smaller the value of $\mathscr{F}$, the better the segmentation performance. According to this criterion, the average $\mathscr{F}$s of BCRFM, OCRFM, and our CRF model are computed as 0.1752, 0.1521, and 0.1079, respectively. Consequently, our CRF model achieves the best segmentation performance.

5. Conclusion

In this paper, we introduced a framework for object segmentation and annotation based on a high-order CRF model in order to capture spatial contextual dependencies of superpixels. First, unary potential functions are trained, at multiple scales, using Adaboost; these functions measure the likelihood of superpixels given object classes. Then, an anisotropic pairwise potential function is introduced and makes it possible to capture superpixel dependencies at 4 different neighboring directions. Finally, region label consistency is enforced using a higher-order potential function based on superpixel grouping. Experiments conducted on the standard MSRC database, clearly show the good performance and the substantial gain of our framework.

References

[1] J. Shotton, J. Winn, C. Rother, and A. Criminisi, “TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation”, in *Proc. ECCV*, pp. 1-15, 2006.

[2] D. Batra, R. Sukthankar, and C. Tsuhan, “Learning class-specific affinities for image labelling”, in *Proc. CVPR*, 2008.

[3] C. Galleguillos, A. Rabinovich, and S. Belongie, “Object categorization using co-occurrence, location and appearance”, in *Proc. CVPR*, 2008.

[4] L. Yang, P. Meer, and D. J. Foran, “Multiple class segmentation using a unified framework over mean-shift patches”, in *Proc. CVPR*, 2007.

[5] L Ladický, C. Russell, and P. Kohli, “Associative Hierarchical CRFs for Object Class Image Segmentation”, in *Proc. ICCV*, 2009.

[6] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, “Objects in context”, in *Proc. ICCV*, 2007.

[7] C. Pantofaru, C. Schmid, and M. Hebert, “Object recognition by integrating multiple image segmentations”, in *Proc. ECCV*, 2008.

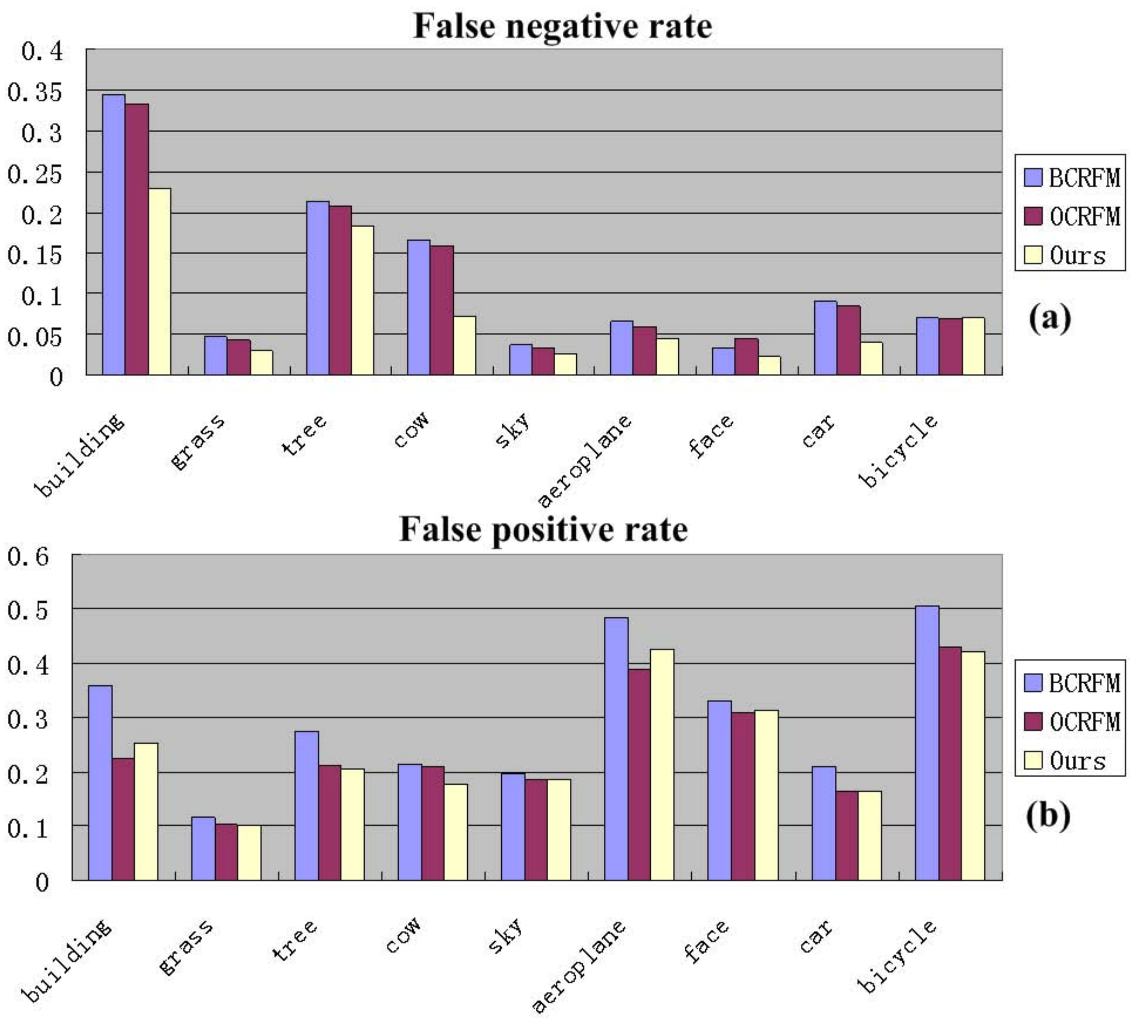

Figure 4: **The quantitative comparison results of BCRFM, OCRFM, and our CRF model.**

[8] J. Reynolds and K. Murphy, "Figure-ground segmentation using a hierarchical conditional random field", in *Proc. Fourth Canadian Conference on Computer and Robot Vision (CRV)*, 2007.

[9] P. Kohli, L. Ladický, P. H. S. Torr, "Robust Higher Order Potentials for Enforcing Label Consistency", in *Proc. CVPR*, 2008.

[10] J. Shotton, M. Johnson, R. Cipolla, "Semantic Texton Forests for Image Categorization and Segmentation", in *Proc. CVPR*, 2008.

[11] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis", *IEEE Trans. PAMI*, 2002.

[12] J. Shi and J. Malik, "Normalized cuts and image segmentation", *IEEE Trans. PAMI*, 2000.

[13] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labelling sequence data", in *Proc. ICML*, 2001.

[14] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation", *IJCV*, 2004.

[15] J. Besag. "On the statistical analysis of dirty pictures", *Journal of Royal Statistical Soc.*, B-48:259-302, 1986.

Dépôt légal : 2010 – 3ème trimestre
Imprimé à Télécom ParisTech – Paris
ISSN 0751-1345 ENST D (Paris) (France 1983-9999)

Télécom ParisTech

Institut TELECOM - membre de ParisTech

46, rue Barrault - 75634 Paris Cedex 13 - Tél. + 33 (0)1 45 81 77 77 - www.telecom-paristech.frfr

Département TSI