Weakly Supervised Semantic Segmentation for Social Images

Wei Zhang, Sheng Zeng, Dequan Wang, and Xiangyang Xue
Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai, China

Image semantic segmentation is the task of partitioning image into several regions based on semantic concepts. In this paper, we learn a weakly supervised semantic segmentation model from social images whose labels are not region-level but image-level; furthermore, these labels might be noisy. We present a joint conditional random field model leveraging various contexts to address this issue.

Suppose that each image I is associated with a label vector $y=[y_1,...,y_L]$, where L is the number of categories, and $y_i=1$ indicates that the i-th category is present in this image, otherwise $y_i=0$. In the training set, y is given; however, it might be incorrect. In the test set, y is unknown. For each image, we firstly employ the existing multi-scale segmentation algorithm [1] to get a set of superpixels $\{x_p\}_{p=1}^{M}$ over multiple quantization levels. Here, M is the total number of superpixels in image I. The label of superpixel x_p is denoted as $h_p \in \{1,2,...,L\}$, and the labels of all superpixels for image I are $h=[h_1,...,h_M]$, which are not available for training.

Our goal is to infer semantic label for each superpixel in an image and the adjacent superpixels sharing the same semantic label are fused as a whole one. We jointly build a conditional random field (CRF) over the image-level label variables y and the superpixel-level label variables h. We leverage label-pair correlation and connect each superpixel to its neighbors to encode local smoothness constraints. Thus we formulate an energy function E with five types of potentials as follows:

$$E(y,h,I)=\sum_{i=1}^{L}\varphi_i(y_i,I)+\sum_{1\le i,j\le L}\varphi_{ij}(y_i,y_j) + \sum_{p=1}^{M}\psi_p(h_p,x_p)+\sum_{(p,q)\in\mathcal{N}}\psi_{pq}(h_p,h_q)+\tau(y,h) \tag{1}$$

where φ_i and ψ_p are the unary potentials for feature-label associations on image-level and superpixel-level respectively, φ_{ij} is the pairwise potential for label correlation, ψ_{pq} is the pairwise potential encoding the spatial context constraints for adjacent superpixels, $\mathcal{N}$ denotes the set of pairs of neighboring superpixels, and τ ensures the coherence between image-level labels and superpixel-level labels.

We formulate the image-level potential for feature-label association φ_i as follows:

$$\varphi_i(y_i,I)=-\ln\frac{\exp\{f_i(y_i,I)\}}{\exp\{f_i(0,I)\}+\exp\{f_i(1,I)\}} \tag{2}$$

where $f_i(y_i,I)$ is the linear support vector machine score for the semantic concept i with features extracted by convolutional neural network (CNN) and latent semantic concept model (LSC). Although the labels of social images for training might be noisy, the potential for feature-label association is robust due to the features learned by the latent semantic concept model which is unsupervised.

Similar with image-level potential, the superpixel-level potential for feature-label association is formulated as follows:

$$\psi_p(l,x_p)=-\ln\frac{\exp\{a_p^\top\theta_a^l+c_p^\top\theta_c^l\}}{\sum_{i=1}^{L}\exp\{a_p^\top\theta_a^i+c_p^\top\theta_c^i\}} \tag{3}$$

where $x_p=[a_p;c_p]$ is the feature vector concatenating the CNN feature and latent semantic concept distribution extracted from the superpixels.

To model the pairwise potential of inter-label correlations, we not only utilize label co-occurrence statistics but also capture visual contextual cues, as shown in Figure 1. The label correlation potential φ_{ij} can be defined as follows:

$$\varphi_{ij}(y_i,y_j)=A(i,j)R(i,j)\mathbf{1}(y_i\neq y_j) \tag{4}$$

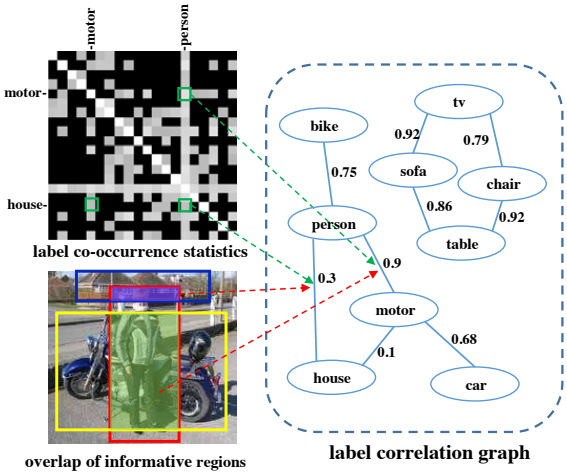

Figure 1: Illustration of the pairwise potential for label correlations which are leveraged from two aspects: label co-occurrence statistics and visual contextual cues.

where $A(i,j)$ and $R(i,j)$ capture label correlations by label co-occurrence statistics and visual contextual cues, respectively, and $\mathbf{1}(\cdot)$ is the indicator function.

Inspired by [2, 3], we also focus on adjacent superpixels in the same quantization level and overlapped superpixels in the neighboring levels, and define the pairwise potentials for superpixels as follows:

$$\psi_{pq}(h_p,h_q)=\begin{cases}\phi_{inter}(h_p,h_q) & \text{if } |lev(p)-lev(q)|=1,\\ \phi_{intra}(h_p,h_q) & \text{if } lev(p)=lev(q),\\ 0 & \text{otherwise}\end{cases} \tag{5}$$

where $lev(p)$ indicates the quantization level for superpixel x_p. ϕ_{inter} encodes that superpixels lying within the same clique are more likely to take the same label. ϕ_{inter} can be used to find the proper segmentation scale for each object.

It is naturally required that superpixel-level labels should be consistent with image-level labels: if any superpixel x_p takes the label i, then the image label indicator $y_i=1$; otherwise $y_i=0$. Such constraints can be encoded by the following potential:

$$\tau(y,h)=C\sum_{i,p}\mathbf{1}(y_i=0\wedge h_p=i) \tag{6}$$

where C is a cost that penalizes any inconsistency between the image-level and superpixel-level labels. Such label consistency potential is a soft constraint, and we can further refine superpixel label and image label via an iterative process.

Experimental results on two real-world image datasets PASCAL VOC2007 and SIFT-Flow demonstrate that the proposed approach outperforms state-of-the-art weakly supervised methods and even achieves accuracy comparable with fully supervised methods.

[1] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, 2014.

[2] Pushmeet Kohli, Philip HS Torr, et al. Robust higher order potentials for enforcing label consistency. *IJCV*, 2009.

[3] Lubor Ladicky, Christopher Russell, Pushmeet Kohli, and Philip HS Torr. Associative hierarchical crfs for object class image segmentation. In *CVPR*, 2009.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.