LaTeX Author Guidelines for CVPR Proceedings

Anonymous CVPR submission

Paper ID ****

Abstract

The ABSTRACT is to be in fully-justified italicized text, at the top of the left-hand column, below the author and affiliation information. Use the word "Abstract" as the title, in 12-point Times, boldface type, centered relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type. Leave two blank lines after the Abstract, then begin the main text. Look at previous CVPR abstracts to get a feel for style and length.

1. Introduction

2. Related Work

3. The Proposed Model

Assume we have a set of weakly labeled images and each image is oversegmented into several superpixels. We fomulate this weakly supervised semantic segmentation as a joint learning problem which we factor into multi-class and binary CRFs.

...

Suppose we have a set of weakly labeled images $\mathcal{I} = \{I^k\}_{k=1}^M$ and each image is oversegmented into m_k superpixels $\mathrm{X}^k = \{x_i^k\}_{p=1}^{m_k}$ by [1]. We describe each superpixel x_i^k by its appearance model and topic model (see Sec. 4 for details). Let $\boldsymbol{y}^k = (y_1^k, ..., y_L^k)^{\mathrm{T}}$ denote a vetor of the L binary label variables, *i.e.* $y_i^k \in \{0, 1\}$, where $y_i^k = 1$ indicates that category i is present in image k, and 0 otherwise. For each superpixel x_p^k, we define a random variable $h_p^k \in \{1, ..., L\}$ to represent its semantic category.

Our goal is to find an optimal label configuration that ... To tackle this problem, we build a conditional random field (CRF) on the image-level label variables $\boldsymbol{y}$ and the superpixel variables $\boldsymbol{h}$. We connect each superpixel variables to its neighbors to encode a local smoothness constraint, . Specifically, let $\mathcal{N}$ donate the superpixel neighborhood, we define an energy function E with five types of potential as follow:

$$\begin{aligned} E(\boldsymbol{y},\boldsymbol{h},I) = &\sum_{i=1}^{L}\psi_G(y_i,I) + \sum_{1\le i,j\le L}\psi_R(y_i,y_j) \\ &+ \sum_{p=1}^{m}\psi_{at}(h_p,x_p) + \sum_{(p,q)\in\mathcal{N}}\psi_S(h_p,h_q) \\ &+ \psi_C(\boldsymbol{y},\boldsymbol{h}) \end{aligned} \tag{1}$$

where ψ_G and ψ_{at} encode the unary potential of global and regional constraints respectively, ψ_R impose labels' correlation and co-occurrence, ψ_S are the spatial context constraints for each superpixel, and ψ_C ensure the consistency between global and regional labels. The details of each potential will be described in the following sections. The posterior distribution $P(\boldsymbol{y},\boldsymbol{h}|I)$ of the CRF can be written as $P(\boldsymbol{y},\boldsymbol{h}|I) = \frac{1}{Z(I)}\exp\{-E(\boldsymbol{y},\boldsymbol{h},I)\}$, where $Z(I)$ is the normalizing constant. Thus, the most probable labelling configuration $\boldsymbol{y}^\star, \boldsymbol{h}^\star$ of the random field can be defined as $\boldsymbol{y}^\star, \boldsymbol{h}^\star = \arg\min_{\boldsymbol{y},\boldsymbol{h}} E(\boldsymbol{y},\boldsymbol{h},I)$.

3.1. Label Consistency

We require that the superpixel labels be consistent with the image labels: if any superpixel x_i takes the label l, then image label indicator $y_l = 1$; otherwise $y_l = 0$. Such constraints can be encode by the following potential:

$$\psi_C(\boldsymbol{y},\boldsymbol{h}) = C \cdot \sum_{l,i} I(y_l = 0 \text{ and } h_i = l) \tag{2}$$

where $I(\cdot)$ is the indicator function and C is a large constant that penalizes any inconsistency between the global and local labels.

3.2. Appearance Model and Topic Model

We include both appearance and topic model as follow:

$$\begin{aligned} \psi_{at}(h_p,x_p) = &-\log\{w_1\phi_a(h_p,a_p,\theta_a) \\ &+ w_2\phi_t(h_p,t_p,\theta_t)\} \end{aligned} \tag{3}$$

where a_p, t_p are the appearance and topic feature vectors extracted from the superpixels, θ_a, θ_t donate the parameters with repect to appearance model and topic model, $\{w_i\}|_{i=1}^2$ are the weighting coefficients for the unary terms. We define the appearance model $\phi_a(h_p, a_p, \theta_a) = f_{h_p}(a_p, \theta_a)$ and topic model $\phi_t(h_p, t_p, \theta_t) = g_{h_p}(t_p, \theta_t)$ measuring how well the local appearance a_p and topic t_p matches the semantic label h_p.

3.3. Spatial Context Constraints

We

$$\psi_S(h_p, h_q) = \begin{cases} & \text{if } l_p = l_q - 1, \\ Sim(x_p, x_q) \cdot R(h_p, h_q) & \text{if } l_p = l_q, \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $Sim(x_p, x_q) \in [0, 1]$ measures the visual similarity between superpixel x_p and x_q, $R(h_p, h_q) \in [0, 1]$ is a learnt correlation between label h_p and h_q. Hence, we pay a high cost for the similar superpixels if they were assigned different labels and for the superpixels which were assigned an irrelevant label to the context.

3.4. Label Correlation and Co-occurrence

3.5. Joint Inference with Alternate Procedure

The energy minimization problem (1) can be solved in the following two alternate optimization steps:

$$\begin{aligned} \boldsymbol{y}^* = \arg\min_{\boldsymbol{y}} & \sum_i \psi_G(y_i, I) + \frac{1}{2}\psi_C(\boldsymbol{y}, \boldsymbol{h}^*) \\ & + \sum_{1 \le i,j \le L} \psi_R(y_i, y_j), \end{aligned} \tag{5}$$

$$\begin{aligned} \boldsymbol{h}^* = \arg\min_{\boldsymbol{h}} & \sum_p \psi_{at}(h_p, x_p) + \frac{1}{2}\psi_C(\boldsymbol{y}^*, \boldsymbol{h}) \\ & + \sum_{(p,q) \in \mathcal{N}} \psi_S(h_p, h_q). \end{aligned} \tag{6}$$

As a standard binary CRF problem, the first subproblem in Eq. (5) has an explicit solution which utilizes min-cut/max-flow algorithms (*e.g.* the Dinic algorithm [4]) to obtain the global optimal label configuration. And the second subproblem in Eq. (6) reduces to an energy minimization for a multiclass CRF. Although finding the global optimum for this energy function has been proved to be a NP-hard problem, there are various approximate methods for fast inference, such as approximate maximum a posteriori (MAP) methods (*e.g.* graph-cuts [2]). In this paper, we adopt move-making approach that finds the optimal α-expansion [2, 6] by converting the problems into binary labeling problems which can be solved efficiently using graph cuts techniques. The energy obtain by α-expansion has been proved to be within a known factor of the global optimum [2]. Considering the two alternate optimization steps together, we summarize our XXXX in Algorithm 1.

Algorithm 1 Energy minimization

```
123
```

4. Appearance and Topic Model Generation

We use Convolutional Neural Network (CNN) to encode the superpixels' appearance. CNN has made a significant breakthrough in object detection and semantic segmentation tasks [5]. As demonstrated in [5], the classification network trained on ImageNet [3] can generalize well to the detection task. We train a classification model on ILSVRC with the same setup to [5], which uses five convolutional layers and three fully-connected layers. We represent each superpixel by the *fc6* layer, which is the first fully-connected layer containing 4096 neurons. Therefore, the appearance representation of each superpixel is a feature vector with 4096 dimentions.

Moreover, we learn the latent category (known as topic model) from the superpixels.

5. Experiments

6. Conclusion

References

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

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(11):1222–1239, 2001. 2

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 2

[4] E. Dinits. Algorithm of solution to problem of maximum flow in network with power estimates. *Doklady Akademii Nauk SSSR*, 194(4):754, 1970. 2

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint arXiv:1311.2524*, 2013. 2

[6] V. Kolmogorov and R. Zabin. What energy functions can be minimized via graph cuts? *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(2):147–159, 2004. 2