Data-Driven Scene Understanding with Adaptively Retrieved Exemplars

Xionghao Liu, Wei Yang, Liang Lin, Qing Wang, Zhaoquan Cai, and Jianhuang Lai, Member, IEEE

Abstract—This article investigates a data-driven approach for semantically scene understanding, without pixelwise annotation and classifier training. Our framework parses a target image with two steps: (i) retrieving its exemplars (i.e. references) from an image database, where all images are unsegmented but annotated with tags; (ii) recovering its pixel labels by propagating semantics from the references. We present a novel framework making the two steps mutually conditional and bootstrapped under the probabilistic Expectation-Maximization (EM) formulation. In the first step, the references are selected by jointly matching their appearances with the target as well as the semantics (i.e. the assigned labels of the target and the references). We process the second step via a combinatorial graphical representation, in which the vertices are superpixels extracted from the target and its selected references. Then we derive the potentials of assigning labels to one vertex of the target, which depend upon the graph edges that connect the vertex to its spatial neighbors of the target and to its similar vertices of the references. Besides, the proposed framework can be naturally applied to perform image annotation on new test images. In the experiments, we validate our approach on two public databases, and demonstrate superior performances over the state-of-the-art methods in both semantic segmentation and image annotation tasks.

Index Terms—scene understanding, semantic segmentation, image retrieval, graphical model, image annotation

I. INTRODUCTION

Significant progresses have been identified in solving the task of semantic image understanding [14], [5]. However, these methods usually build upon supervised learning with fully annotated data that are expensive and sometimes limited in large-scale scenarios [9], [7]. Several weakly supervised methods were proposed [17] to reduce the overload of data annotating, which can be trained with only image-level labels indicating the classes presented in the images. Recently, data-driven approaches [10], [11] receive increasing attentions, which tend to leverage knowledges from auxiliary data in weakly supervised fashions, and demonstrate very promising applications. Following this trend, one interesting but challenging problem arises for the scene understanding: How to parse the raw images in virtue of the strength of numerous unsegmented but tagged images, as the image-level tags can be achieved easier. In this work, we investigate this problem by developing a unified framework, in which the two following steps perform iteratively, as Fig. 1 illustrates.

In Step. 1, we search for similar images as the exemplars (i.e. references) matching to the target image from the auxiliary database (in Fig. 1 (b)), and these references are required to share similar semantic concepts with the target. Moreover, we enforce the representation to be semantically meaningful: The references that are selected should contain consistent tags. The tags of the target image can be also taken into account during the iteration, as they can be determined by the last label assignment step (in Step. 2). We solve this step using the proximal gradient method.

In Step. 2, we assign labels to the pixels of the target by propagating semantics from the selected references. We create a graphical model, in which the vertices are the superpixels from the target image and its references. There are two types of edges defined over the graph, which is inspired by [6]: (i) the inner-edges connecting the spatial adjacent vertices...
within the target; (ii) the outer-edges connecting the vertices of the target to those of its references. The potentials are then derived into an MRF form by aggregating the two types of edge connections, which can be fast solved by the Graph Cuts algorithm [5].

The two above steps are mutually conditional, providing complementary information to each other. We present a novel probabilistic Expectation-Maximia (EM) formulation making the two steps bootstrapped by each other to conduct results in a self-driven manner. In addition, the proposed framework can also be directly applied on new test image to perform multi-label image annotation. Our approach is evaluated on several benchmarks, and outperforms other state-of-the-art methods.

II. RELATED WORK

Traditional efforts for scene understanding mainly focused on capturing scene appearances, structures and spatial contexts by developing combinatorial models, e.g., CRF [14], 5, Texton-Forest [13], Graph Grammar [8]. These models were generally founded on supervised learning techniques, and required manually prepared training data containing labels at pixel level.

Several weakly supervised methods are proposed to indicate the classes presented in the images with only image-level labels. For example, Winn et al. [16] proposed to learn object classes based on unsupervised image segmentation. Zhang et al. [17] learned classification models for all scene labels by selecting representative training samples, and multiple instance learning was utilized in [15].

Some nonparametric approaches have been also studied that solve the problems by searching and matching with an auxiliary image database. For example, an efficient structure-aware matching algorithm was discussed in [10] to transfer labels from the database to the target image, but the pixelwise annotation was required for the auxiliary images.

III. PROBLEM FORMULATION

In this section, we phrase the problem in a probabilistic formulation, and then discuss the Expectation-Maximization (EM) inference framework for optimization.

A. Probability Model

Let \( \Delta = \{I_k, L_k\}_{k=1}^{N} \) denote a set of images \( \{I_k\} \) with image-level labels \( \{L_k\} \). Each image \( I_k \) is represented as a set of superpixels \( \{x^k_i\}_{i=1}^{n_k} \), where \( n_k \) is the number of superpixels in \( I_k \).

Given the target image \( I_t \), our task is to predict its image-level labels \( L_t \), as well as to assign each superpixel \( x^t_i \) a label \( y^t_i \in L_t \). Let \( Y_t \) denote the whole label assignment, i.e., \( Y_t = \{y^t_i\}_{i=1}^{n_t} \), we can define the joint probability distribution of target image \( I_t \) and the label assignment \( Y_t \).

We also define a binary-valued correspondence variable \( \alpha = \{\alpha_{k}\}_{k=1}^{N} \) such that \( \alpha_k = 1 \) if image \( I_k \) is selected as a reference for the target image. \( \alpha \) is treated as a hidden variable.

The complete probability model is defined as follows,

\[
P(I_t, Y_t, \alpha | \Delta) = P(I_t, Y_t | \alpha, \Delta)P(\alpha),
\]

and we further derive it by summing out \( \alpha \) as,

\[
P(I_t, Y_t | \Delta) = \sum_{\alpha} P(I_t, Y_t | \alpha, \Delta)P(\alpha).
\]

Then the optimal label assignment \( Y_t^* \) by maximizing the probability,

\[
Y_t^* = \arg\max_{Y_t} P(I_t, Y_t | \Delta),
\]

and we propose to solve it iteratively under an Expectation-Maximization (EM) framework.

B. The EM Iterations

It has been shown that estimating \( Y_t^* \) from \( P(I_t, Y_t | \Delta) \) is equivalent to minimize the following energy function [12]:

\[
\mathcal{L}(Q, Y_t) = -\sum_{\alpha} Q(\alpha) \ln P(I_t, Y_t, \alpha | \Delta) + \sum_{\alpha} Q(\alpha) \ln Q(\alpha),\]

where \( Q(\alpha) \) is the posterior of the latent variable \( \alpha \).

Since the second term in Eq. (4) is a constant, the optimization iterates with two steps: (i) The \( E \)-step minimizes the energy \( \mathcal{L}(Q, Y_t) \) with respect to \( Q(\alpha) \) with \( Y_t \) fixed. (ii) The \( M \)-step minimizes the energy \( \mathcal{L}(Q, Y_t) \) with respect to \( Y_t \) with \( Q(\alpha) \) fixed.

(i) The \( E \)-step: Approximating \( Q(\alpha) \):

The posterior of the latent variable \( Q(\alpha) \) is defined as,

\[
Q(\alpha) = P(\alpha | I_t, Y_t, \Delta) = \frac{1}{Z} \exp\{-E_{\alpha}(\alpha, I_t, Y_t, \Delta)\},
\]

where \( Z \) is the normalization constant of the probability. The energy \( E_{\alpha} \) evaluates the appearance and semantics consistency, which is specified as,

\[
E_{\alpha}(\alpha, I_t, Y_t, \Delta) = E_{Sc}(\alpha, I_t, \Delta) + \gamma E_{Sa}(\alpha, Y_t, \Delta),
\]

The first term \( E_{Sc} \) measures the appearance similarity between \( I_t \) and images in \( \Delta \), defined as,

\[
E_{Sc} = \frac{1}{2} \| F(I_t) - B\alpha \|_2^2 + \beta \| \alpha \|_1,
\]
where $\beta$ is the tradeoff parameter used to balance the sparsity and the reconstruction error. $F(\cdot)$ is an $m$-dimensional global feature of an image, and $B \in \mathbb{R}^{m \times N}$ is a matrix consisting of all the features of images in $\Delta$.

The second term $E_{Sa}$ in Eq. (6) measures semantic consistency, defined as,

$$E_{Sa} = \frac{1}{2} \sum_{i,j \in N} S_{ij} \| \frac{\alpha_i}{\sqrt{A_{ii}}} - \frac{\alpha_j}{\sqrt{A_{jj}}} \|^2 + \lambda \alpha^T D \alpha$$

$$= \alpha^T \mathcal{L} \alpha + \lambda \alpha^T D \alpha,$$

(8)

where $S_{ij}$ measures the semantic similarity between $(I_i, I_j) \in \Delta$, as,

$$S_{ij} = \frac{|L_i \cap L_j|}{|L_i \cup L_j|}.$$  

(9)

and $A$ in Eq. (8) is a diagonal matrix where $A_{ii} = \sum_{j=1}^{N} S_{ij}$ and $L = A^{-1/2} (A - S) A^{-1/2}$, in which $L$ is the normalized Laplacian matrix. Images with similar semantics should be encoded with similar activations. In other words, if two images have common features of an image, and their activation codes should be small. We initialize $\alpha$ as the whole label set of the database.

During the later iterations.

The M-step performs to minimize the following energy function with respect to $Y_t$:

$$E_M(Y_t) = -\sum_{\alpha} Q(\alpha) \ln P(I_t, Y_t, \alpha | \Delta).$$

(11)

However, summing out $\alpha$ for all possibilities demands very expensive computational cost, particularly to process a large number $N$ of data. Instead, we seek a lower-bound of $E_M(Y_t)$.

Assume that we can infer $\alpha^*$ with the maximized probability $Q(\alpha^*)$ by the E-step. Then we can define the joint distribution of $(I_t, Y_t)$ conditioned on $Q(\alpha^*)$, and we have

$$\sum_{\alpha} P(I_t, Y_t | \Delta; \alpha^*) > \sum_{\alpha} P(I_t, Y_t, | \Delta)$$

(12)

It is straightforward in the context of our task, as the cumulative density of assigning labels from good references (i.e. given $\alpha^*$) is higher than that with general cases. Thus, we set the lower-bound as,

$$E_M(Y_t) > -\sum_{\alpha} Q(\alpha) \ln P(I_t, Y_t, | \Delta; \alpha^*),$$

(13)

where $Q(\alpha)$ is fixed by the last E-step. The energy to be minimized can be further simplified as,

$$E_M(Y_t) = -\ln P(I_t, Y_t | \Delta, \alpha^*),$$

(14)

where we will specify $-\ln P(I_t, Y_t | \Delta, \alpha^*)$ with a combinatorial graph model in Sec. IV-B.

IV. INFERENCE AND IMPLEMENTATION

Within the EM formulation, the inference algorithm iterates with two steps: (i) computing $\alpha^*$ in the E-step for reference retrieval and (ii) solving the optimal labeling $Y_t^*$ with the selected references in the M-step.

Algorithm 1 Adaptive Reference Retrieval

**Input:** Target image feature $F(I_t)$, codebook $B$, semantic constrains $A$, and the threshold $\sigma$ for stop.

**Output:** Semantical sparse coding coefficient $\alpha^*$.

**Initial:** $\alpha^*$ in randomly , and $k = 1$. Denote $g(\alpha) = \frac{1}{2} \| F(I_t) - B \alpha \|^2 + \frac{1}{2} \gamma \alpha^T \Lambda \alpha$, so Eq. (15) can be reformulated as $E_{\alpha} = g(\alpha) + \beta \| \alpha \|_1$.

1: while $\| \alpha^{k+1} - \alpha^k \|_2 > \sigma$ do
2: Compute the gradient of $g(\alpha)$ at $\alpha^k$, $\nabla g(\alpha^k) = B^T (B \alpha^k - F(I_t)) + \gamma \Lambda \alpha^k$.
3: $z_L^k = \arg \min_{z \in \mathcal{L}} (z - \alpha^k)^T \nabla g(\alpha^k) + \beta \| z \|_1 + \frac{1}{2} \gamma \| z - \alpha^k \|^2$, where $L > 0$ is a parameter.
4: Iteratively increasing $L$ by a constant factor until the condition $g(z_L^k) \leq M_2 \alpha^k, z_L^k := g(\alpha^k) + \nabla g(\alpha^k)^T (z_L^k - \alpha^k) + \frac{L}{2} \| z_L^k - \alpha^k \|^2$ is met, else return step 3.
5: Update $\alpha^{k+1} := \alpha^k + \nu_k (z_L^k - \alpha^k)$, where $\nu_k \in (0, 1]$
6: $k:=k+1$
7: end while
8: $\alpha^* = \alpha^k$

A. Adaptive Reference Retrieval

Maximizing $Q(\alpha)$ is equivalent to minimizing the energy defined in Eq. (6) w.r.t. $\alpha^* = \arg \min_{\alpha} E_{\alpha}(\alpha, I_t, Y_t, \Delta)$. Notice that $E_{\alpha}(\alpha, I_t, Y_t, \Delta)$ can be regarded as a semantic-aware sparse representation, where we jointly model the appearance reconstruction with semantic consistency. Fig. 2 intuitively illustrates this model, and it can be rewritten as,

$$E_{\alpha} = \frac{1}{2} \| F(I_t) - B \alpha \|^2 + \beta \| \alpha \|_1 + \frac{1}{2} \gamma \alpha^T \Lambda \alpha,$$

(15)

where $\Lambda = 2(\mathcal{L} + \lambda D)$. The semantic associated terms in Eq. (15) can be phrased in convex forms, thus we can use the proximal gradient method to solve this problem efficiently. The optimization process is shown in Algorithm 1.

Given the optimized $\alpha^*$, we can simply select the references according to coding co-efficiencies, e.g., select by thresholding. And we set $\nu_k = 0$ if image $I_k$ is not selected.
where the target and its neighboring superpixels connected by outer-edges in the reference image \( I_k \), thus it implicitly exhibits the probability that \( x_i^t \) sharing the same labels with its reference \( I_k \).

**Algorithm 2** Overall procedure of our framework

**Input:** Target \( I_t = \{ x_i^t \}_{i=1}^{n_t} \), and auxiliary \( \Delta = \{ I_k, L_k \}_{k=1}^{N} \).

**Output:** Label of each superpixel \( Y_t = \{ y_i^t \}_{i=1}^{n_t} \)

**Initial:** \( L_t^1 \) contains all labels, and \( n = 1 \).

1. while \( L_t^{n+1} \neq L_t^n \) do
2. Minimize \( E_\alpha \) defined in Eq. (15) using Alg. 1
3. Sort \( \alpha^* \) in descend order, select the images corresponding to the \( p \)-first nonzero coefficients, as a set \( B \).
4. for all \( x_i^t \) in \( I_t \) do
5. for all image \( I_k \) in \( B \) do
6. Select the \( q \)-most similar superpixels \( O^k = \{ x_j^k \}_{j=1}^{q} \).
7. Construct \( O_{x_i^t} = \bigcup_k O_{x_i^t}^k \)
8. end for
9. Add \((x_i^t, x_j^k)\) to \( \omega \) for all \( x_j^k \in O_{x_i^t} \).
10. Add \((x_i^t, x_j^k)\) to \( \xi \) for all neighbors \( \{ x_j^k \} \) of \( x_i^t \), \( i \neq j \).
11. end for
12. Minimize Eq. (16). Optimize the latent label \( Y_t^* \) using alpha-beta swap algorithms of graph cuts.
13. Update \( L_t^{n+1} \) as the unique set of \( Y_t^* \).
14. \( n := n+1 \)
15. end while

The pairwise potentials, i.e. \( \phi(y_i^t, y_j^k, x_i^t, x_j^k) \) in Eq. (16), encourages the smoothness between neighboring superpixels within the target, as,

\[
\phi(y_i^t, y_j^k, x_i^t, x_j^k) = \| f(x_i^t) - f(x_j^k) \|_2 \delta(y_i^t \neq y_j^k),
\]

where \( \delta(\cdot) \) is the indicator function.

Thus the approximate solutions Eq. (16) can be found using alpha-beta swap algorithms of graph cuts. The sketch of our framework is shown in Algorithm 2.

**C. Image Annotation**

We propose a simple method to transfer \( n \) labels to a test image \( I_t \) from the query’s \( K \) nearest neighbors in the training set. For a given test image \( I_t \), the sparse reconstruction coefficient vector \( \alpha \) is determined by solving the problem in Eq. (15), where we set \( \lambda = 0 \), and set other parameters as the same as described in section V-B1. The optimal sparse coefficient solution denote as \( \tilde{\alpha} \), then let its top \( K \) largest value denote as \( \tilde{\pi} \in \mathbb{R}^K \) corresponding with image label indicator \( l_i \in \mathbb{R}^C \), \( i = 1, 2, \ldots, K \). The label vector probability of test image can then be obtained as:

\[
z_t = \sum_{i=1}^{K} \tilde{\pi}_i l_i
\]

where \( \tilde{\pi} \) is the \( i \)-th component of vector \( \tilde{\pi} \). The labels corresponding to the top few largest values in \( z_t \) are considered as the final annotations of the test image.

We compare the following two annotation methods, and find out that the sparse coefficient \( \alpha \) is extremely useful for image
annotation. (i) **weighted**: That is the annotation weighed by sparse reconstruction coefficient $\hat{\pi}_i$. (ii) **unweighted**: We set $\hat{\pi}_i = 1, i = 1, \cdots, K$ in manual.

Besides, we also compared with classical works for image annotation, the proposed method here have the following characteristics: (i) the propagation process is robust and less sensitive to the image noises owing to the semantic constraints in image retrieval step. (ii) the proposed algorithm is scalable to large-scale, and retrieval images by jointly matching their appearances as well as the semantics.

V. EXPERIMENT

In this section, we conduct extensive experiments to validate the performance of our method and discuss the experimental analysis. We also conduct an empirical study on the effectiveness of the proposed EM iterations.

**Implementation details**: Five parameters are required to be set in our framework. We set $q = 20$ to construct the $q$-nearest graph, and set $p = 10$ to retrieval 10 images as reference for each test image. In the experiment we also set $\lambda = 1$ empirically. The other parameters $\beta$ and $\gamma$ are introduced in Sec. (V-B1).

![Image](image_url)

**Fig. 6.** Illustration of the decrease energy $E_\alpha$ decrease w.r.t. time. $x$-axis indicates the number of iteration, and the $y$-axis shows the energy $E_\alpha$ of Eq. (15). The results randomly selected from test set.

A. Datasets

To verify the effectiveness of our method, we conduct experiments on two challenging datasets, i.e. MSRC [14] and VOC 2007 [2], by comparing with state-of-the-art. We use the standard average per-class measure (average accuracy) to evaluate the performance. For each test image, we use the training set as the auxiliary data for our framework.

B. Exp-I: Image Semantic Segmentation

1) **Parameter Analysis**: Specifically, we focus on the effects of $\beta$ and $\gamma$ which control the influence of appearance term and semantic term in Eq. (15), and these two parameters are crucial to our results. The range of $\beta$ and $\gamma$ are both set to $\{0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$. The semantic segmentation performance is used to tune parameters.

We used MSRC dataset to finetune the parameters. The results of changing the parameter values are presented in Fig. 7 from which we can observe the following conclusions:

- When $\beta$ and $\gamma$ increase from small values to large values, the performance varies apparently, which shows that the sparse term and semantic constraint term have great impacts on the performance.

- Mean average precision (MAP) reach the peak points (0.71) when $\beta = 0.1$ and $\gamma = 0.2$ on MSRC which lie in the middle range and the precision do not increase monotonically when $\beta$ and $\gamma$ increase. In the following experiments, we adopt the best parameter settings on all datasets.

![Image](image_url)

**Fig. 7.** Parameter tuning results of parameters $\beta$ and $\gamma$ for MSRC dataset.

2) **Experiments on MSRC dataset**: Given this insight, we compare the proposed method with the following stae-of-the-art algorithms: MIM [15], and K. Zhai [17].

Table I shows that our algorithm outperforms the others. Benefit from the semantic constraints incorporated in our approach, we achieve a significant improvements for certain difficult classes, e.g., chair and cat. Serveral visualized results with the corresponding ground-truths are presented in Fig. 4(a) and more semantic segmentation results are in supplementary material as to the limited space of article.

3) **Experiments on VOC 2007 dataset**: Few performance on VOC 2007 dataset is reported, due to the 20 extremely challenging categories it contains. Here we compare with the weekly supervised STF[13] by running the code provide by the author. We also compare our method with [17]. Results are reported in Table I and our methods outperforms [17] by 3%.

It takes about 8 seconds per image with an un-optimized matlab implementation for semantic segmentation, on a 64-bit system with Core-4 3.6 GHz CPU, 4GB memory (extracting features: 1s; sparse coding with semantic constraints: 5s; optimization by GraphCuts: 2s).

Moreover, we validate the effectiveness of the proposed EM iterations from two aspects. First, we plot the energy $E_\alpha$ in each iteration, which is the energy of semantic-aware spare coding defined in Eq. (15), as shown in Fig. 6. We also present some intermediate results during the EM iterations$^3$ as Fig.

$^3$Generally, the iteration is complete after two or three steps since the average number of labels for each image is 3 in MSRC or VOC2007 dataset.
Fig. 4. Some final results (a) and some intermediate results of semantic segmentation (b) on the MSRC dataset. The original image and its ground truth are shown on the left, and the semantic segmentation result by our method is on the right. It’s encouraged to be viewed in color.

| Method          | building | grass    | tree | cow | sheep | sky | airplane | water | face | car | bicycle | flower | sign | bird | book | chair | road | cat | dog | body | boat | average |
|-----------------|----------|----------|------|-----|-------|-----|----------|-------|------|-----|---------|--------|------|------|------|-------|------|-----|-----|------|------|---------|
| MIM             | 12       | 83       | 79   | 81  | 93    | 84  | 91       | 55    | 97   | 87  | 92      | 82     | 69   | 51   | 61   | 59    | 66   | 53  | 44  | 9    | 58   | 67      |
| K. Zh           | 63       | 93       | 92   | 62  | 75    | 78  | 79       | 64    | 95   | 79  | 93      | 62     | 76   | 32   | 95   | 48    | 83   | 63  | 38  | 68   | 15   | 69      |
| Ours            | 45       | 73       | 65   | 79  | 81    | 66  | 71       | 87    | 75   | 84  | 73      | 73     | 94   | 51   | 89   | 85    | 83   | 81  | 66  | 32   | 71   | 55      |

| Method          | aeroplane | bicycle | bird | boat | bottle | bus | car | cut | chair | cow | diningtable | dog | horse | motorbike | person | potophot | sheep | sofa | train | tvmonitor | average |
|-----------------|------------|---------|------|------|--------|-----|-----|-----|-------|-----|-------------|-----|--------|-----------|---------|----------|--------|------|-------|------------|---------|
| Shotton,weakly  | 14         | 8       | 11   | 0    | 17     | 46  | 5   | 13  | 4     | 0   | 30          | 29  | 12     | 18        | 40      | 6        | 17     | 14   | 9     | 16         |         |
| K. Zh           | 48         | 20      | 26   | 25   | 3      | 7   | 23  | 13  | 38    | 19  | 15          | 39  | 17     | 18        | 25      | 47       | 9      | 41   | 17    | 33         | 24      |
| Ours            | 68         | 14      | 12   | 16   | 4      | 27  | 18  | 12  | 28    | 16  | 7           | 46  | 36     | 11        | 78      | 18       | 29     | 11   | 47    | 41         | 27      |

TABLE I

Accuracies (%) of our method for each category on MSRC and VOC 2007 dataset, in comparison with other algorithms. The last column is the average accuracy over all categories.
Compared with the traditional supervised learning methods, MLkNN and ML-LOC are the state-of-the-art multi-label annotation algorithms in literature. They have been reported to outperform most other multi-label annotating algorithms, such as RankSVM [1]. Thus, we do not plan to further implement the latter two in this work. We evaluate and compare among the three algorithms over two datasets, MSRC and VOC 2007, each of which is randomly and evenly split into training and testing subset. The image annotation performance is measured by mean average precision, which is widely used for evaluating the performances of ranking related tasks.

2) Results and Analysis: The weighed method is outperforms the unweighed one as Table II shown. It notices that the sparse coefficient $\alpha$ is useful to improve the image annotation performance, and useful for image semantic segmentation apparently, as we do the image retrieval by jointly matching their appearance as well as the semantics. The larger $\alpha$ means the more similar in semantics between the test image and image $I_i$ (i.e. sharing the more common labels).

The weighed method proposed outperforms the three classical methods listed in Table II. Some example image annotation results from the MSRC and VOC 2007 dataset are shown in Fig. 5. Here we only display the top 3 or 2 labels for MSRC and VOC 2007, since the average number of labels for each image in MSRC and VOC 2007 is 3 and 2 respectively.

![Image](image.png)

**Fig. 5.** Some example results on image annotation from the MSRC (left) and VOC 2007 dataset (right).

VI. CONCLUSIONS

In this paper proposes a new framework for data-driven semantic image segmentation where only image-level labels are available, and it is also useful for image annotation. Compared with the traditional supervised learning methods, our framework is more flexible for real applications such as online image retrieval. In the experiments, we demonstrate very promising results on the standard benchmarks of scene understanding. In future work, we can improve the algorithm efficiency by utilizing parallel implementation and validate our approach on larger scale datasets.

REFERENCES

[1] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In NIPS, volume 14, pages 681–687, 2001.
[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.
[3] S.-J. Huang, Y. Yu, and Z.-H. Zhou. Multi-label hypothesis reuse. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 525–533. ACM, 2012.
[4] S.-J. Huang, Z.-H. Zhou, and Z. Zhou. Multi-label learning by exploiting label correlations locally. In AAAI, 2012.
[5] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Graph cut based inference with co-occurrence statistics. In ECCV, Springer, 2010.
[6] L. Lin, X. Liu, and S.-C. Zhu. Layered graph matching with composite cluster sampling. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32(8):1426–1442, 2010.
[7] L. Lin, P. Luo, X. Chen, and K. Zeng. Representing and recognizing objects with massive local image patches. Pattern Recognition, 45(1):231–240, 2012.
[8] L. Lin, X. Wang, W. Yang, and J.-H. Lai. Discriminatively trained and-or graph models for object shape detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014.
[9] L. Lin, T. Wu, J. Porway, and Z. Xu. A stochastic graph grammar for compositional object representation and recognition. Pattern Recognition, 42(7):1297–1307, 2009.
[10] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In CVPR, IEEE, 2009.
[11] P. Luo, X. Wang, L. Lin, and X. Tang. Joint semantic segmentation by searching for compatible-competitive references. In Proceedings of the 20th ACM international conference on Multimedia, pages 777–780. ACM, 2012.
[12] R. M. Neal and G. E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In Learning in graphical models. Springer, 1998.
[13] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In CVPR, IEEE, 2008.
[14] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In ECCV, Springer, 2006.
[15] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised semantic segmentation with a multi-image model. In ICCV. IEEE, 2011.
[16] I. Winn and N. Jojic. Locus: Learning object classes with unsupervised segmentation. In ICCV. IEEE, 2005.
[17] K. Zhang, W. Zhang, Y. Zheng, and X. Xue. Sparse reconstruction for weakly supervised semantic segmentation. In IJCAI. AAAI Press, 2013.
[18] M.-L. Zhang and Z.-H. Zhou. ML-knn: A lazy learning approach to multi-label learning. Pattern recognition, 40(7):2038–2048, 2007.
Xionghao Liu is currently a graduate student majoring in Pattern Recognition and Computer Vision from Sun Yat-sen University, and has received his B.E. degrees in the School of Information Science and Technology, Sun Yat-sen University, Guangzhou, P. R. China, in 2012.

Wei Yang received his B.E. degree in Software Engineering, Sun Yat-sen University. He received his master degree in the Department of Computer Science, Sun Yat-sen University. He is currently a PhD student in the Department of Electronic Engineering, The Chinese University of Hong Kong. His research interests include computer vision and machine learning.

Liang Lin is a Professor with the School of Advanced Computing, Sun Yat-Sen University, China. His research focuses on new models, algorithms and systems for intelligent processing and understanding of visual data. He has published more than 60 papers in top tier academic journals and conferences, and has served as an associate editor for journal Neurocomputing and The Visual Computer.

Qing Wang is an Associate Professor of Sun Yat-Sen University, China, Ph.D. in Computer Science and Member of SIGCHI, researcher on Human Computer Interaction, User Experience, Collaborative Software, and Web Usability, and especially interested in utilizing browser history on collaboration.

ZhaoQuan Cai was born in 1970, and is now a professor in Huizhou University, China. His research interest include computer networks, intelligent computing, and database systems.

Jianhuang Lai is currently a Professor and the Dean of the School of Information Science and Technology. His research focuses on image processing, pattern recognition, multimedia communication, wavelet, and its applications. He serves as a Standing Member of the Image and Graphics Association of China and a Standing Director of the Image and Graphics Association of Guangdong.