Weakly Supervised Graph Propagation Towards Collective Image Parsing

Si Liu, Shuicheng Yan, *Senior Member, IEEE*, Tianzhu Zhang, Changsheng Xu, *Senior Member, IEEE*, Jing Liu, and Hanqing Lu, *Senior Member, IEEE*

***Abstract*—In this work, we propose a weakly supervised graph propagation method to automatically assign the annotated labels at image level to those contextually derived semantic regions. The graph is constructed with the over-segmented patches of the image collection as nodes. Image-level labels are imposed on the graph as weak supervision information over subgraphs, each of which corresponds to all patches of one image, and the contextual information across different images at patch level are then mined to assist the process of label propagation from images to their descendent regions. The ultimate optimization problem is efficiently solved by Convex Concave Programming (CCCP). Extensive experiments on four benchmark datasets clearly demonstrate the effectiveness of our proposed method for the task of collective image parsing. Two extensions including image annotation and concept map based image retrieval demonstrate the proposed image parsing algorithm can effectively aid other vision tasks.**

***Index Terms*—Concept map-based image retrieval, convex concave programming (CCCP), image annotation, nonnegative multiplicative updating, weakly supervised image parsing.**

I. Introduction

IMAGE parsing, which aims to decompose an image into semantically consistent regions, is a fundamentally challenging problem [1]–[4]. An effective image parsing facilitates many higher level image processing tasks, such as image editing [5], region-based image retrieval [6], and image synthesis [5]. Most existing image parsing methods suppose a training dataset with region-level annotations is provided and then either learn a region-based model which combines appearance and scene geometry model [7] or resort to nonparametric approaches to transfer the annotations from training images to query image [2]. However, it is generally too expensive and exhausting to annotate region level labels. Fortunately, with the popularity of online photo sharing sites, e.g., Flickr [8] (Flickr now hosts more than four billion images!), a large amount of images with user-provided labels become available. These labels can be even further refined by exploiting the correlations across images [9]. Therefore, a natural question arises of whether these labels can benefit the semantic parsing of corresponding images. In this work, we propose a graph-based approach to tackle the weakly supervised image parsing task, as illustrated in Fig. 1, where the inputs are a collection of images with annotations and the outputs are semantic consistent regions with corresponding labels.

The whole weakly supervised image parsing process can be broken down into two successive steps. First, adjacent pixels are merged to generate semilocal semantically consistent patches; second, labels are assigned to those patches. The first step is necessary here due to the following two reasons. First, the number of image patches is significantly smaller than that of pixels, which brings much computational convenience. Second, patches are bigger and can be better semantic carriers. Features extracted from patches are more discriminative than raw pixel color values. In this work, bottom-up image segmentation is utilized to generate patches. So far, our weakly supervised image parsing task has been transformed into how to assign image labels to these over-segmented patches. The major stumbling block of the label assignment task is that image labels and patch labels lie at different granularity levels.

To facilitate the label propagation, we start from the following three observations. First, even though patches cannot directly inherit labels from their parent images, they can still be constrained in a weakly supervised way. The image labels provide a candidate set for its patches. If the image labels are complete and correct, the image labels should be the same as the union of all of its patch labels. For example, in Fig. 1, if the image "B" is labeled with "grass, cow," then all of its patches in subgraph B can only be labeled by either "grass" or "cow." Other labels beyond this scope are not allowed. What's more, each image label must interpret at least one image region. In other words, it is forbidden to label all patches of image "B" as "grass," leaving the label "cow" unexplained.

Second, if there are in total k labels in one image, and the number of patches in an image is n, then there exists approximately k^n possible kinds of patch label assignments. Picking out the optimal assignment from such a large candidate set is

Manuscript received December 19, 2010; revised July 16, 2011; accepted October 19, 2011. Date of publication November 04, 2011; date of current version March 21, 2012. This work was supported in part by the "NExT Research Center" funded by MDA, Singapore, under Grant WBS:R-252-300-001-490, 973 Program under Project 2010CB327905 and Project 2012CB316304, and the National Natural Science Foundation of China under Grant 60903146, Grant 60833006, and Grant 90920303. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ketan Mayer-Patel.

S. Liu, C. Xu, J. Liu, and H. Lu are with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the China-Singapore Institute of Digital Media, Singapore 119613, Singapore (e-mail: sliu@nlpr.ia.ac.cn; csxu@nlpr.ia.ac.cn; liujingmgm@gmail.com; luhq@nlpr.ia.ac.cn).

T. Zhang is with the Advanced Digital Sciences Center (ADSC), Singapore 138632, and also with the China-Singapore Institute of Digital Media, Singapore 119613, Singapore (e-mail: tzzhang@nlpr.ia.ac.cn).

S. Yan is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576, and also with the China-Singapore Institute of Digital Media, Singapore 119613, Singapore (e-mail: eleyans@nus.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2011.2174780

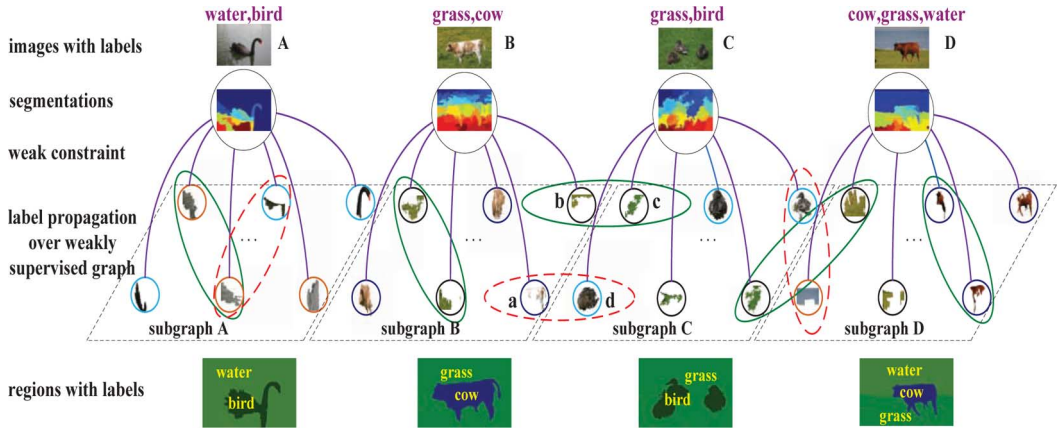


Fig. 1. Schematic illustration of the weakly supervised graph-based label propagation process. The inputs are images with labels and outputs are regions with labels. The images are first segmented into semantically unique patches, which constitute the nodes of a graph, then by mining contextual relationships among patches and taking the labels of individual images as weak supervision information over subgraphs, the labels of patches are inferred over the entire graph.

a severely under-constrained problem for a single image. To assist individual image parsing, contextual information among different image patches can be explored. Specifically, two types of context, namely consistency and incongruity, can be mined. Consistency indicates two patches share the same label with high probability. For example, in Fig. 1, if patch "b" from image "B" is visually similar with patch "c" from image "C," we can infer that patch "b" and "c" stand a good chance to be annotated with the same label, i.e., "grass." Incongruity indicates that two patches are annotated by different labels with high probability. In most cases, if patch "a" is significantly different from patch "d," they are highly unlikely to share common labels. These two types of contextual information supplement each other and are both indispensable in the collective image parsing task.

Third, the number of semantically meaningful regions is still quite large with the consideration of the large amount of images, therefore, it is critical for the proposed solution to be computationally efficient.

To tackle the above issues, we propose a weakly supervised graph (WSG)-based algorithm towards collective image parsing. First, we segment each image into multiple smaller size patches and harness them as the basic processing units. Then, a WSG is constructed by treating the patches from the image collection as nodes, where two types of edges are linked to describe the consistency and incongruity contextual relationships among patches. In Fig. 1, these two types of edges are denoted by green (online version) solid and red (online version) dotted ellipses, respectively. With these two types of contextual information and the image labels serving as weak constraints (shown as purple (online version) solid arcs between images and their patches), the labels for all patches can be inferred by solving a constrained optimization problem. By merging those semantically consistent patches within each image, the ultimate image parsing is achieved. To solve the constrained optimization problem, a Convex Concave Programming (CCCP) [10] based iterative procedure is adopted, and, for each iteration of CCCP, a nonnegative multiplicative updating procedure is derived for further reducing the computational burden.

The main contributions of this work can be summarized here.

- We are the first to formulate the collective image parsing as a weakly supervised graph propagation problem. The labels of each image are imposed as weak overall supervision information over a subgraph corresponding to its oversegmented patches. In addition, contextual information across different images at patch level are mined to facilitate the label propagation from images to their descendant patches.
- We solve the constrained optimization problem using the CCCP and a nonnegative multiplicative updating procedure, the latter of which is quite efficient and thus scalable for large datasets.
- Our image parsing algorithm has many applications and extensions, such as image annotation and semantic image retrieval. Extensive experiments demonstrate that comprehensive image parsing can effectively aid other vision tasks.

The rest of the paper is organized as follows. In Section II, we briefly review existing work in image parsing field. The proposed weakly supervised collective image parsing method is introduced in Section III. Section IV shows how to efficiently solve the proposed formulation. Extensive image parsing results in four publicly available datasets: MSRC, COREL-100, NUS-WIDE and VOC-07 datsets are reported in Section V. Some applications of image parsing are demonstrated in Section VI. Finally, the conclusion and future work are discussed in Section VII.

II. Related Work

The image parsing problem has received wide interests in the computer vision community, and numerous methods have been proposed. These methods could be roughly divided into two categories. The first category focuses on unsupervised or weakly supervised learning techniques [3], [9], [11]–[13]. Chen *et al.* [3] propose to learn explicit shape models from a single image. Winn *et al.* [11] show that object labels can be learned based on the results of automatic image segmentation. An extension is presented by Cao *et al.* in [13], which applies the spatially coherent latent topic model to conduct multilabel image

segmentation and classification. Eran *et al.* [14] learned to perform tag-based segmentation using only unsegmented training examples. In [15], the authors construct bags of multiple segmentation of an image and learn to extract the one containing an object. These algorithms, however, can only handle images either with a single major object or without occlusions between objects. In contrast, the research in this paper aims to process more challenging images containing multiple objects and with possible inter-object occlusions.

The second category is found on supervised learning techniques [2], [5], [16]–[18]. Liu *et al.* [2] propose a nonparametric approach for object recognition and image parsing by using dense scene alignment. Yuan *et al.* [16] use conditional random fields (CRF) and max-margin Markov networks (M3N) to solve the scene understanding task. Shotton *et al.* [5] learn a discriminative model of object labels, incorporating appearance, shape, and context information efficiently. Gould *et al.* [17] propose a method for capturing global information from inter-class spatial relationships and encoding it as a local feature. Tu *et al.* [18] present a Bayesian framework for parsing images into their constituent visual patterns. The parsing algorithm optimizes the posterior probability and outputs a scene representation in a "parsing graph." All of these methods need pixel-level labels for training purpose, however, which are very tedious to be manually annotated in practice. On the contrary, the proposed solution in this paper is based on image-level labels, which can be easily obtained from the public photograph-sharing websites with a proper label-refinement process.

Graph-based label propagation has been widely used in semisupervised learning tasks, where each sample is used as a node and the edge weights provide measures on the similarities between nodes. He *et al.* [19] adopt the manifold ranking algorithm for image retrieval. Pan *et al.* [20] propose to preform random walk on a graph to find the correlations between image features and keywords. These methods assume that the labels of some nodes are available, and this information is then propagated from the labeled data to unlabeled ones. In our problem, as introduced afterward, the labels of all nodes (image patches) are not exactly known; instead, only weakly supervised label information at image level is available (acting as label set for a subset of nodes within a graph), and thus the task is more challenging.

Our work also resembles cosegmentation task that tries to find the common structures between a pair of images and segment a similar object in two images simultaneously [21]–[25]. Mukherjee's cosegmentation algorithms can also be extended to handle multiple images, but cosegmentation is generally designed for the segmentation of foreground and background, yet not suitable for direct collective multilabel image parsing. What's more, the key difference is that the cosegmentation requires histograms of all segmented foregrounds to be alike, while we only require each patch to be similar to few patches of its label. That is why we claim "our algorithm loosens this requirement."

A more related work is done by Liu *et al.* [1], who also fulfill the weakly supervised scene parsing task (named label-to-region in [1]) and propose a bi-layer sparse coding formulation for uncovering how an image or semantic region can be robustly reconstructed from the over-segmented image patches of an image set. As we shall introduce later in the experiment part, the work in [1] suffers from the high computational cost, and the boundaries of the segmented regions are often not quite precise.

III. WSG Propagation

Here, we first introduce some notations, and then further explain how to construct the so-called WSG, which encodes two types of contextual information among image patches, i.e., consistency and incongruity. Finally, the collective image parsing task is formulated as a constrained optimization problem.

Notations

Given a data collection $\{X_1, \ldots, X_m, \ldots, X_M\}$, where X_m denotes the mth image, and its corresponding label information is denoted as an indicator vector $y_m = \left[y_m^1, \ldots y_m^c \ldots, y_m^C\right]^T$, where $y_m^c = 1$ if the image m has the cth label, otherwise $y_m^c = 0$, and C denotes the total number of unique image labels. After image over-segmentation with a certain approach, e.g., [26], the mth image is represented as $X_m = \{x_{m,1}, \ldots, x_{m,j}, \ldots x_{m,n_m}\}$, where n_m is the number of patches for the mth image, $x_{m,j}$ is the representation for the jth patch of image X_m, and its label information is also denoted as an indicator vector $f_{mj} = \left[f_{mj}^1, \ldots, f_{mj}^c \ldots, f_{mj}^C\right]^T$, where $f_{mj}^c = 1$ if image $x_{m,j}$ has the cth label, otherwise $f_{mj}^c = 0$. Let $N = \sum_{m=1}^{M} n_m$ be the total number of patches in the whole data collection. In the whole graph propagation process, the image labels y are known, and the patches's labels f are to be inferred.

A. Graph Construction

For graph-based label propagation algorithms, how to construct the graph is critical. In this work, the nodes are oversegmented image patches, and the ideal edge weights should measure the semantic relationships among the nodes. Here, the semantic relationships include two types of contextual information, one is the consistency relationship, and the other is the incongruity relationship.

1) Consistency Relationship Mining: Several recent algorithms [27]–[29] show that sparse coding is an effective way to uncover the consistency relationship among the nodes. Therefore, we employ sparse coding to build the consistency relations among image patches. Specifically, we reconstruct each image patch (based on the extracted texton features [5], which encode both texture and color information) as a sparse linear combination of the rest image patches coming from images with at least one common label. The image patches with nonzero reconstruction coefficients are considered to be similar with the reconstructed patch.

Let $\hbar$ denote all of the feature vectors of the image patches, where $\hbar \in R^{d \times N}$. Note that $\hbar$ is columnly normalized with unitary ℓ_2-norm. For any given $\hbar_k$, namely, the kth column vector of $\hbar$, we reconstruct $\hbar_k$ as a sparse linear combination of the rest column of $\hbar$, i.e., $\mathbb{Z}_k = [\hbar_1, \ldots, \hbar_{k-1}, \hbar_{k+1}, \ldots, \hbar_N]$. Let φ denotes the coefficient of the derived sparse coding. Then, φ is derived by solving the following optimization problem:

$$\min \|\varphi\|_{\ell_1}, \quad s.t. \quad \mathbb{Z}_k \varphi = \hbar_k. \tag{1}$$

This optimization problem is convex and can be transformed into a general linear programming problem. Thus, there exists a globally optimal solution, which can be solved efficiently using the classical ℓ_1-norm optimization toolkit [30]. With the derived sparse coding coefficients φ from (1), the similarity between $\hbar_k$ and $\hbar_i$ is defined by: $S_{ki} = \varphi_i$ if $i < k$, and $S_{ki} = \varphi_{i-1}$ otherwise. To ensure the symmetry of the similarity matrix, the final similarity between $\hbar_k$ and $\hbar_i$ is defined as $W_{ki} = S_{ik} + S_{ki}/2$. Note that, to ensure the nonnegativity property of W, the nonnegativity constraints for the reconstruction coefficients are posed for the sparse coding formulation.

2) Incongruity Relationship Mining: To comprehensively mine the contextual information among the image patches, we introduce another graph characterizing the incongruity relationship. In this graph, the edge weight denotes patch dissimilarity. The higher the edge weight is, the less likely the nodes at the two ends are to be assigned with the same label. The weight used in this work is the traditional χ^2 distance defined as: $\chi^2(v^1, v^2) = \sum_{i=1}^{d} \left(v_i^1 - v_i^2\right)^2 / v_i^1 + v_i^2$, where v^1 and v^2 are two histogram feature vectors of these image patches and d is the feature dimension. To guarantee robustness to noises, we sparse the graph by maintaining the K farthest neighbors for each image patch and setting other weights to be zero. Note that, for each patch, we only require its K most dissimilar patches to be labeled differently. We assume that in most cases, the extremely dissimilar patches should be from different labels. Therefore, we put this observation into our formulation.

B. Label Propagation

Based on the derived consistency relationship graph and incongruity relationship graph, our task is to propagate labels from images to patches. To obtain the mathematical formulation for this task, the following factors need be taken into consideration.

1) Patch Label Self-Constraints: Directly inferring the indicator vector f_{mj} is difficult, and, for this reason, we relax its value range to be $[0, 1]$. Since f_{mj}^c represents the probability of assigning label c to patch x_{mj}, it is natural to require that $\sum_{c=1}^{C} f_{mj}^c = 1, \forall\ m,\ j$. In matrix form, it is $Pe_1 = e_2$, where $e_1 = \mathbf{1}_{C\times 1}$, $e_2 = \mathbf{1}_{N\times 1}$, $P = [f_{11}, \ldots, f_{1n_1}, \ldots, f_{Mn_1}, \ldots, f_{Mn_M}]^T$ and $P \in R^{N\times C}$ represents the derived patch labels. As f_{mj} is a probability vector, it is natural to constrain that $P \geq 0$.

2) Patch–Patch Contextual Relationships: To integrate the consistency relationship between patches into formulation, we introduce the following patch–patch mutual consistency loss:

$$O_{con} = \sum_{m_1,m_2} \sum_{j_1,j_2} W_{m_1 j_1, m_2 j_2} \left\| f_{m_1 j_1} - f_{m_2 j_2} \right\|^2 \quad (2)$$

where the patch similarity matrix $W_{m_1 j_1, m_2 j_2}$ is obtained by sparse coding as introduced in Section III-B1, m_1, m_2 are image indexes and j_1, j_2 are patch indexes. In matrix form, we have

$$O_{con} = Tr\left(P^T L P\right) \quad (3)$$

where L is a Laplace matrix $L = D - W$, where D is the degree matrix and $Tr(\cdot)$ representing matrix *trace* operator. It is a smoothness constraint enforcing adjacent points in feature space to share similar labels.

We also introduce an incongruity loss term

$$O_{\text{incong}} = \sum_{m_1,m_2} \sum_{j_1,j_2} B_{m_1 j_1, m_2 j_2} \left(f_{m_1 j_1} \circ f_{m_2 j_2}\right) \quad (4)$$

where the operator $\circ$ is the inner-product and $B_{m_1 j_1, m_2 j_2}$ denotes the χ^2 distance between $x_{m_1 j_1}$ and $x_{m_2 j_2}$, as shown in Section III-B2. If $B_{m_1 j_1, m_2 j_2}$ is large, $f_{m_1 j_1} \circ f_{m_2 j_2}$ should be small to minimize this loss term, that is, $x_{m_1 j_1}$ and $x_{m_2 j_2}$ can not contain common labels. During this sparse incongruity graph construction process, each patch only links with its K most dissimilar patches. Therefore, we only weakly constrain very dissimilar patches to be labeled differently. With $B \in R^{N\times N}$, in matrix form, we have

$$O_{\text{incong}} = Tr\left(BPP^T\right). \quad (5)$$

3) Image-Patch Inclusion Supervision: We summarize the weakly supervised label information imposed by image labels with the following objective function:

$$O_{\text{inc}} = \sum_{m=1}^{M} \sum_{c=1}^{C} \left| \max_{x_{mj}\in X_m} f_{mj}^c - y_m^c \right|. \quad (6)$$

Equation (6) makes sense due to the following reasons. If $y_m^c = 1$, namely, the mth image contains the cth label, then at least one of its patches should interpret the label, that is $\max f_{mj}^c$ must be close to one. If the image labels are complete and $y_m^c = 0$, namely the image does not include the cth label, then none of its patches can be assigned with the label, which is equal to require that $\max f_{mj}^c$ be close to 0. In the image labels missing cases, our algorithm can still parse image by introducing the "background" label. However, directly dealing with the absolute symbol in (6) is difficult, we take $f_{mj}^c \in [0, 1]$ into consideration and decompose (6) into two separate parts according to different value of y_m^c as

$$\left| \max_{j\in X_m} f_{mj}^c - y_m^c \right| = \begin{cases} 1 - \max_{x_{mj}\in X_m} f_{mj}^c & \text{if } y_m^c = 1 \\ \max_{x_{mj}\in X_m} f_{mj}^c & \text{otherwise} \end{cases}. \quad (7)$$

By inserting (7) into (6), we can get

$$\begin{aligned} O_{\text{inc}} = & \sum_m \sum_c \left(1 - y_m^c\right) \max_{x_{mj}\in X_m} f_{mj}^c \\ & + \sum_m \sum_c y_m^c \left(1 - \max_{x_{mj}\in X_m} f_{mj}^c\right). \end{aligned} \quad (8)$$

The first term of (8) can be further approximated by $\sum_m \sum_c \left(1 - y_m^c\right) \sum_{x_{mj}\in X_m} f_{mj}^c = \sum_m \sum_c \left(1 - y_m^c\right) h_c P^T q_m$, where h_c is a $1 \times C$ indicator vector whose all elements, except for the cth element, are zeros and $q_m = [\underbrace{0, \ldots, 0}_{1,\ldots,m-1}, \underbrace{1, \ldots, 1}_{m}, \underbrace{0, \ldots, 0}_{m+1,\ldots,M}]^T$ is an $N \times 1$ vector whose all elements, except for those elements corresponding to the mth image, are zeros. An intuitive explanation is that, if we require the maximum value of a

positive data set to be zero, we actually require all the data in the set to be zero. Therefore, the (8) could be rewritten as

$$O_{\text{inc}} = \sum_m \sum_c (1 - Y_{mc})\, h_c P^T q_m + \sum_m \sum_c Y_{mc} \left(1 - \max_{x_{mj} \in X_m} g_{mj} P h_c^T\right) \quad (9)$$

where $Y = [y_1, \ldots, y_m, \ldots y_M]^T$, g_{mj} is a $1 \times N$ indicator vector whose elements corresponding to the jth patch in the mth image are ones and other elements are zeros. Note that the second term in (9) is concave, and we will discuss how to optimize it in detail in Section IV-A.

4) Unified Objective Function: Aiming to carry out all of the objectives aforementioned, we optimize a unified objective function $O_{\text{all}} = \lambda_1 O_{\text{con}} + \lambda_2 O_{\text{incong}} + O_{\text{inc}}$ as

$$\begin{aligned} \min_P \quad & \lambda_1 Tr\left(P^T L P\right) + \lambda_2 Tr\left(B P P^T\right) \\ & + \sum_m \sum_c (1 - Y_{mc})\, h_c P^T q_m \\ & + \sum_m \sum_c Y_{mc} \left(1 - \max_{x_{mj} \in X_m} g_{mj} P h_c^T\right) \\ s.t. \quad & P \geqslant 0, P e_1 = e_2 \end{aligned} \quad (10)$$

where λ_1, λ_2 are two positive parameters for balancing all three objectives.

IV. Optimization Procedure

Considering (10), the first three terms are convex, while the last term is concave. Thus, we adopt the popular CCCP [10] to search for the suboptimum solution.

A. Optimization Via CCCP

CCCP is guaranteed to converge [10] although working in an iterative way. At each iteration, the first-order Taylor expansion is used to approximate the nonconvex function, and the problem is thus approximated by a convex optimization problem locally. The suboptimum solution is achieved by iteratively optimizing the convex subproblem until convergence.

Note that $\max(f)$ is a nonsmooth function with respect to f. To use CCCP, we have to replace the gradients by the subgradients. Since the last term in (10) is a summarization term, we consider only the term related with Y_{mc}. Let $l = [f_{m1}^c, \ldots, f_{mj}^c, \ldots f_{mn_m}^c]^T$, we pick the subgradient of l with η, which is an $n_m \times 1$ vector and its jth element is given by

$$\eta_j = \begin{cases} \frac{1}{n_\alpha}, & \text{if } {l_j}^{(t)} = \max\left({l_m}^{(t)}\right) \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

where n_α is the number of patches with the largest label value $\max l^{(t)}$. At the $(t+1)$th iteration, we estimate the current l based on $l^{(t)}$ and the corresponding $\eta_j^{(t)}$. As $\eta^T l^{(t)} = \sum_j \eta_j l_j^{(t)} = \max l^{(t)} \sum_{\eta_j \neq 0} \eta_j = \max l^{(t)}$, for the function $\max(l)$, its first-order Taylor expansion is approximated as $(\max l)_{l^{(t)}} \approx \max l^{(t)} + \eta^T \left(l - l^{(t)}\right) = \max l^{(t)} + \eta^T l - \max l^{(t)} = \eta^T l$, which could also be rewritten in matrix form as

$$\sum_m \sum_c Y_{mc} \left(1 - h_c \eta U_m P h_c^T\right) \quad (12)$$

where U_m is an $N \times N$ diagonal block matrix, $U_m = diag\,(u_1, \ldots, u_m, \ldots u_M)$, $u_k = 0_{n_k \times n_k}$ for $k = 1, \ldots, m-1, m+1, \ldots M$ and $u_m = I_{n_m \times n_m}$, I indicates identity matrix. β is a $C \times N$ matrix, $\beta = [\beta_1, \ldots, \beta_m, \ldots, \beta_M]$, each $\beta_m = [\beta_{m1}^T, \ldots, \beta_{mc}^T, \ldots, \beta_{mC}^T]^T$ is a $C \times n_m$ matrix corresponding to image m and $\beta_{mc} = \eta^T$. With (12), (10) could be rewritten as the following quadratic programming problem:

$$\begin{aligned} \min_P \quad & \mathrm{Q(P)} = \lambda_1 Tr\left(P^T L P\right) + \lambda_2 Tr\left(B P P^T\right) \\ & + \sum_m \sum_c (1 - Y_{mc})\, h_c P^T q_m \\ & + \sum_m \sum_c Y_{mc} \left(1 - h_c \beta G_m P h_c^T\right) \\ \text{s.t.} \quad & P \geqslant 0, P e_1 = e_2. \end{aligned} \quad (13)$$

Equation (13) can be directly solved using classical quadratic programming methods [31]. However, the computational cost is extremely high for a large-scale dataset. Therefore, we propose an efficient nonnegative multiplicative updating procedure to iteratively refine the solution and reduce the computational burden.

B. Nonnegative Multiplicative Updating Procedure

Specifically, we solve the subproblem in (13) with the general nonnegative data factorization approach introduced in [32]. The efficient updating procedure is available based on the following theorem.

Theorem 4.1: [32] Given a quadratic function $Q(P)$ with its first-order derivative as $\partial Q(P)/\partial P = \sum_{l=1}^K A_l P E_l + H$, where A_l, E_l, and H are real-value constant matrices. By decomposing the matrix A_l, B_l, and H as the difference of two nonnegative parts (one part is zero), denoted as $A_l = A_l^+ - A_l^-$, $E_l = E_l^+ - E_l^-$, $H = H^+ - H^-$, respectively, the update rule to search for the local optimum is as follows:

$$P_{ij} \leftarrow P_{ij} \times \frac{\left[\sum_{l=1}^K \left(A_l^+ P E_l^- + A_l^- P E_l^+\right) + H^-\right]_{ij}}{\left[\sum_{l=1}^K \left(A_l^+ P E_l^+ + A_l^- P E_l^-\right) + H^+\right]_{ij}}. \quad (14)$$

By relaxing the hard constraints of $Pe_1 = e_2$ in (13) into soft ones, we obtain the soft-version objective formulation

$$\min_P Q(P) + \delta \|P e_1 - e_2\|^2, s.t.\ P \geqslant 0 \quad (15)$$

where δ is a tunable weighting parameter. Then, we obtain the following corollary to the solution to (15). The proof can be easily derived by calculating the derivative of the objective function in (15) with respect to P and then replacing the matrices of A_1, B_l and C in Theorem 4.1 with the corresponding values.

Corollary 4.2: Equation (15) can be optimized based on the nonnegative multiplicative updating rule as

$$P_{ij} \leftarrow P_{ij} \times \frac{\left[2\lambda_1 WP + 2\delta e_2 e_1^T + \sum_m \sum_c Y_{mc} G_m^T \beta^T h_c^T h_c\right]_{ij}}{\left[2\lambda_1 DP + 2\lambda_2 BP + 2\delta P e_1 e_1^T + + \sum_m \sum_c (1 - Y_{mc}) q_m h_c\right]_{ij}}. \tag{16}$$

The entire algorithm for solving the problem (10) is summarized in Algorithm 1, where t_1 and t_2 are the maximum iteration number for outer and inner iterations, respectively, and ε_1 and ε_2 are two thresholds to determine the minimum solution differences of two successive iterations for the outer and inner iterations to stop. In all experiments, t_1, t_2, ε_1 and ε_2 are set to 25, 20, 0.01, and 0.1, respectively.

Algorithm 1 Procedure to solve problem **(10)** via CCCP

1: Initialize $P, T_1, T_2, \varepsilon_1, \varepsilon_2, t_1 = 0, t_2 = 0$.

2: **while** $\left\|P_1^{t_1} - P\right\| > \varepsilon_1$ and $t_1 < T_1$ **do**

3: $t_1 = t_1 + 1, P_1^{t_1} = P, t_2 = 0$.

4: Calculate β^{t_1} based on the η from **(11)**.

5: **while** $\left\|P_2^{t_2} - P\right\| > \varepsilon_2$ and $t_2 < T_2$ **do**

6: $t_2 = t_2 + 1, P_2^{t_2} = P$.

7: Update P by **(16)**.

8: **end while**

9: **end while**

10: Obtain P as the solution to problem **(10)**.

V. Image Parsing Experiments

Here, we systematically evaluate the effectiveness of the proposed WSG-based label propagation algorithm in image parsing task over four widely used datasets.

A. Image Datasets

We perform the experiments on four publicly available benchmark image datasets: MSRC [5], COREL-100 [6], NUS-WIDE [33], and VOC-07 [34].

The first dataset used in the experiments is the MSRC image database which contains 591 images from 23 labels and scene categories with region-level ground truths. Around 80% images are associated with more than one label, and there are around three labels per image on average. However, some labels have only few positive samples, and some are only coarsely associated with one or few labels. We select the same database as in [1], focusing on relatively well-labeled labels, yielding 355 images and 18 different labels: "$building$," "$grass$," "$tree$," "cow," "sky," "$sheep$," "$mountain$," "$airplane$," "$water$," "$bird$," "$road$," "$boat$," "$sign$," "$book$," "$flower$," "$chair$," "cat," and "dog". The MSRC dataset provides pixel-level groundtruth, and thus we may evaluate the image parsing performance at the pixel level.

The second dataset is COREL-100. Each image is annotated with about 3.5 labels on average, and region-level groundtruth is provided for evaluation. It contains the images from seven categories of: "$grass$," "$snow$," "cow," "sky," "$bear$," "$ground$," and "$water$."

The third dataset named NUS-WIDE is a more challenging collection of real-world social images from Flickr. It contains 269 648 images in 81 categories and has about two labels per image. We select a subset of this dataset, focusing on images containing at least five labels and obtain a collection of 18 325 images with 41 989 unique labels. We refer to this social image dataset as NUS-WIDE-SUB, and the average number of labels for each image is 7.8.

The fourth dataset named VOC-07 Segmentation Dataset [34]. This dataset contains 21 extremely challenging labels including background and contains 209 training data, 213 validation data, and 210 testing images. We utilize the labels of all training, validation, and testing images to infer pixels' labels. Even supervised methods do not yield satisfactory results on this data. We also report results of our algorithms on this dataset.

B. Experiment Setup

Three algorithms are implemented as baselines for comparison to evaluate the effectiveness of the proposed image parsing algorithm, given here.

1) The first one is the classical binary support vector machine (BSVM). To handle a multilabel situation, C one-versus-all binary classifiers are trained. Note that classifiers are trained at the image level and tested at the patch level. During training, one image is considered as a positive sample if it contains the specific label. During testing, we apply all classifiers on each patch and combine the results to generate the C-dimensional label confidence vectors. The Gaussian kernel is used by setting the kernel parameter as 1. The result is denoted as BSVM.
2) The kNN algorithm. For each segmented patch of an image, we first select its k nearest neighbors from the whole semantic patch collection and collect the images containing these patches into an image subset. Then, we count the occurrence number of each label in the obtained image subset and choose the most frequent label as the label of the given patch. We apply this baseline with different parameter setting, i.e., $k = 49$ and 99, and thus can obtain two results from this baseline. The results are denoted as kNN-1 and kNN-2, respectively.
3) The bi-layer sparse coding (bi-layer) algorithm in [1]. This algorithm uses a bi-layer sparse coding formulation to uncover how an image or semantic region can be robustly reconstructed from the oversegmented image patches of an image set and is the state-of-the-art algorithm for weakly supervised image parsing assignment task. The performance is measured by pixel-level accuracy which is the percentage of pixels with agreement between the assigned labels and the groundtruth.

The proposed algorithm starts with image oversegmentation to divide each image into multiple homogeneous patches. Here we choose to use the mean shift segmentation approach proposed in [26]. We divide each image into around 20 coherent patches. Note that the proposed solution is general and not tied

TABLE I
COLLECTIVE IMAGE PARSING ACCURACY COMPARISON ON MSRC, COREL-100, AND VOC-07 DATASETS

Dataset	BSVM	kNN-1	kNN-2	Bi-Layer [1]	**Ours**
MSRC	0.24	0.45	0.37	0.63	**0.71**
COREL-100	0.33	0.52	0.44	0.61	**0.64**
VOC-07	0.15	0.22	0.20	0.32	**0.38**

to any specific image segmentation algorithm. After image oversegmentation, texton features, which encode both texture and color information, are extracted. The training images are first convolved with a 17-dimensional filter-bank. Then, K-means clustering is performed. Finally, each pixel in each image is assigned to the nearest cluster center. Based on the texton map, normalized texton histogram within a region is computed as the patch feature descriptor. In our experiments, 350 textons are used and the descriptor is therefore of 350 dimensions. When constructing the consistency relationship by sparse coding, we set the tolerance factor as 0.003 and the maximum number of primal-dual iterations as 50. For the incongruity relationship mining, we set the number of farthest neighbors as $K = 600$ on MSRC, NUS-WIDE-SUB and VOC-07 dataset ,and $K = 300$ on COREL-100 dataset.

C. Results and Analysis

Table I shows the accuracy comparison between baseline algorithms and our algorithm on MSRC, COREL-100 and VOC-07 datasets. From these results, we can draw the following conclusions.

1) Compared with baselines, the proposed WSG algorithm achieves much higher accuracies of 0.71, 0.64, and 0.38 on the MSRC, COREL-100, and VOC-07 dataset respectively. This clearly demonstrates the effectiveness of the proposed method in the image parsing task.
2) Since the BSVM classifier is trained at the image level and tested at the patch level, it performs worst. It shows that cross-level label inference is not trivial, and straightforward propagating labels from images to patches is not applicable. A more sophisticated method is required to weakly impose image labels upon their descendent patches.
3) The contextual image parsing algorithms, including KNN, the bi-layer, and the proposed WSG-based algorithms, all outperform the BSVM-based counterpart. This is because the former three harness the contextual information among the semantic regions in the image collection.
4) The WSG-based algorithm clearly beats the state-of-the-art bi-layer sparse coding algorithm, which owes to the fact that the weakly supervised information of graph avoids the ambiguities among the smaller patches in the bi-layer sparse coding algorithm and WSG can make use of both consistency and incongruity relationships among patches while the bi-layer method mainly focuses on the consistency relationship.

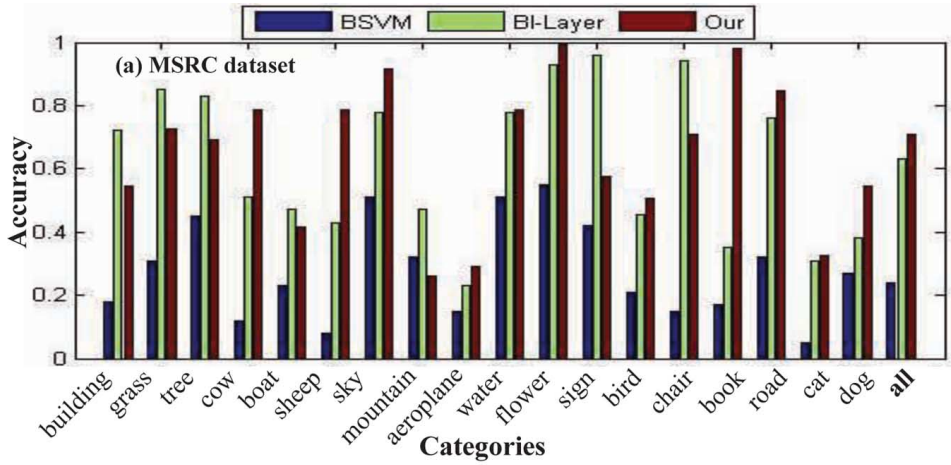
The detailed comparison results for individual labels are illustrated in Fig. 2. On MSRC dataset, we have 11 out of total 18 labels better than the bi-layer method; on COREL-100 dataset, we have five out of a total of seven labels better and in VOC-07 dataset, we have 17 out of 21 labels better than the bi-layer baseline. The results demonstrate the superiority of the proposed algorithm over baseline algorithms. Some image parsing results are displayed in Fig. 3. These results over various conditions further validate the effectiveness of the proposed solution. Also, if we compare these results with those reported in [1], we may observe that the boundaries of the segmented regions are much more precise. We do not further compare our solution with those algorithms for classifying and localizing objects in images [2], [5], [6], because the proposed solution works without region-level label annotation, which is, however, the general prerequisite for most typical algorithms.

As we solve the WSG-based label propagation problem via CCCP, which is an iterative optimization method, the convergence speed is critical. We show in Fig. 5(a) the accuracies of three datasets with respect to the iterative number. We can see that on all datasets, the accuracies continuously increase till convergence. From these results, we can conclude that during iteration, information is propagated through the graph effectively. To further show the optimization progress of the proposed algorithm, we give some detailed intermediate image parsing results for different iterations in Fig. 4, and we can observe that the image parsing results become better and better as the iteration goes. Thanks to the propagation, visually consistent patches are enforced to merge and visually dissimilar patches are labeled differently. Therefore, the object boundaries become more and more meticulous during the iteration.

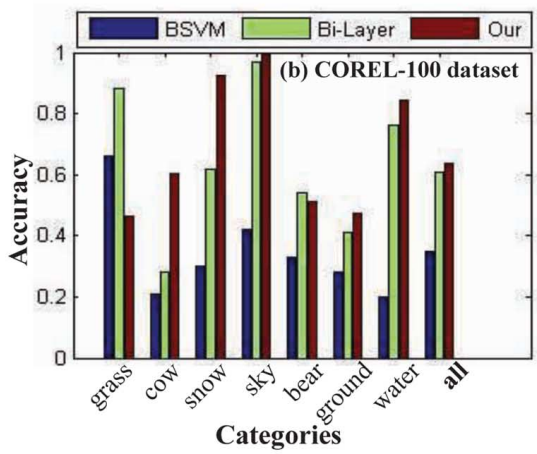
All of the experiments are run on a server with Intel Xeon(R) 2.4-GHz CPU 16-GB memory, and the code is run in MATLAB platform. As CCCP converges quite fast and the nonnegative multiplication updating procedure only involves matrix multiplication, WSG achieves the local optimum very fast. Fig. 5(b) shows the comparison of computational cost between WSG and Bi-layer [1] for three datasets. In detail, only 9.9, 1.3, and 17.6 min are spent on MSRC, COREL-100, and VOC-07 datasets, respectively. Therefore, the proposed algorithm is scalable to large-scale applications. Because the pixel-label groundtruth label is not provided for the NUS-WIDE-SUB dataset, we can not quantitatively report the image parsing results. However, as shown in Section VI, image parsing facilitates image annotation task, which can be quantitatively evaluated. Superior performance over baselines in image annotation task in return validates the effectiveness of our method in the earlier image parsing step.

D. Discussions

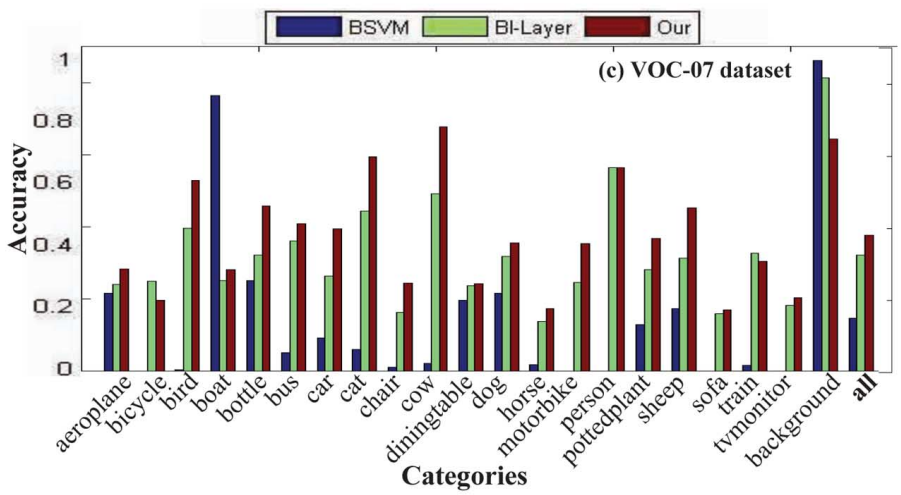
Our work focuses on how to propagate the image labels to their regions and assumes that the image labels have been properly labeled. Unfortunately, in real applications, e.g., for images at flickr.com, image labels are provided by users and inevitably noisy. In this situation, we can first preprocess the image labels with the state-of-the-art label-refinement algorithm [35]–[37] and then run our algorithm based on the refined and cleaner labels.



(a)



(b)



(c)

Fig. 2. Detailed collective image parsing accuracies for (a) MSRC dataset, (b) COREL-100 dataset, and (c) VOC-07 dataset. The horizontal axis shows the name of each label and the vertical axis represents the collective image parsing accuracy.

VI. Extensions

Here, we extend the proposed image parsing framework from two aspects, including image annotation and semantic image retrieval.

A. Extension I: Image Annotation

Automatic image annotation is of immense interest as it allows one to exploit the fast indexing and retrieval architecture of existing search engines. There are two key differences between the weakly supervised image parsing task and image annotation task. First, instead of knowing all images' labels in advance as in image parsing task, only training images are labeled while test images' labels are left to be estimated. Second, we are more concerned about whole test images' labels, but not detailed patches labels.

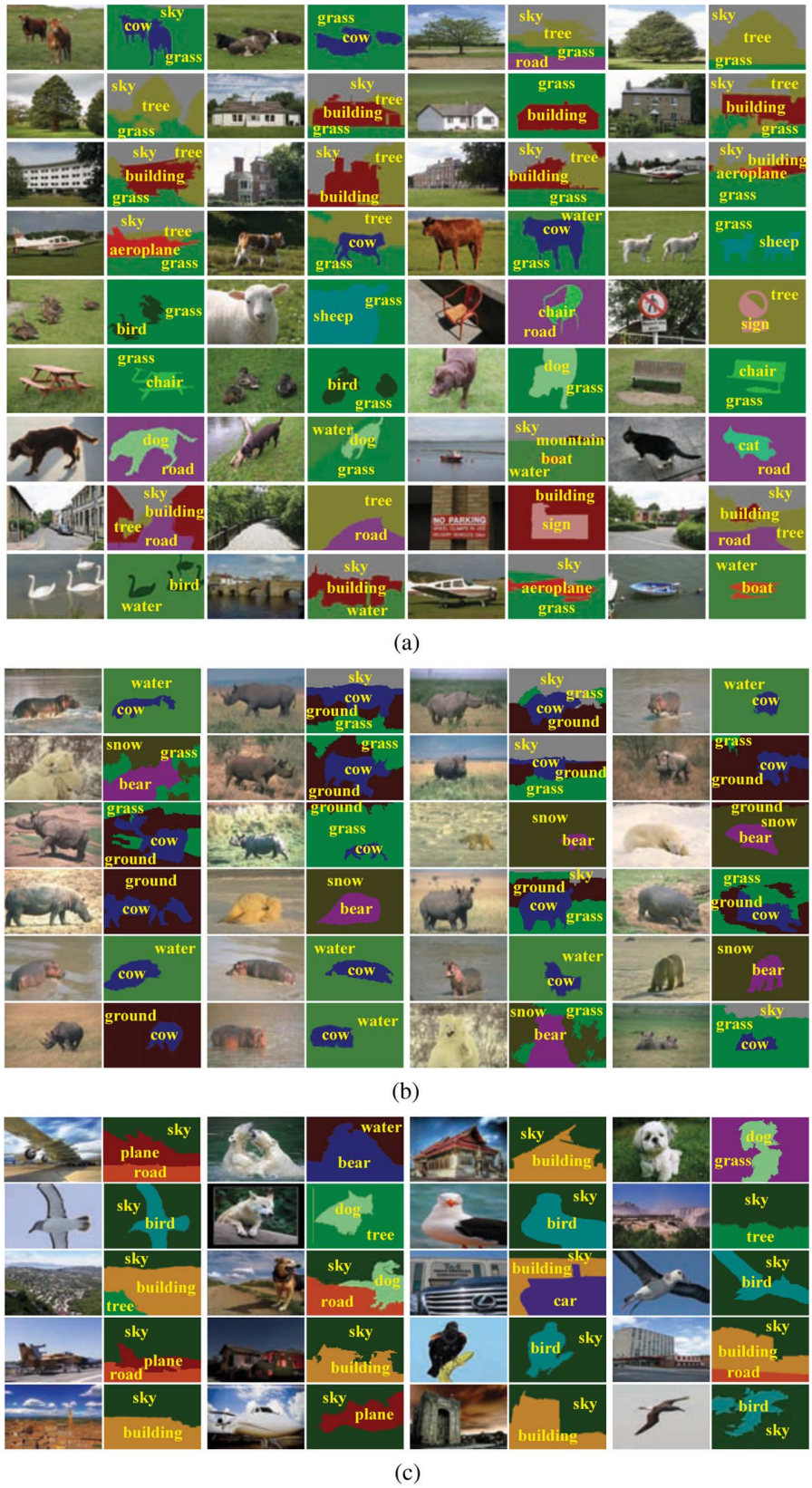


Fig. 3. Example final image parsing results. For clarity, textual labels have been superimposed on the resulting regions. The original input images are shown in columns 1, 3, 5, and 7 and the corresponding image parsing results are shown in columns 2, 4, 6, and 8. Each color in the labeled images denotes one label of localized region (by merging patches with common label). Results from (a) MSRC, (b) COREL-100, and (c) NUS-WIDE.

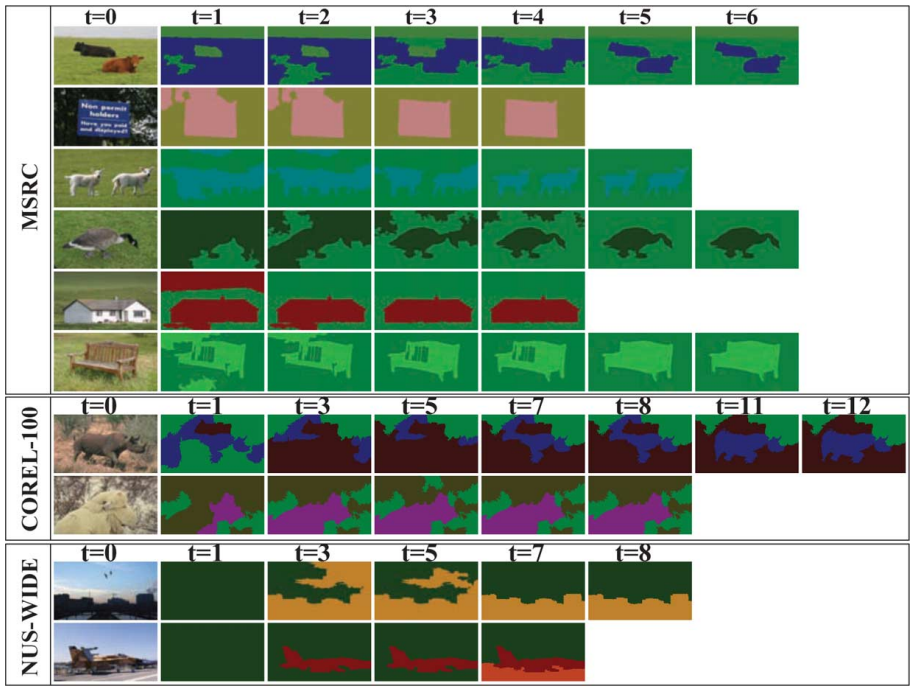


Fig. 4. Illustration of example intermediate image parsing results at different iterations for three datasets. The t indicates the iteration number.

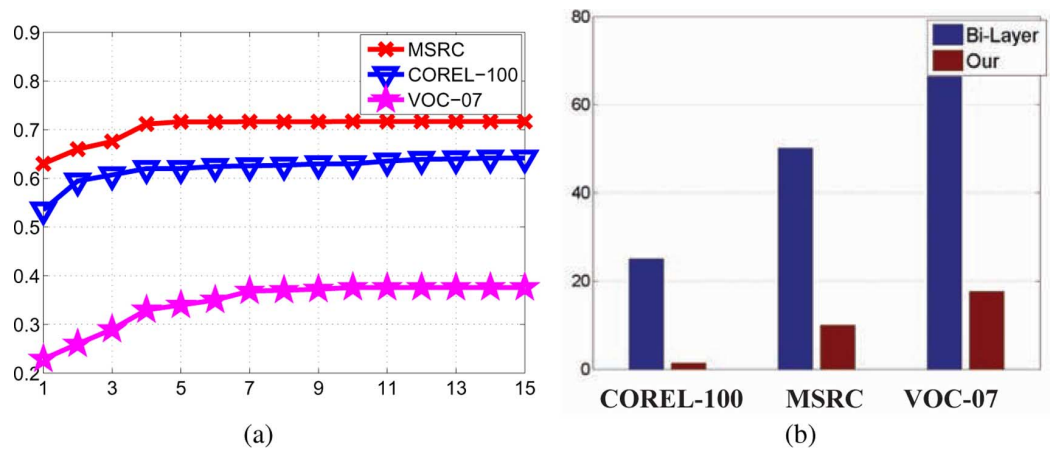


Fig. 5. (a) The collective image parsing accuracies on COREL-100, MSRC and VOC-07 datasets with different iteration numbers for the proposed algorithm. (b) Computation time comparison (in minutes) between the bi-layer method [1] and the proposed algorithm.

1) Annotation Algorithm: In order to fit the particular requirement of image annotation, we adapt our image parsing framework (10) to the following form:

$$\begin{aligned}\min_{P} \quad & \lambda_1 Tr\left(\tilde{P}^T\tilde{L}\tilde{P}\right)+\lambda_1 Tr\left(B\tilde{P}\tilde{P}^T\right)\\ & +\sum_m\sum_c\left(1-Y_{mc}\right)h_c P^T q_m\\ & +\sum_m\sum_c Y_{mc}\left(1-\max_{x_{mj}\in X_m} g_{mj}Ph_c^T\right)\\ \text{s.t.} \quad & \tilde{P}\geqslant 0, \tilde{P}e_1=e_2\end{aligned} \tag{17}$$

where $\tilde{L}_{(n_L+n_U)\times(n_L+n_U)}$ is Laplace matrix between patches from both labeled and unlabeled images, n_L and n_U are the numbers of labeled and unlabeled images, respectively, $\tilde{P}_{(n_L+n_U)\times C}$ is the label matrix of all patches and $P_{n_L\times C}$ has the same meaning as in (10), which is an upper left part submatrix of $\tilde{P}$. In (17), training images' labels are imposed

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT AUTOMATIC IMAGE ANNOTATION ALGORITHMS ON DIFFERENT DATASETS

Dataset	Method	Precision	Recall
MSRC	ML-LGC	0.62	0.55
	SMSE	0.67	0.62
	MISSL	0.63	0.60
	WSG	**0.72**	**0.63**
NUS-WIDE-SUB	ML-LGC	0.28	0.29
	SMSE	0.32	0.32
	MISSL	0.27	0.33
	WSG	**0.40**	**0.39**

Fig. 6. Semantic retrieval results. The first column is the user-provided concept map, and the top ten retrieved results of each query are shown from the second to the eleventh column. Incorrect results are shown in red bounding boxes.

on their descendent patches by the third and forth terms, while the patch level contextual relationships are defined upon all patches from both labeled and unlabeled patches, indicated by the first two terms. The optimization problem can be solved in the same way as (10).

2) Image Datasets: We evaluate and compare among four multilabel automatic tagging algorithms on the MSRC, NUS-WIDE-SUB datasets. Since COREL-100 dataset is smaller and VOC-07 datasets is a typical image segmentation datset, so we do not validate the image annotation task in these datasets. For the MSRC dataset, there are only three images annotated with label "$horse$," so we remove this label and evaluate different algorithms on the remaining 22 labels. For the NUS-WIDE-SUB dataset, we use all 81 semantic concepts.

3) Experiment Setup: We compare the proposed WSG-based algorithm against the following established semi-supervised multilabel learning algorithms for automatic image tagging.

1) The Multi-Label Local and Global Consistency algorithm (ML-LGC) proposed by Zhou *et al.* [38]. It aims to design a classifying function which is sufficiently smooth with respect to the intrinsic structures collectively revealed by both labeled and unlabeled points.
2) The Semi-supervised Multi-label learning algorithm by solving a Sylvester Equation (SMSE) proposed by Chen *et al.* [39]. Two graphs are first constructed on instance level and category level. Then a regularization framework combining two regular terms for the two graphs is used.
3) The Multiple-Instance Semi-Supervised Learning algorithm (MISSL) proposed by Rahmani *et al.* [40]. It transforms multiple instance problem into an input for a graph-based single instance semi-supervised learning method.

In all baseline implementations, the suggested parameter setting strategies in the original work are used. Training data and test data is obtained by randomly and evenly splitting the dataset. To evaluate the performance of different methods, we calculate precision and recall on each label and average them to measure the final performance.

4) Results and Analysis: Table II shows the performance comparison of the above four algorithms over the two datasets. From the results, we have the following observations.

1) The proposed method outperforms ML-LGC and SMSE algorithms on the dataset. It indicates that patches based graph can capture richer and more detailed information than image based graph, and is more suitable for multi-label image annotation task. This is quite obvious. Compared with BoWs holistic image representation, which fuses all local information indiscriminately into one histogram, semi-local patches are much more semantically independent and unique, being a better agent for label propagation.
2) The proposed algorithm outperforms the MISSL which is also based on the semantic regions, as our algorithm directly infers a C-dim label vector together instead of decomposing it into multiple binary classification problem, which preserves the correlations among labels to some extent.

B. Extension II: Image Retrieval by Concept Map

Concept map-based image retrieval [41] is a new kind of image retrieval mode. Besides multiple query words, users can also provide a blank canvas to indicate the desired spatial positions of the query words, as shown in the first column of Fig. 6. In this setting, each concept is represented by a rectangular, whose size indicates the influence scope of the keyword. A concept map captures the users' intension more accurately. Image parsing provides a high-level comprehensive understanding of an image by telling both the relative positions and scales of different objects, which matches the needs of concept map based image retrieval exactly. When users specify a particular concept map, it is matched with all images parsing results in the training set. The images with highest matching score are returned.

1) Retrieval Algorithm: We denote a concept map as $\{(k_c, b_c)\}_{c=1,\ldots,C}$, where k_c and b_c is the cth keyword and the associated rectangle by users, and C is the number of concepts in the concept map. The desired spatial distribution of the cth key word is estimated by a Gaussian distribution $G_c(x, y)$. Note that $G_c(x, y)$ of different c will overlap. To cut off the long tail, an overall distribution D_c of keyword c is defined by

$$D_c(x,y) = \begin{cases} G_c(x,y), & \text{if } G_c(x,y) = \max_{j=1,\ldots,C} G_j(x,y) \\ 0, & \text{otherwise} \end{cases}$$

where $G_c(x, y)$ is a $2D$ Gaussian distribution, with the mean $u_c = [x_c; y_c]^T$ and the covariance matrix $\Sigma_c = Diag((\theta w_c)^2, (\theta h_c)^2)$. Here, h_c, w_c and (x_c, y_c) are height, width and center coordinates of rectangle b_c respectively. To make the distribution degrade to a half near the boundary of the rectangle, θ is set to a constant $\sqrt{(2\log(2))^{-1}}$.

2) Image Dataset: As an illustration example, only the MSRC dataset is used in this task.

3) Results and Analysis: To evaluate the proposed system, we design eight different concept maps as shown in the first column of Fig. 6. Among all of the concept maps, two maps involve two concepts, five maps involve three concepts and one map involves four concepts. From the result, we can see most retrieved images are highly correlated with user provided concept maps. The incorrectly returned images (shown in red bounding boxes) can trace back to their corresponding inaccurate image parsing results.

VII. Conclusion and Future Work

In this paper, we tackle the collective image parsing problem under the weakly supervised setting with a novel WSG-based label propagation method. Different from traditional image parsing methods, the proposed approach only requires image-level label annotations as input. WSG can absorb weak label information from images and propagate them among patches simultaneously. Based on the image parsing framework, image annotation and concept map based image retrieval can be effectively implemented. In our future work, we plan to further exploit the label–level spatial contextual relationship for further boosting collective image parsing accuracy. For example, image parsing may gain more from contexts that the sky often appears above water in images.

References

[1] X. Liu, B. Cheng, S. Yan, J. Tang, T.-S. Chua, and H. Jin, "Label to region by bi-layer sparsity priors," in *Proc. ACM Multimedia*, 2009, pp. 115–124.

[2] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 115–124.

[3] Y. Chen, "Unsupervised learning of probabilistic object models (poms) for object classification, segmentation and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[4] B. Russell, A. Efros, J. Sivic, W. Freeman, and A. Zisserman, "Segmenting scenes by matching image composites," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2009, pp. 1580–1588.

[5] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *Int. J. Comput. Vis.*, vol. 81, 2009.

[6] J. Yuan, J. Li, and B. Zhang, "Exploiting spatial context constraints for automatic image region annotation," in *Proc. ACM Multimedia*, 2007, pp. 595–604.

[7] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1–8.

[8] [Online]. Available: http://www.flickr.com/

[9] L.-J. Li, R. Socher, and L. Fei-Fei, "Towards total scene understanding:classification, annotation and segmentation in an automatic framework," in *Proc. IEEE Conf. Comput. Vis. Recognit.*, 2009, pp. 2036–2043.

[10] A. Yuille and A. A. Rangarajan, "The concave-convex procedure," *Neural Computation*, vol. 15, no. 4, pp. 915–936, 2003.

[11] J. Winn and N. Jojic, "Locus: Learning object classes with unsupervised segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005.

[12] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *Proc. IEEE Eur. Conf. Comput. Vis. Workshop Stat. Learning Comput. Vis.*, 2004, pp. 17–32.

[13] L. Cao and F. Li, "Spatially coherent latent topic model for concurrent object segmentation and classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[14] E. Borenstein and S. Ullman, "Learning to segment," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2004, pp. 315–328.

[15] C. Galleguillos, B. Babenko, A. Rabinovich, and S. Belongie, "Weakly supervised object localization with stable segmentations," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2008, pp. 193–207.

[16] J. Yuan, J. Li, and B. Zhang, "Scene understanding with discriminative structured prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[17] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller, "Multi-class segmentation with relative location prior," *Int. J. Comput. Vis.*, vol. 80, no. 3, pp. 300–316, 2008.

[18] Z. Tu, X. Chen, A. L. Yuille, and S.-C. Zhu, "Image prsing: Unifying segmentation, detection, and recognition," *Int. J. Cmput. Vs.*, vol. 63, no. 2, pp. 113–140, 2005.

[19] J. He, M. Li, H. Zhang, H. Tong, and C. Zhang, "Manifold-ranking based image retrieval," in *Proc. ACM Multimedia*, 2004, pp. 9–16.

[20] J. Pan, H. Yang, C. Faloutsos, and P. Duygulu, "Gcap: Graph-based automatic image captioning," in *Proc. 4th Int. Workshop Multimedia Data Document Eng.*, 2004, p. 146.

[21] L. Mukherjee, V. Singh, and C. R. Dyer, "Half-integrality based algorithms for cosegmentation of images," in *Proc. CVPR*, 2009, pp. 2028–2035.

[22] D. S. Hochbaum and V. Singh, "An efficient algorithm for co-segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 269–276.

[23] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for cmage co-segmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1943–1950.

[24] W.-S. Chu, C.-P. Chen, and C.-S. Chen, "Momi-cosegmentation: Simultaneous segmentation of multiple objects among multiple images," in *Proc. Asian Conf. Comput. Vis.*, 2010, pp. 355–368.

[25] L. Mukherjee, V. Singh, and J. Peng, "Scale invariant cosegmentation for image groups," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 2028–2035.

[26] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[27] H. Cheng, Z. Liu, and J. Yang, "Sparsity induced similarity measure for label propagation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 317–324.

[28] S. Yan and H. Wang, "Semi-supervised learning by sparse representation," in *Proc. SIAM Int. Conf. Data Mining*, 2009, pp. 858–866.

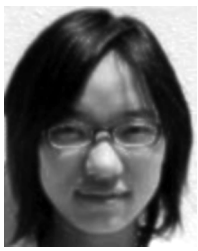
[29] B. Cheng, J. Yang, S. Yan, and T. Huang, "Learning with ℓ_1 graph for image analysis," *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 858–866, Apr. 2010.

[30] [Online]. Available: http://www.acm.caltech.edu/l1magic/

[31] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[32] X. Liu, S. Yan, J. Yan, and H. Jin, "Unified solution to nonnegative data factorization problems," in *Proc. IEEE Conf. Data Mining*, 2009, pp. 307–316.

[33] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: A real-world web image database from national university of singapore," in *Proc. CIVR*, 2009, pp. 48–55.
[34] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge 2007 (voc2007) Results." [Online]. Available: http://www.pascalnetwork.org/challenges/VOC/voc2007/workshop/index.html
[35] D. Liu, S. Yan, Y. Rui, and H. Zhang, "Unified tag analysis with multi-edge graph," in *Proc. Int. Conf. Multimedia*, 2010, pp. 25–34.
[36] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *Proc. Int. Conf. Multimedia*, 2010, pp. 461–470.
[37] D. Liu, X. Hua, L. Yang, M. Wang, and H. Zhang, "Tag ranking," in *Proc. Int. Conf. World Wide Web*, 2009, pp. 747–762.
[38] T. L. J. W. D. Zhou, O. Bousquet, and B. Scholkopf, "Learning with local and global consistency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 321–328.
[39] G. Chen, Y. Song, F. Wang, and C. Zhang, "Semi-supervised multi-label learning by solving a sylvester equation," in *Proc. SIAM Int. Conf. Data Mining*, 2008, pp. 410–419.
[40] R. Rahmani and S. Goldman, "Missl: Multiple-instance semi-supervised learning," in *Proc. Int. Conf. Mach. Learning*, 2006, pp. 705–712.
[41] H. Xu, J. Wang, X. Hua, and S. Li, "Image search by concept map," in *Proc. SIGIR*, 2010, p. .



Si Liu received the bachelor's degree from Beijing Institute of Technology, Beijing, China. She is currently working toward the Ph.D. degree at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

She is currently a Research Assistant with the Learning and Vision Group, National University of Singapore, Singapore. Her research interests includes computer vision and multimedia.

Shuicheng Yan (SM'11) received the Ph.D. degree from Peking University, Beijing, China.

He is currently an Assistant Professor in the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, and the founding lead of the Learning and Vision Research Group. His research areas include computer vision, multimedia and machine learning, and he has authored or coauthored over 200 technical papers over a wide range of research topics.

Prof. Yan is an associate editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and has been serving as the guest editor of special issues for the IEEE TRANSACTIONS ON MULTIMEDIA and *Computer Vision and Image Understanding*. He received the Best Paper Awards from ACM MM'10, ICME'10, and ICIMCS'09, the winner prize of the classification task in PASCAL VOC'10, the honorable mention prize of the detection task in PASCAL VOC'10, 2010 TCSVT Best Associate Editor (BAE) Award, 2010 Young Faculty Research Award, 2011 Young Scientist Award Singapore, and the coauthor of the Best Student Paper Awards of PREMIA'09 and PREMIA'11.

Tianzhu Zhang received the B.S. degree in communications and information technology from the Beijing Institute of Technology, Beijing, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2011.

He is a Postdoctoral Fellow with the Advanced Digital Sciences Center (ADSC), Singapore. He does extensive research on computer vision and multimedia, such as action recognition, video surveillance, and object tracking.

Changsheng Xu (M'97–SM'99) received the Ph.D. degree from Tsinghua University, Beijing, China.

He is a Professor with the Institute of Automation, Chinese Academy of Sciences, Beijing, China, and Executive Director of the China-Singapore Institute of Digital Media. His research interests include multimedia content analysis, indexing and retrieval, computer vision and pattern recognition. He holds 30 granted/pending patents and has authored or coauthored over 200 refereed research papers in those areas.

Dr. Xu a member of the Association for Computing Machinery. He is an associate editor of *Multimedia Systems Journal* and received the Best Editorial Member Award in 2008. He served as Program Chair of ACM Multimedia 2009. He has served as associate editor, guest editor, general chair, program chair, area/track chair, special session organizer, session chair and TPC member for over 20 IEEE and ACM prestigious multimedia journals, conferences and workshops. He is a founding member and Director of Programs of ACM SIG Multimedia Beijing Chapter.

Jing Liu received the B.E. and M.E. degrees from Shandong University, Shandong, China, in 2001 and 2004, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2008.

She is an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. Her current research interests include machine learning, image content analysis and classification, and multimedia.

Hanqing Lu (M'05–SM'06) received the Ph.D. degree from Huazhong University of Sciences and Technology, Wuhan, China, in 1992.

He is currently a Professor with the Institute of Automation, Chinese Academy of Sciences. His research interests include image similarity measures, video analysis, and object recognition and tracking. He published more than 100 papers in those areas.