Label to Region by Bi-Layer Sparsity Priors

Abdul Liu†‡
ECE, NUS, Singapore
elelxb@nus.edu.sg
Jinhui Tang
SOC, NUS, Singapore	
tangjh@comp.nus.edu.sg
Bin Cheng
ECE, NUS, Singapore
g0800415@nus.edu.sg
Tat-Seng Chua
SOC, NUS, Singapore
chuats@comp.nus.edu.sg
Shuicheng Yan
ECE, NUS, Singapore
eleyans@nus.edu.sg
Hai Jin
SCTS, HUST, China
hjin@hust.edu.cn

Abstract
In this work, we investigate how to automatically reassign the manually annotated labels at the image-level to those contextually derived semantic regions. First, we 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. We then harness it for the automatic label to region assignment of the entire image set. The solution to bi-layer sparse coding is achieved by convex \( \ell^1 \)-norm minimization. The underlying philosophy of bi-layer sparse coding is that an image or semantic region can be sparsely reconstructed via the atomic image patches belonging to the images with common labels, while the robustness in label propagation requires that these selected atomic patches come from very few images. Each layer of sparse coding produces the image label assignment to those selected atomic patches and merged candidate regions based on the shared image labels. The results from all bi-layer sparse codings over all candidate regions are then fused to obtain the entire label to region assignments. Besides, the presenting bi-layer sparse coding framework can be naturally applied to perform image annotation on new test images. Extensive experiments on three public image datasets clearly demonstrate the effectiveness of our proposed framework in both label to region assignment and image annotation tasks.

Categories and Subject Descriptors
I.4.9 [Computing Methodologies]: Image Processing and Computer Vision—Applications

General Terms
Algorithm, Performance, Experimentations

Keywords
Label-to-Region Assignment, Bi-layer Sparse Coding, Image Annotation, Image Parsing.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Abstract
We investigate how to automatically reassign the manually annotated labels at the image-level to those contextually derived semantic regions. First, we 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. We then harness it for the automatic label to region assignment of the entire image set. The solution to bi-layer sparse coding is achieved by convex \( \ell^1 \)-norm minimization. The underlying philosophy of bi-layer sparse coding is that an image or semantic region can be sparsely reconstructed via the atomic image patches belonging to the images with common labels, while the robustness in label propagation requires that these selected atomic patches come from very few images. Each layer of sparse coding produces the image label assignment to those selected atomic patches and merged candidate regions based on the shared image labels. The results from all bi-layer sparse codings over all candidate regions are then fused to obtain the entire label to region assignments. Besides, the presenting bi-layer sparse coding framework can be naturally applied to perform image annotation on new test images. Extensive experiments on three public image datasets clearly demonstrate the effectiveness of our proposed framework in both label to region assignment and image annotation tasks.

1. Introduction

Keywords based queries have been found to be the most efficient and effective for Internet image search. Beyond simply harnessing the indirect surrounding texts of web images for query matching, the more desirable technique is to annotate the images with their associated semantic concepts/labels. To achieve reliable and visible content-based image retrieval, it is critical to obtain the correspondence between the image labels and their precise regions within an image. In practice, it is very tedious to manually annotate the image labels to the corresponding image regions, and a more feasible alternative is to annotate the labels at the image-level. Therefore, it is interesting and practically valuable to investigate how to automatically reassign the labels annotated at the image-level to those contextually derived image regions, i.e., the label to region assignment (LRA) problem.

Although the LRA problem has not been essentially studied before, there are some related works in computer vision community, known as simultaneous object recognition and image segmentation. These algorithms can be roughly divided into two categories. The first category focuses on unsupervised learning techniques [11, 13, 2, 27]. Leibe et al. [2] propose to perform object localization, namely image segmentation along with object classification, by using an implicit shape model, which was further extended by Chen et al. in [27] to learn explicit shape model from single image. Both of them focused on single object category or assumed there is no overlapping between multiple objects in the training images. Winn et al. in [11] propose to learn object classes based on the results of automatic image segmentation. A recent extension is presented
by Cao et al. in [13], which applied the spatially coherent latent
topic model to conduct multi-label image segmentation and classi-
fication. These algorithms, however, can only handle images either
with single major object or with clean background and without oc-
cclusions between objects. In contrast, the research in this paper
aims to process more challenging images containing multiple ob-
jects and with possible inter-object occlusions.

The second category is generally founded on supervised learning
techniques. The typical efforts are the classifier-based methods [15,
17, 3, 21], which usually first learn image classifiers to character-
ize concepts (or keywords) based on the training images, and then
identify the images belonging to the specific category. These algo-
rithms are very limited when encountering cases with semantically
overlapped labels or imbalanced data from different semantic la-
bles, which will heavily impair the discriminative power of these
algorithms. There are also approaches which focus on learning the
relation between the visual features and semantic concepts, in-
cluding CMRM [8] and its extended versions [25, 23, 9]. In addi-
tion, some works [6, 16, 22] are proposed to additionally harness
the label correlation for label ranking and choosing the proper key-
words as semantic annotations, and most of them use the image-
to-image visual similarities to predict the image labels. However,
there are usually multiple semantic concepts within one image and
two different images containing a common object may contain dif-
f erent other objects at the same time. For example, in Figure 1,
the image with objects "cow", "sky" and "mountain" may be vi-
sually different from the images with only "sky" or " mountain".
Therefore, it is not reliable to directly compare the features of two
images that may contain different number of objects from different
categories.

Compared with the above efforts for simultaneous image anno-
ation and parsing, LRA instead elicits a more challenging prob-
lem, characterized by: 1) the optimal partition of the input images
to semantic regions and the correspondence between the annotated
labels and image regions are unknown, which makes most state-
of-the-art classifier-based methods [15, 17, 3] inapplicable; and 2)
all the spatially connected objects within an image need be assigned
with individual labels, which may challenge those conventional un-
supervised learning algorithms as aforementioned. Figure 1 illus-
trates the problem inputs, i.e., images annotated with labels at the
image-level, and the problem outputs, i.e., semantic regions with
labels, for the label-to-region assignment task.

To address the LRA problem, we propose to propagate the la-
belso annotated at the image-level to those local semantic regions
merged from the over-segmented atomic image patches of the en-
tire image set. Generally, one label of an image only characterizes
a single local semantic region, and two images with common la-
ables often share similar semantic regions. Inversely, if two local
semantic regions from different images are visually similar, these
two images are likely to share certain common label. Thus, if the
region-to-region correspondences have been given for all the im-
age pairs, we can assign the common image labels to those cor-
sponded local regions, which then translates the label-to-region
assignment problem into a problem to uncover the region-level cor-
respondence. In practice, these semantic regions corresponding to
certain labels cannot be directly obtained, but those smaller-size
spatially coherent image patches are easy to derive with classical
image over-segmentation approach. One semantic region generally
comprises of multiple such atomic patches, but it is infeasible to un-
cover the patch-to-region relations by merging those visually simi-
lar patches, due to the underlying large within-region variations.
In this work, we instead propose to first construct the so-called can-
dicate regions, initially grouping from those local spatially coherent

Figure 2: Sketch of our proposed solution to automatic label-
to-region assignment task. This solution contains four steps: 1) patch
extraction with image over-segmentation algorithm; 2) image recon-
struction via bi-layer sparse coding, 3) label prop-
agation between candidate region and selected image patches
based on the coefficients from bi-layer sparse coding, and 4)
post-processing for deriving both semantic regions and associ-
ated labels.

patches, and then use those atomic patches from other input images
to reconstruct the candidate regions, with the hypothesis that those
selected atomic patches for reconstruction shall come from few se-
mantically similar regions. Finally the cross-image patch-to-region
 correspondences are used for the ultimate label-to-region assign-
ment purpose. Note that: (1) we cannot directly use visual similarity
between the candidate region and atomic patch to select patches
for the reconstruction purpose, since an atomic patch is only part of
a region and their similarity cannot convey the inclusion relations;
and (2) an intuitive way to improve the accuracy of cross-image
region-to-region correspondence is to enforce the usage of atomic
patches from few images for this reconstruction, from which those
selected atomic patches from one image may have higher possibility
to form a semantic region.

More specifically, the reconstruction of a semantic region from
a set of images is achieved by the proposed bi-layer sparse
coding formulation. The basic philosophy is that an image or se-
manic region can be sparsely reconstructed via the image patches
belonging to the images with common image labels. We addi-
tionally introduce another type of constraints, namely, to select patches
from as few images as possible, which brings the second layer of sparsity to improve the fidelity in label-to-region assignment.
Based on the sparse reconstruction coefficients, we assign the
common image labels to the selected patches, and then further fuse all
the assignment results to distribute the image labels to those con-
textually derived semantic regions merged from multiple atomic
patches. The proposed label-to-region assignment process has the
following characteristics: 1) the bi-layer sparse coding aims to en-
force the usage of merged patches within an image to reconstruct
the reference image or semantic region, which ensures the reliabil-
ity of label propagation; 2) the process does not require exact image
object/concept parsing, which is still far from satisfactory on real
world images; and 3) no generative model for each label/concept is
learnt, and thus it is scalable to applications with large label set. In
addition, the proposed bi-layer sparse coding formulation can also
be directly applied on new test images to perform multi-label image
annotation. Figure 2 illustrates the overall sketch of this idea.

The remainder of this paper is organized as follows. We first
formulate the label-to-region task within the bi-layer sparse cod-
ing framework in Section 2 and introduces how to use the bi-layer
sparse coding for direct image annotation in Section 3. The de-
tailed comparison experiments are then demonstrated in Section 4.
Section 5 presents the conclusive remarks along with discussion for
future work.
2. LABEL TO REGION ASSIGNMENT BY BI-LAYER SPARSITY PRIORS

2.1 Overview of Problem and Solution

The ability to annotate images with related text labels at the semantic region-level is invaluable for boosting keyword based image search with the awareness of semantic image content. However, it is tedious if not impossible to manually annotate labels at the region-level for large-scale image set. We therefore study in this work on how to utilize the cross-image label contexts to automatically reassign the image labels to those contextually merged image patches in a group manner. As illustrated in Figure 4, image \( y \) comprises an ensemble of image patches, each of which may partially characterize one image label, e.g., tree, building, etc. Two images annotated with common labels are likely to contain some similar patches. However it is generally difficult to directly derive those semantically similar patch pairs between two images. Thus instead we use a group of atomic patches to reconstruct an image or semantic region, and then harness the reconstruction coefficients for propagating the image labels to those localized image patches. Meanwhile, to reduce the influence of image noises and robustly derive label-to-region relations, we propose to enforce that the selected atomic patches should come from as few images as possible for the reconstruction purpose. Consequently, we obtain a bi-layer sparse coding framework, where each image is reconstructed using a few localized atomic patches from a few related images. Note that in this work, we only consider such localized labels, i.e., the so-called flat labels.

2.2 Over-Segmentation and Representation

As aforementioned, the main purpose of this work is to propagate the semantic labels annotated at the image-level to image regions merged from image patches. Each homogeneous patch comprises of the pixels that are spatially coherent and perceptually similar with respect to certain appearance features, such as intensity, color and texture, etc.

Our proposed solution starts with an initial image over-segmented by a reliable segmentation algorithm into multiple homogeneous atomic patches. Here we choose to use the graph-based segmentation algorithm in [19], which incrementally merges smaller-size patches with similar appearances and with small minimum spanning tree weights. This method is of nearly linear computational complexity in the number of neighboring pixels. In this work, we use a modified version of [19] to obtain coherent patches which contain only one single image label each. First, we resize all the images into the resolution of \( 320 \times 240 \) pixels and initialize each pixel as one atomic patch. Then, we use the color features to describe the appearance of an initial image patch and apply the algorithm [19] for merging the smaller patches into the larger ones. Thereinto, we propose to stop the merge if the patch size is larger than a predefined threshold, denoted as \( M_1 \). Here, we set \( M_1 = 300 \). Thus, for each input image, we generally obtain about \( 40 \sim 50 \) atomic patches. Based on Intel Xeon X5450 workstation with 3.0GHz CPU and 16GB memory, it takes less than 0.2 second to segment one image. Figure 3 shows an exemplary result of an over-segmented image.

The goal of the image over-segmentation step is to ensure that the segmented patch is involved within an object/concept, and these over-segmented patches shall be merged to constitute semantic regions. Note that our proposed solution is general and not tied to any specific image segmentation algorithms. This way of using the image patches makes our algorithm less vulnerable to the quality of the image segmentation step.

Based on the image over-segmentation results, we can obtain the feature representations for those atomic patches. Let \( X = \{ x_1, z_1, \cdots, x_i, z_i, \cdots, z_N \} \) denotes the annotated image set, where \( N \) is the total image number, \( z_i \in \mathbb{R}^{N_C} \) indicates the label vector. The binary element \( z_i(c) \) takes 1 if the \( i \)th image contains the \( c \)th label and 0 otherwise. \( N_C \) is the total number of image labels. As aforementioned, each image \( x_i \) contains an ensemble of atomic patches, denoted as \( X_i = [ x_{i,1}, x_{i,2}, \cdots, x_{i,n_i} ] \), where \( x_{i,n_i} \in \mathbb{R}^m \) is an \( m \)-dimensional feature descriptor and \( n_i \) is the number of patches belonging to the \( i \)th image. We then arrange all the patch representations as column vectors of the matrix \( A = [ X_1, X_2, \cdots, X_N ] \in \mathbb{R}^{m \times \sum_{i=1}^N n_i} \). Thus, for any specific images \( Y = [ y_1, \cdots, y_{m_y} ] \in \mathbb{R}^{m \times m_y} \), the target of sparse image coding is to represent the full or partial sum of the column vectors as the linear combination of the column vectors in matrix \( A \). In other words, an image or its initially merged candidate region is reconstructed from a set of over-segmented image patches.

We describe each atomic patch by using Bag-of-Words (BOW) features. The generation of visual words comprises of three steps: i) we apply the Difference-of-Gaussian filter on the gray-scale image to detect a set of salient points; ii) we then compute the Scale-Invariant-Feature-Transform (SIFT) [4] features over the local areas defined by the detected salient points; and iii) we perform the vector quantization on SIFT region descriptors to construct the visual vocabulary by K-Means clustering approach. In this work we generate 5000 clusters, and thus the dimension of the BOW feature vector is \( m = 500 \).

2.3 1: Sparse Coding for Candidate Region

The core component of the solution to label-to-region assignment task is to find out the semantically-similar region-pair from two images that contains common labels/concepts. We can then derive the label information for the region-pair from the shared images labels. A dilemma is that the over-segmentation step only produces smaller-size patches instead of semantic regions. In this work, we propose a sparse coding framework to implicitly uncover the semantic region correspondence by explicitly uncovering how an image or its candidate region can be reconstructed from the over-segmented patches of other input images. Mathematically, denote \( y \) as the feature representation of an image or its candidate region merged from over-segmented patches. If sufficient training samples are available for each label, it is possible to represent \( y \) as a sparse and linear combination of the patch representations from other input images, namely,

\[
y = A \alpha_0 \in \mathbb{R}^m,
\]

where \( \alpha_0 \) is the coefficient vector whose entries are expected to be zeros except for those samples associated with common label(s) with \( y \). For ease of representation, we use \( A \) again here for all the patch representations from all other input images.

Theoretically, \( \alpha_0 \) can be obtained by solving the linear system of equation \( y = A \alpha_0 \), but when \( m < \sum_{i=1}^N n_i \), there exist infinite
number of possible solutions. A possible way to select a solution is to minimize the $\ell^2$-norm of the solution, namely,

$$\hat{\alpha}_2 = \arg \min_{\alpha} ||\alpha||_2, \quad s.t. \quad A \alpha = y.$$  

(2)

The solution $\hat{\alpha}_2$, although is easy to obtain, is dense and thus not informative for reconstructing $y$. Essentially, the sparser the recovered $\hat{\alpha}_0$, the easier it will be to accurately determine the correspondence of $y$ and the selected patches. Thus, it is reasonable to seek the sparsest solution to $y = A\alpha$ by solving the following optimization problem:

$$\hat{\alpha}_0 = \arg \min_{\alpha} ||\alpha||_0, \quad s.t. \quad A \alpha = y,$$

(3)

where $|| \cdot ||_0$ denotes the $\ell^0$ norm, which counts the number of nonzero elements in a vector. However, this problem is NP-hard. Fortunately, recently developed theories on sparse representation reveals that if the solution $\hat{\alpha}_0$ is sparse enough, the solution from the $\ell^2$-minimization can be recovered by the solution to the following $\ell^1$-norm minimization problem:

$$\hat{\alpha}_1 = \arg \min_{\alpha} ||\alpha||_1, \quad s.t. \quad A \alpha = y.$$  

(4)

This optimization problem is convex and can be transformed into a general linear programming problem. There exists a globally optimal solution, which can be solved efficiently using the classical $\ell^1$-norm optimization toolboxes, like [7].

Furthermore, the real world images are often noisy, and thus it may be impossible to express $y$ exactly as a sparse superposition of the column vectors of $A$. To explicitly account for those often sparse noises, we rewrite the sparse coding formulation in Eq. (1) as follows:

$$y = A\alpha + \epsilon,$$

(5)

where $\epsilon \in \mathbb{R}^m$ is a noise vector. The sparse solution can again be recovered by solving the following robust $\ell^1$-norm minimization problem:

$$[\hat{\alpha}_1, \hat{\epsilon}_1] = \arg \min_{\alpha, \epsilon} ||\alpha||_1 + ||\epsilon||_1, \quad s.t. \quad y = A\alpha + \epsilon,$$

(6)

which simultaneously imposes the sparse constraints on both reconstruction coefficients and noises. Similarly, this problem can also be solved by classical $\ell^1$-norm optimization toolboxes.

### 2.4 II: Sparsity for Patch-to-Region

The ultimate of sparse coding in this work is to build pairwise semantic region correspondence, which is then used for label propagation from image-level to region-level. Generally each semantic region comprises of several over-segmented patches, and thus it is natural to enforce the possibility of merging atomic patches into semantic regions within individual image, which motivates an extra layer of sparsity, called the sparsity for patch-to-region.

Let $\beta_{i,j}$ denote the normalized importance weight of the $j$th patch for the $i$th input image, and we bring another set of coefficients, $\gamma_1, \gamma_2, \cdots, \gamma_N$, to measure the total importance weights for individual image,

$$\gamma_i = [\gamma_1, \gamma_2, \cdots, \gamma_N]^T,$$

(7)

where $\beta_{i,j}$ is calculated according to the size of the $j$th atomic patch and normalized by the image size of the $i$th image, and the index for $\alpha$ is rearranged according to the patch index within each image. Here, we define a matrix $B \in \mathbb{R}^{N \times \sum_n n_i}$ using $\beta_{i,j}$ as:

$$B = \begin{bmatrix} \beta_{1,1} & \cdots & \beta_{1,n_1} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \beta_{N,1} & \cdots & \beta_{N,n_N} \end{bmatrix},$$

and then we can rewrite Eq. (7) as

$$\gamma = B\alpha.$$  

(8)

Finally, we obtain the following optimization problem:

$$[\hat{\alpha}_1, \hat{\epsilon}_1, \hat{\gamma}_1] = \arg \min_{\alpha, \epsilon, \gamma} ||\alpha||_1 + ||\epsilon||_1 + ||\gamma||_1, \quad s.t. \quad y = A\alpha + \epsilon, \quad \gamma = B\alpha.$$  

(9)

Let

$$y' = \begin{bmatrix} y \\ 0_{N \times n} \end{bmatrix}, \quad \alpha' = \begin{bmatrix} \alpha \\ \epsilon \\ \gamma \end{bmatrix}, \quad A' = \begin{bmatrix} A_{1 \times m \times m} & 0_{m \times N} \\ B_{N \times m \times m} & -I_{N \times N} \end{bmatrix},$$

and then we can reformulate the bi-layer sparse coding as the solution of $\ell^1$-norm optimization below:

$$\hat{\alpha}'_1 = \arg \min_{\alpha'} ||\alpha'||_1, \quad s.t. \quad y' = A'\alpha'.$$

(10)

The derived $\hat{\alpha}_1$ and $\hat{\gamma}_1$ are both sparse, and thus $y$ is reconstructed from a set of sparsely selected column vectors in $A$, which belong to few images. This result is in accordance with the real observations and the entire algorithm is called "Bi-layer Sparse Coding". Figure 4 illustrates the exemplary explanation of the bi-layer sparse coding formulation. Figure 5 shows the comparison results on the distribution of the reconstruction coefficients from bi-layer and one-layer sparse codings. We can observe that based on the bi-layer sparse coding, the selected patches tend to gather from within a few images. Figure 6 displays some examples on how a candidate region within an image is reconstructed from the over-segmented atomic patches guided by bi-layer sparsity priors.
Figure 5: Two exemplar comparison results for bi-layer sparsity (a, c) vs. one-layer sparsity (b, d). The subfigures are obtained based on 20 samples randomly selected from the MSRC dataset[12] used in the experiment part. The horizontal axis indicates the index for the atomic image patch and the vertical axis shows the values of the corresponding reconstruction coefficients (We only plot the positive ones for ease of display).

Figure 6: Exemplary results of bi-layer sparse coding for sparse image reconstruction from the MSRC database. For each row, the left subfigure shows the initially merged candidate region and its parent image, and the right subfigure shows the top five selected images and their selected patches.
choose an image to-region assignment. and the reference image or candidate regions. The fusion of the directional label propagation between the selected atomic patches results from all such reconstructions yields the procedure for label- selected atomic patch with the common labels shared by the image X segmentations patches) of the reference image x.

2.5 Contextual Label-to-Region Assignment

In this subsection, we further introduce how to utilize the bi-layer sparse coding for label-to-region assignment, namely, the simultaneous semantic region merging from atomic patches and region label assignment. The proposed procedure is motivated by the observation that, if candidate region representation y of image x is reconstructed by using the patch x_{i,j} of the image x_i with the coefficient α_{i,j}, then the patch x_{i,j} is likely to contain the content for the labels shared by the image x and x_i. Moreover, the larger the reconstruction coefficient α_{i} is, the more likely the patch x_{i} contains the shared labels. This observation naturally leads to a bi-directional label propagation between the selected atomic patches and the reference image or candidate regions. The fusion of the results from all such reconstructions yields the procedure for label-to-region assignment.

The label assignment procedure contains four iterative steps: (1) choose an image x_i and corresponding label vector z_i from the input image set, then collect and aggregate the atomic patches of the remaining images in matrix A; (2) derive the bi-layer sparse solution α of the equation y_i = Aα by ℓ²-norm minimization, where y_i is the representation for a candidate region (merged from over-segmented patches) of the reference image x_i; (3) select each atomic patch with the common labels shared by the image x_i and the image that the patch belongs to; and (4) assign the labels to the candidate region based on the selected patches and the coefficient vector α.

These four steps iterate by choosing each input image as reference image in turn, and the label vector z_{i,p} of each atomic patch x_{i,p} is obtained by cumulatively summing the label vector propagated to it in each iteration. Note that each candidate region comprises of several atomic patches, and the patch-level label vectors for the involved patches are updated in a cumulative way. In practice, after choosing an input image as reference image, we use a simple algorithm described in [19] to merge the spatially coherent and perceptually similar atomic patches to form the relatively larger-size candidate regions. We stop the merging if the region size is larger than a constant threshold, denoted as M2, which is set as 6000 pixels in this work. Figure 6 shows some generated candidate regions from MSRC dataset [10] in the first column.

Algorithm 1 details the procedure for label-to-region assignment based on the bi-layer sparse construction. The system inputs are the images with labels and the outputs are the merged semantic regions with assigned region-level labels. Here, we would like to highlight some aspects of Algorithm 1 as follows:

1. Step 2 calls the algorithm in [19] to obtain the over-segmented patches for each input image. Note that we use a label vector, instead of a single label, to describe and accumulate the label coefficients propagated from images.
2. Step 3 initializes the label vector of atomic patch using the annotated labels of its parent image, which have been manually annotated. The experiments empirically show the gain in algorithmic robustness achieved from this initialization.
3. The iterative procedure in Step 4 implements the one-vs-else label propagation scheme. Note that ∧ denotes the and operator between two vectors.

Suppose now the label vector of each atomic patch in image x_i has been derived by Algorithm 1, we adopt the K-means clustering approach over all the vectors to generate K = K_i clusters, where K_i is the number of labels annotated for image x_i. Here, each cluster corresponds with one single semantic region and thus should be assigned with a different label. Algorithm 2 summarizes the overall procedure.

3. DIRECT IMAGE ANNOTATION BY BI-LAYER SPARSE CODING

In this section, we show how the proposed bi-layer sparse coding formulation can be used for direct image annotation on new test images by propagating the labels from a set of training images with the annotated labels. For a given test image with the patch representations as Y = [y_1, · · · , y_N], we set the reference representation y = ∑_i y_i. We then determine the sparse reconstruction coefficient matrix α_1 and g_1 by solving the problem in Eq. (10).
The label vector of the test image can then be obtained as:

\[ z_y = Z \gamma_1, \]

where \( Z \in \mathbb{R}^{N_x \times \sum_{i=1}^{N_y} n_i} \) is the label matrix for all the training images. The labels corresponding to the top few largest values in \( z_y \) are considered as the final annotations of the test image. Note that, as the annotation task requires only unidirectional label propagation, namely assigning the labels of the training images to the testing image, we extract the reference representation \( y \) from the whole testing image, instead of candidate regions, to save the computational cost and to reduce the possible errors caused by the over-segmentation step.

Compared with classical works for image annotation, the proposed bi-layer sparse coding based image annotation algorithm has the following characteristics: 1) the propagation process is robust and less sensitive to the image noises owing to the bi-layer sparse coding formulation; and 2) the proposed algorithm is scalable to large-scale, even web-scale, image retrieval by first selecting a set of visually related images and then performing bi-layer sparse coding over those roughly selected images.

4. EXPERIMENTS

In this section, we systematically evaluate the effectiveness of our proposed bi-layer sparse coding formulation for both label-to-region assignment and image annotation tasks.

4.1 Data Sets


The third dataset, COREL Collection, is the most broadly adopted dataset in the community of image retrieval, which however has a lack of diversity among the images that stored in the same group. Therefore, we select a moderate scale images subset to avoid this problem. Following the collection strategy in [12], we eventually obtain an image set that contains a total of 4000 images from 11 categories. Each image is annotated with about 3.5 labels on average. Besides, as the original COREL collection only provides image-level labels, we further randomly select 100 images and manually annotate the region-level groundtruth for evaluation. This subset, named COREL-100, contains the images from 7 categories of: ‘grass’, ‘cow’, ‘snow’, ‘sky’, ‘bear’, ‘ground’ and ‘water’. These three datasets provide the image-level annotations and hence can all be used for the experiments on image annotation task. MSRC and COREL-100 additionally provide the region-level annotations, and thus they can both be used for the exams on label-to-region assignment.

All the experiments are performed on an Intel Xeon X5450 workstation with 3.00 GHz CPU and 16 GB memory. The code is implemented on MATLAB platform. Algorithm 1 can process 350 images (each of which is segmented into about 40 50 atomic patches) within 50 minutes. For the image annotation task, our method can reconstruct and predict a new test image (320 × 240 pixels) within 10 seconds (using the patch set collected from 350 images).

4.2 Exp-I: Label-to-Region Assignment

4.2.1 Parameters, Benchmarks and Metrics

We implement the proposed label to region assignment algorithm using the \( \ell^1 \)-Magic package [7]. It first translates Eq. (10) into a linear programming problem and then adopts the primal-dual algorithm to perform the optimization. In the implementation, we set the tolerance factor as 0.003 and the maximum number of primal-dual iterations as 50. Another two free parameters are the maximal patch size \( M_1 \) and maximal region size \( M_2 \), both of which are used to control the segmentation algorithms [19]. The selection of these two parameters essentially makes a tradeoff between algorithmic performance and efficiency. Basically, the decrease of the patch size or region size shall increase the computational cost or the iteration number for the reconstruction step in Algorithm 1, but may potentially increase the algorithmic performance. Therefore, in experiments, we resize all the images to the resolution of 320 × 240 pixels and empirically set the two parameters as \( M_1 = 300 \) pixels and \( M_2 = 6000 \) pixels respectively.

Two algorithms are implemented as baselines for comparison to evaluate the effectiveness of the proposed bi-layer sparse coding formulation in label to region assignment task. One is the classical binary Support Vector Machine (BSVM), which translates the \( m \)-class multi-label classification problem into \( m \) binary classification problems. For each classifier, the image is considered as positive sample if it contains the specific concept/label, otherwise it is set as negative sample. In the training stage, we choose equally number of positive and negative samples and remove the overabundant ones to balance the training of SVM. In the testing stage, we first apply each classifier on the atomic patch to obtain the probability of the patch to be positive sample. The results from the \( m \) classifiers are then fused to generate the \( m \)-dimensional label confidence vectors, which are further processed by Algorithm 2 to obtain the labels of those merged regions. Note that the training and testing procedures work at two different levels of the images, and the goal is to eventually obtain the semantic annotations at the region-level. For a fair and reliable comparison, we apply this baseline on atomic patches generated by the modified segmentation algorithm [19] with different allowed maximal sizes, i.e., \( M_1 = 150, 200, 400 \) and \( 600 \) pixels. The binary SVM is implemented based on the lib-SVM library [20] and the Gaussian Radial Basis Function kernel is used by setting the kernel parameter as 1.

The second baseline algorithm is a simplification of the proposed solution, called one-layer sparse coding, which is used to demonstrate the improvement brought by the proposed bi-layer sparse coding formulation. The overall procedure is similar to Algorithm 1, except that the system of equation to perform the optimization is the Eq. (6) in Section 2.4. We set the parameters the same as that for the algorithm based on bi-layer sparse coding formulation.

The label-to-region performance is evaluated in both qualitative and quantitative ways. The quantitative label-to-region assignment
accuracy measures as the percentage of pixels with agreement between the assigned label and ground truth.

4.2.2 Results and Analysis

Table 1 shows the accuracy comparison between the baseline algorithms and our proposed algorithm on the MSRC and COREL-100 datasets. The detailed comparison results for individual classes are illustrated in Figure 7. From these results, we can have the following observations. (1) The proposed algorithm achieves much higher accuracies of 0.63 and 0.61 on the MSRC and COREL-100 dataset respectively as compared to the SVM-based baseline. This clearly demonstrates the effectiveness of the bi-layer sparse coding for relating the candidate regions with the atomic paths. 2) Bi-layer sparse coding based algorithm outperforms the one-layer sparse based algorithm over both two datasets. This is because the former enforces the usage of merged patches for the reconstruction of candidate regions. It greatly improves the quality of construction for label propagation and thus boosts the accuracy for label-to-region assignment. (3) On analysis of detailed class-level results shown in Figure 7, it is noted that our algorithm seems to be less effective for handling the categories for foreground objects, such as dogs, cows, and cats, compared to background regions such as streets, trees and sky etc. This is because the labels of object classes have lower weights (normalized region size) as compared to the background regions. Although we can learn individual classifiers for these specific objects as in [13, 17, 3] and then apply them to detect and localize objects in images, these algorithms generally require training data with ground-truths at the region-level and thus are not applicable for this general prior-free label-to-region assignment task.

Note that we do not compare our solution to label-to-region assignment task with those typical algorithms for simultaneously classifying and localizing objects in images, for three reasons: a) our proposed solution works under the assumption that no region-level label annotation is provided, which is however the general prerequisite for most typical algorithms; b) most typical algorithms are tailored to specific objects and can only work on cases with limited number of object categories; and c) for each image label, those typical algorithms need to learn an individual detector, which is thus very time consuming and more difficult to be applied for large-scale applications.

Some example results on label-to-region assignment are displayed in Figure 8 and 9 for the MSRC and COREL-100 datasets, respectively. These results over various conditions well validate the effectiveness of our proposed solution. Note that our proposed algorithm is scalable to large-scale applications, though we do not report results on larger-size dataset (mainly due to the tedious annotation for providing ground-truths, which is also the main motivation of this work) here. The algorithm is essentially fast if we utilize the priors to remove the input images without common labels with the reference image, and also the entire algorithm is suitable for parallel computation. Accordingly, more samples with abundant labels are able to provide more contextual information, which shall further boost the overall performance of Algorithm 1.

4.3 Exp-II: Image Annotation on Test Images

4.3.1 Benchmarks and Metrics

Three popular algorithms are implemented as benchmark baselines for the image annotation task. (1) Binary SVM [14] (BSVM), which translates the \( m \)-class multi-label classification problem into \( m \) binary classification problems. We use the same setting as that for the label-to-region assignment experiments. (2) The Correlated Label Propagation algorithm (CLP) proposed by Feng et al. in [6]. It provides several choices for the kernel type in the objective function, and in this work, we use the exponential function with parameter \( \alpha = 0.6 \) (not the \( \alpha \) in Eq. (10)), which usually achieves the best performance as reported in [6]. (3) The KNN based multi-label learning algorithm (MLKNN) proposed by Zhuang et al. [16]. We set the number of nearest neighbors to 15 and keep other parameters the same as in [16].

CLP and MLKNN 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 rank-SVM [11] and boost.MH [5]. Thus, we do not plan to further implement the latter two in this work. We evaluate and compare among the four algorithms over three datasets, MSRC, COREL-4K and NUS-WIDE, 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.
Figure 8: Example results on label-to-region assignment. The images are from the MSRC dataset. The original input images are shown in the columns 1, 3, 5, 7 and the corresponding labeled images are shown in the columns 2, 4, 6, 8. Each color in the labeled images denotes one class of localized region.

Figure 9: Example results on label-to-region assignment from the COREL dataset.
4.3.2 Results and Analysis
The comparison results on image annotation performance are reported in Table 2, where each row shows the achieved MAP (mean average precision) over different datasets. From these results, we can derive the following observations. (1) The proposed method based on bi-layer sparse formulation outperforms the three baselines over all datasets. (2) The latter three methods all use the label contextual information, and achieve much higher performance than the SVM-based algorithm. This also accords with the motivation of our proposed solution, which takes the advantage of the label contextual information for label propagation. Figure 10 illustrates some exemplary image annotation results from the NUS-WIDE dataset. The images are challenging due to the large intra-class variations and the usually inter-class occlusions.

Table 2: Image label annotation MAP (Mean Average Precision) comparisons among four algorithms on three different datasets.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MSRC</td>
<td>0.30</td>
<td>0.54</td>
<td>0.65</td>
<td>0.70</td>
</tr>
<tr>
<td>COREL-4K</td>
<td>0.41</td>
<td>0.53</td>
<td>0.61</td>
<td>0.72</td>
</tr>
<tr>
<td>NUS-WIDE</td>
<td>0.56</td>
<td>0.64</td>
<td>0.63</td>
<td>0.76</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS AND FUTURE WORK
In this paper, we proposed a novel sparse coding technique for addressing an interesting task of label-to-region assignment, which only requires image-level label annotations. With the popularity of the photo sharing websites, the community-contributed images with rich tag information are becoming much easier to obtain, it is predicted that the keyword query based semantic image search can greatly benefit from applying our proposed technique for label-to-region assignment on these tagged images.

Our proposed solution for both label-to-region assignment and image annotation tasks is applicable to handling large-scale dataset. For the label-to-region assignment task: 1) the images can first be clustered according to the label information, and the proposed solution can then be applied within each cluster; and 2) the priors can be utilized to remove the input images without common labels with the reference image for the sparse reconstruction of the reference image or its candidate regions. For the image annotation task, we can first roughly select a sufficiently large set of visually similar images of the reference image, and then apply the bi-layer sparse coding formulation on these selected images for image annotation.

We are planning to further study the label-to-region assignment problem from two aspects. (1) Currently our proposed solution can only handle flat labels, namely one label corresponds to one local image region. We shall further study how to handle semantically overlapped labels. (2) We shall study how to take the advantage of those images with partial region-level labels.

6. ACKNOWLEDGEMENT
This work is supported by NRF/IDM Program, under research Grant NRF2008IDM-0DM004-029.

7. REFERENCES