Nonparametric Label-to-Region by Search [*]

Xiaobai Liu[†,‡], Shuicheng Yan[‡], Jiebo Luo[§], Jinhui Tang[‡], Zhongyang Huang[¶] and Hai Jin[†]
[†] Huazhong University of Science and Technology, China,
[‡] National University of Singapore, Singapore
[§] Kodak Research Laboratories, Eastman Kodak Company, Rochester, USA
[¶] Panasonic Singapore Laboratories Pte Ltd, Singapore

Abstract

In this work, we investigate how to propagate annotated labels for a given single *image from the image-level to their corresponding semantic regions, namely Label-to-Region (L2R), by utilizing the auxiliary knowledge from Internet image search with the annotated image labels as queries. A nonparametric solution is proposed to perform L2R for single image with complete labels. First, each label of the image is used as query for online image search engines to obtain a set of semantically related and visually similar images, which along with the input image are encoded as Bags-of-Hierarchical-Patches. Then, an efficient two-stage feature mining procedure is presented to discover those input-image specific, salient and descriptive features for each label from the proposed Interpolation SIFT (iSIFT) feature pool. These features consequently constitute a patch-level representation, and the continuity-biased sparse coding is proposed to select few patches from the online images with preference to larger patches to reconstruct a candidate region, which randomly merges the spatially connected patches of the input image. Such candidate regions are further ranked according to the reconstruction errors, and the top regions are used to derive the label confidence vector for each patch of the input image. Finally, a patch clustering procedure is performed as post-processing to finalize L2R for the input image. Extensive experiments on three public databases demonstrate the encouraging performance of the proposed nonparametric L2R solution.*

1. Introduction

Label-to-Region (L2R) refers to the task of assigning the labels or keywords annotated at image-level to the unknown local semantic regions within an image. This task is beneficial for improving keyword based image search with the awareness of semantic image content. However, it is usually laborious to manually annotate the image labels, both at region-level and image-level. In this work, we present a novel framework on L2R assignment for single input image, by utilizing the raw outputs of Internet image search results. In fact, for a given image label, we can use it as query on an image search engine, such as BING or GOOGLE, to obtain a set of semantically related and visually similar images, which share a common label/category with each other. This cross-image label context can be utilized to derive the label-specific representations, which are then used to parse the input image into local semantic regions. Figure 1 illustrates the input and output of the proposed Label-to-Region framework, characterized by its remarkable simplicities: i) the input is one single image with label annotation, and ii) it does not require any database of training images that are manually prepared to build object models.

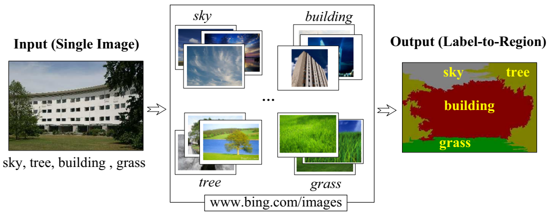

Figure 1. Illustration of the proposed nonparametric L2R-by-Search task.

In the literature, although the specific task of Label-to-Region has not been extensively studied before, there are many related works about image parsing and region labeling [6]. In general, these algorithms involve building object appearance models, as well as higher-level spatial context models in order to overcome the limitations of appearance models. Both types of parametric models require manually prepared training data containing labels at the region level, even though they do not require the knowledge of what labels are present. In particular, there are some related works, known as simultaneous object recognition and image segmentation [13, 26]. These algorithms usually assume that there is only one single label contained in each image, or require manual efforts to prepare the training data [23, 8]. Also, other methods [15, 19] aim to explore the inter-label or label-to-scene correlation, and thus can handle only some specific categories. In contrast, the proposed nonparametric solution has fewer limitations to the input data or the label annotations.

[*] This work was performed when Xiaobai Liu was working as an Research Engineer in National University of Singapore, Singapore.

978-1-4244-6985-7/10/$26.00 ©2010 IEEE

There are also some previous efforts that focus on how to learn object models from the Internet image searching results [18] or unlabeled image collections [27]. These methods are usually based on parametric models, and thus the system performance is limited by the generalization ability of the learned models. Wang *et al*. [24] formulated the image annotation task in a divide-and-conquer framework and proposed a model-free method for image annotation by mining the image search results. Their method, however, is hindered by one strong assumption, i.e., for each input image at least one near-duplicate can be detected in the image dataset, which cannot be always satisfied in practice. In fact, to our best knowledge, no previous efforts ever investigate how to parse a single image into semantic regions by utilizing the outputs of image search engine. It is also worth noting that L2R is different from learning to annotate regions based on weakly labeled data [4] because no object models need to be learned in our case.

In general, the difficulties of L2R-by-Search task lie in the following aspects. First, the searching results usually contain a significant number of "noise" images, which are semantically somewhat related but perceptually different with the input image. Second, there are usually large intra-label variations in the obtained images, meaning that two image regions belonging to the same label may have dramatically different visual appearance characteristics.

We propose a nonparametric solution that addresses the above issues. In our work, an image is represented as a Bag-of-Patches (BOP) at different scales to exploit the rich cues in an image. We also propose a variant of the widely used SIFT [5] feature. Based on the possibly sparse interest points detected by SIFT, a set of new points are interpolated for enhancing the image description capability. We refer to it as Interpolation SIFT (iSIFT). After encoding the BOP representation with iSIFT, an efficient feature mining procedure is introduced for each label to prune the noise images as well as the image patches that are not characteristic for the specific label. As a result, the remaining features are distinctive and descriptive for a specific label and can be used to select relevant patches for a specific label of the input image. However, the image patches in BOP representation are usually of different scales and it is not appropriate to directly compare the feature similarity between such patch pairs. Also, since semantic regions cannot be directly obtained, in this work, we instead propose to first extract candidate regions, by merging from the spatially coherent and perceptually similar image patches in the input image, and then use the detected patches of each label to reconstruct the candidate regions, under the hypothesis that those patches selected for reconstruction should come from a small number of semantically similar images. The reconstruction errors are then used to predict the confidence of containing one specific label, for candidate regions.

The Label-to-Region assignment is facilitated by cross-image correlation and the reconstruction of a candidate semantic region from a set of image patches is achieved by a sparse coding formulation. The basic philosophy is that an image region/patch can be sparsely reconstructed using other image patches belonging to the same semantic label. In addition, an intuitive way to obtain robust correspondence between the input image and the Internet images is to enforce that the matched image patches are spatially connected to each other. Therefore, we additionally introduce a continuity-biased prior to favor larger size patches for reconstruction of candidate regions. Based on the reconstruction coefficients, we can calculate for one image region/patch the confidence of belonging to each label and then fuse all the results to distribute the image labels to those contextually derived semantic regions merged from multiple patches. The proposed L2R-by-Search process has the following characteristics and advantages: i) the sparsity and continuity-biased priors are used to ensure the reliability of label assignment, ii) it does not require exact image parsing, which remains an open problem for real world images, and iii) no generative or discriminative models need be learned for each label, and thus it is extremely scalable for applications with large-scale image sets as well as large semantic ontology.

2. Nonparametric Label-to-Region by Search

2.1. Overview of Problem and Solution

Figure 2 shows the overall flowchart of the proposed solution to L2R-by-Search. First, each given label of the input image is used as a query for the Internet image search engines to obtain a set of semantically related images. Second, we segment both input image and online images returned from image search engine into local atomic image patches to obtain the so-called bag-of-patches (BOP) representation. Then, a label-specific feature mining procedure is employed for each label to discover distinctive and descriptive features from the proposed Interpolation SIFT (iSIFT) feature pool. These features are used to discover the patch-level label-specific representations. Next, we construct the candidate regions, by initially clustering the spatially connected image patches within the input image, and then propose a sparse coding formulation to reconstruct each candidate region. In reconstruction, with the multi-scale representation of BOP, the continuity-biased sparsity prior is introduced to select a small number of patches from the online images with preference to larger patches. The candidate regions are further ranked based on the reconstruction errors and the top ones are used to derive the label confidence vector for each atomic patch of the input image. Finally, a patch clustering procedure is performed on the input image as a post-processing step to obtain the ultimate L2R assignments. It is worthy noting that the entire parsing procedure for a single input image is performed automatically and in-

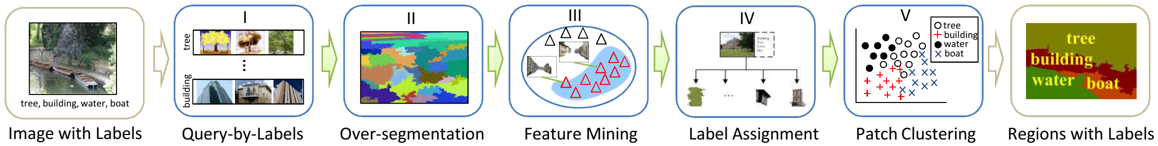

Figure 2. The flowchart of L2R-by-Search procedure. More details refer to the text in Section 2.1.

dividually, making it less laborious and extremely scalable for large-scale image set.

2.2. Image Representation

2.2.1 Bag-of-Patches

An image usually contains a set of semantic regions that are merged from the atomic patches. Each homogeneous patch consists of the pixels that are spatially coherent and perceptually similar with respect to certain appearance features, such as intensity, color and texture. In order to capture rich cues contained in an image, like in [3], we represent an image by constructing a hierarchical tree with the atomic patches as the leaf nodes. Each node of the tree represents a localized image patch that is either further divided into smaller patches, or merged with other patches at the same level to form the parent node. We also remove the links between nodes to obtain an ensemble of image patches at different scales, collectively called the Bag-of-Patches.

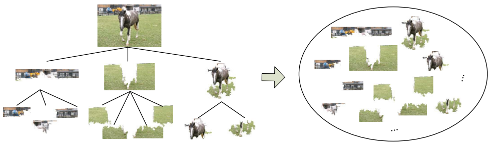

Figure 3. Illustration of patch tree (left) and bag-of-patches (right).

In the implementation, we use the graph-based segmentation algorithm in [16], which incrementally merges smaller-sized patches with similar appearances and with small minimum spanning tree weights. We slightly modify the original algorithm [16] as follows. First, we resize all the images into a roughly equal resolution 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 [16] to merge the smaller patches into larger ones. This step iterates until all the image patches are merged into one single patch, namely the original image. At each iteration, multiple patch pairs are merged and labeled the same depth in the final hierarchical patch tree. 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 this segmentation step. Note that our proposed solution is general and not tied to any specific image segmentation algorithms. The only assumption of this step is that each atomic patch, i.e., the leaf node of the patch tree, is entirely within an object/label. This makes our overall algorithm less sensitive to the quality of the image segmentation step.

2.2.2 Interpolation SIFT features

SIFT (Scale-Invariant-Feature-Transformation) [5] feature has been utilized for many vision and multimedia problems, including stereo matching, object recognition, and image retrieval. Its success is due to its robustness to image noises and scale changes. SIFT feature, however, is generally sparsely detected, that is, given one input image, only the salient interest points are described and the rest of the image, although also potentially informative, are ignored. In this work, in order to generate more informative description of the input image, we propose a variant of the SIFT descriptor based on linear interpolation in the scale space and refer to it as interpolation SIFT (iSIFT).

The basic idea of iSIFT is to interpolate some new interest points between the sparse interest points detected by the standard SIFT detector [5] to enhance the image description capability. Collecting these initially detected points as source anchors, we perform a 2D Delaunay triangulation [14] to obtain a set of non-overlapping triangles and assume the three vertices of each triangle fall on the same object plane [1]. Thus, the parameters of these interest points, including location and scale, should vary smoothly among different vertices of one triangle. For each triangle, one new interest point is generated and the corresponding parameters of location and scale are set as the median of those for the triangle vertices. Figure 4 (a) illustrates the interpolation procedure, where we indicate the scales and orientations of the detected SIFT features by circles of different sizes and directions. The red crosses denote the interest points detected by the algorithm in [5] and the blue crosses denote the newly added iSIFT feature points. Herein, we set the minimal size of triangles for interpolation as 10 pixels. Figures 4 (b-c) show the initially detected interest points and the newly interpolated ones by red crosses and blue crosses, respectively. Generally, the advantages of iSIFT feature include: i) it is the denser version of the original sparse SIFT features and thus can capture more image information, ii) linear interpolation does not add much computation cost over the original dense SIFT features [2], and iii) it is intuitively more informative than the dense SIFT descriptor of fixed scale [2].

2.3. Label-Specific Feature Mining by Search

In this work, an image is described as a visual document composed of repeatable and distinctive basic visual ele-

[1] This assumption is clearly correct when all the triangle vertices fall on the same object. If the vertices fall on different objects, we can also reasonably assume that the points on the border should be smooth on the two connected planes.

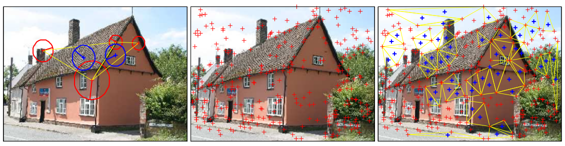

Figure 4. Interpolation SIFT. See more details in Section 2.2.2.

ments that are indexable, namely the Bag-of-Words (BOW). Generally, visual words are usually specific for each label and the visual representation of different category should be distinct from each other. After extracting iSIFT features from image sets, we adopt the method in [22] to build the visual words vocabulary, and set the total number of visual words as N_W=5000. As Figure 4 illustrates, many visual words in the right image appear on the cluttered background instead of the foreground region, namely "building". In order to obtain the informative visual representation for one specific label, we should remove these irrelevant words from the obtained visual vocabulary.

In practice, visual words in each label are only a portion of the entire vocabulary, which means that only a part of the vocabulary is descriptive or informative for the corresponding label. In order to capture objects or scenes, the visual representations should have the following properties: i) the visual words should appear on the input image, ii) the visual words that are informative for a specific label should appear more frequently than other words in the images containing the label, or they should be less frequent in the images not containing the label, and iii) the descriptive visual words should be located on the objects or scenes.

Based on these motivations, we develop a two-stage procedure for deriving the label-specific and input-image specific visual words. First, we remove the words that do not appear in the input image. This hard constraint can prune most of the invalid visual words from the online images returned by the image search engine. Second, we formulate the visual words mining process in a probabilistic inference framework to model two important cues: 1) frequency of each visual word and 2) co-occurrence of each word with other words. Letting $W = \{W_1, W_2, \ldots, W_{N_W}\}$ denote the visual words dictionary, for each label c, we have,

$$P(W|c) \propto \exp\{\sum_{n}^{N_W} \phi(W_n) + \lambda^f \sum_{m \neq n} \phi(W_n, W_m)\} \qquad (1)$$

where the term $\phi(W_n)$ denotes the frequency of the visual word W_n itself, the term $\phi(W_n, W_m)$ denotes the co-occurrence of words W_n and W_m, and λ^f is the tunable tradeoff parameter (we fix it as 1 in this work). With this formulation, we can transform the task of label-specific feature mining task into maximizing $P(W|c)$, which can be efficiently solved by a number of inference methods. In this work, we use the Belief Propagation (BP) algorithm [7] from the Bayes Net Toolbox [12]. According to $P(W|c)$, the label representation can be generated by selecting the top ranked candidates or choosing the ones with rank values larger than a threshold. In this work, we select the top twenty percent of these ranked words as the final label-specific representation. Figure 5 shows in top row of each case the selected visual words after the first step, and in bottom row the final selected words after the second step. Clearly, most of the final selected words are located on the objects, indicating that they are descriptive to the specific label. Note that the mining results are only used to choose the image patches that contain the salient points, whereas the feature descriptor for each patch is still based on the original generated visual words set.

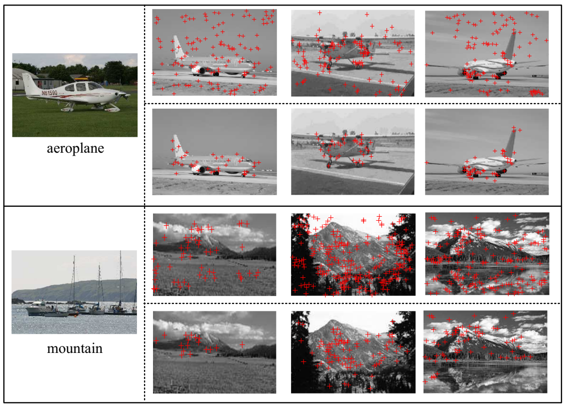

Figure 5. Feature mining results. For each plot, 1^{st} column: input image and query label; 2-4^{th} columns: returned images by search along with the interest points selected after the first step (top row) and the second step (bottom row).

Let I denote the input image and $z \in R^{N_C}$ the label confidence vector of an image patch within I, where N_C is the total number of labels. The component $z(c)$ indicates how likely the image patch contains the c-th label. We generate candidate regions within I, and denote their corresponding feature representations as $\{y_1, \ldots, y_i, \ldots\}, y_i \in R^{N_W}$. We denote N_O as the number of the online images (fixed for each category), and $I_{c,j}$ the j-th online image related to the c-th label. After applying the proposed feature mining procedure on those images, for each category c, the obtained feature patches in $I_{c,j}$ are arranged in a matrix, denoted as $X_{c,j} = [x_{c,j,1}, \ldots, x_{c,j,n_{c,j}}] \in R^{N_W \times n_{c,j}}$, where $n_{c,j}$ is the number of selected patches within $I_{c,j}$. We further collect and arrange the visual words representations of all the N_C labels into one single matrix, given by $A = [X_1, \ldots, X_c, \ldots, X_{N_C}]$, where $X_c = [X_{c,1}, X_{c,2}, \ldots, X_{c,N_O}]$, and use A as the basis dictionary in the following reconstruction step.

2.4. Sparse Region Coding with Continuity-Prior

We propose a sparse coding formulation to discover the cross-image region/patch correspondence. This is achieved by predicting how well a candidate region can be reconstructed from the patches generated by the feature mining step discussed in previous subsection. In fact, if sufficient patch samples are available for each label, it is possible to represent a candidate region as a sparse and linear combination of the patch representations. Letting y denote the feature descriptor of the candidate region, we have,

$$y = A\,\alpha_0 + \epsilon, \tag{2}$$

where α_0 is the coefficient vector, whose entries are expected to be zeros except for those samples containing the same label as y, and $\epsilon \in R^{N_W}$ is a noise vector which explicitly accounts for the possible sparse noises.

Theoretically, α_0 can be obtained by solving the linear system of equation $y = A\alpha$, but when $N_W < N = \sum_{c,j} n_{c,j}$, there exist infinite number of possible solutions. A possible way to select a sparse solution is to minimize the ℓ_0-norm of the solution, and a recent development in theories on sparse representation [10] reveals that if the ℓ_0-norm solution $\hat{\alpha}_0$ is sparse enough, the solution from the ℓ_0-norm minimization can be recovered by the solution to the ℓ_1-norm minimization problem, namely,

$$\arg\min_{\alpha,\epsilon} ||\alpha||_1 + ||\epsilon||_1, \textit{s.t.}\;\; A\,\alpha = y + \epsilon. \tag{3}$$

This optimization problem is convex and can be transformed into a general linear programming problem. There exists a globally optimal solution that can be solved efficiently using the classical ℓ_1-norm optimization toolkit [1].

The reconstructions of candidate regions are with sparsity prior, which means that we prefer to select as few patches as possible. Since our goal is to discover the cross-image correspondence, it is natural to additionally enforce that the selected patches are perceptually and spatially coherent. This motivation leads to the preference to image patches with larger size, namely the continuity-biased prior. Mathematically, let $B \in R^{N\times N}$ denote the correlation matrix, in which, if the element $B_{ij} = 1$ for $i \neq j$, then the i-th patch is the grand-parent node of the j-th patch; otherwise, $B_{ij} = 0$. Also, the diagonal elements of B are set as 1. Then, we rewrite Eq. (3) as,

$$\arg\min_{\alpha,\epsilon} ||\alpha||_1 + ||\epsilon||_1 + ||B\alpha||_1,\; \textit{s.t.}\;\; y = A\,\alpha + \epsilon, \tag{4}$$

which additionally imposes the continuity prior by setting the weight of a node as the summation of the corresponding child atomic patches. Due to the sparsity prior, minimizing the ℓ_1-norm term $||B\alpha||_1$ in Eq. (4) shall result in the preference to the larger size patches, namely the upper-level nodes in the patch tree, since the selection of a subset of smaller size patches shall bring larger ℓ_1-norm than the selection of one single larger size patch.

Letting $\gamma = B\alpha$,

$$y' = \begin{bmatrix} y \\ 0_{N\times 1} \end{bmatrix}, \alpha' = \begin{bmatrix} \alpha \\ \epsilon \\ \gamma \end{bmatrix}, A' = \begin{bmatrix} A, I_{N_W\times N_W}, 0_{N_W\times N} \\ B, 0_{N\times N_W}, -I_{N\times N} \end{bmatrix},$$

we can reformulate (4) as,

$$\hat{\alpha}'_1 = \arg\min_{\alpha'} ||\alpha'||_1, \quad \textit{s.t.}\;\; y' = A'\alpha', \tag{5}$$

where the derived coefficient $\hat{\alpha}_1$ is both sparse and continuity-biased. Figure 6 demonstrates one example comparison result on how a candidate region is reconstructed from the bag of patches guided by different priors. For ease of display, we rank the selected image patches according to the reconstruction coefficients, and plot only the top five ones. From the results, we can observe that adding the continuity-biased prior leads to selecting larger and more meaningful patches in BOP representations.

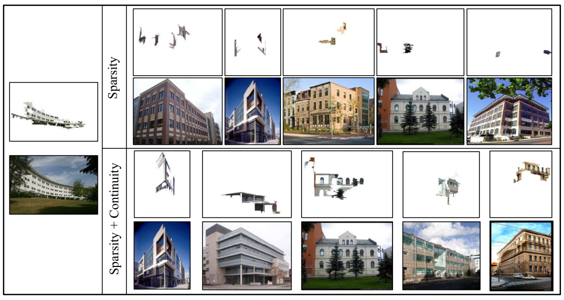

Figure 6. Example reconstruction results with different priors. The 1^{st} column shows a candidate region and its source image. In other columns, the top row shows the patches selected using sparsity prior only, and the bottom row shows the patches selected using both the sparsity and continuity-biased priors. The input image is from MSRC dataset [9] and the online images are from BING.

Algorithm 1 . Label-to-Region Assigning via Sparse representation.

```
1: Input: selected patch-level label representation $A = [X_1, \ldots, X_{N_C}]$; feature of one candidate region $y \in R^{N_W}$;
2: Normalize the columns of $A$ and $y$ to have unit $\ell_2$-norm; Initialize the label confidence vector $z_y \in R^{N_C}$ of $y$, as $z_y = 0$;
3: Solve the optimal solution $\hat{\alpha}$ according to (5);
4: For each label $c$ annotated with the input image, calculate for $y$ the confidence of belonging to the $c$-th label based on the reconstruction residual, namely, $z_y(c) \propto exp\{-\|y - A\hat{\alpha}_c\|_2\}$;
5: Output: $z_y \in R^{N_C}$;
```

Algorithm 2 . Post-processing for L2R-by-Search

```
1: Input: label confidence vector $z$ of the input image $I$; label confidence vectors for all the atomic patches in image $I$, denoted as $\{z_i\}$, $i = 1, \ldots, N_A$, where $N_A$ is the number of atomic patches in the BOP representation of $I$;
2: Set $K$ as the number of image labels provided for $I$;
3: Cluster the atomic patches by grouping all the patch-level label confidence vectors $\{z_1, \ldots, z_{N_A}\}$ into $K$ clusters, denoted as $\{O_1, \ldots, O_K\}$;
4: For each cluster $O_c \in \{O_1, \ldots, O_K\}$
    4.3.1: Let $z_m$ denote the summed label vector for each cluster, calculated as $z_m = \sum_{z_j \in O_c} z_j$;
    4.3.2: Set $z_m$ as the label vector of each atomic patch belonging to the cluster $O_c$;
5: Merge those patches within the same cluster to form a semantic region, and set its label as the one with the largest value in the label vector and without overlapping the label with other regions.
6: Output: Merged patches with semantic labels;
```

2.5. L2R Assignment via Sparse Representation

Given a candidate region y of the input image and the feature basis matrix A, we first compute its sparse representation $\hat{\alpha}$ by solving (5). Then, we classify y based on how well the coefficients associated with all image patches of each label reproduce y. Letting $\hat{\alpha}_c$ be a new vector whose nonzero entries are the entries in $\hat{\alpha}$ that are associated with the label c, one can approximate the given candidate region by using $\hat{\alpha}_c$, as $\hat{y}_c = A\hat{\alpha}_c$, and calculate the label confidence of y for the c-th label based on these approximations between y and $\hat{y}_c$,

$$||y - A\hat{\alpha}_c||_2, \quad c = 1, 2, \ldots, N_C. \tag{6}$$

Algorithm 1 summarizes the entire L2R procedure.

Suppose the label confidence vector for each atomic patch in the input image has been derived by Algorithm 1, we adopt the K-means clustering approach over all the confidence vectors of the atomic image patches to generate K clusters, where K is the number of labels annotated for the input image. Thus, each cluster corresponds to one single semantic region and the patches belonging to the same cluster should be assigned with one identical label. Algorithm 2 summarizes the entire post-processing procedure.

3. Experiments

In this section, we systematically evaluate the effectiveness of our proposed iSIFT feature pool, feature mining procedure and the continuity-biased sparse coding formulation for Label-to-Region assignment task.

3.1. Datasets

We use three publicly available datasets, MSRC [9], COREL, and the dataset collected by Stephen *et al.* [20] as test or input images in this work. The MSRC dataset contains 591 images from 23 categories/labels and region-level ground-truths. There are about 3 labels on average for each image. We remove the images which are annotated with only one single label, resulting in about 500 images. The second dataset is from COREL collection, the most broadly adopted dataset for image retrieval. We use the subset provided in [11], which includes $4,000$ images from 8 labels and the corresponding ground truth of region-level annotations. The third database is collected by Stephen et al. [20], which contains 715 images from LabelMe, MSRC and PASCAL VOC. The ground truth of region-level annotations are available and the images are from 7 labels. We resize the images to the extent of $400/max(width, height)$, and set the minimal image patches in BOP representation as 400 pixels in a tradeoff between efficiency and performance. Thus, there are about 300-400 image patches contained in each BOP. We also randomly select 800 images from above 3 databases and utilize their corresponding iSIFT features to generate the N_W=5000 visual words. All the experiments are performed on an Intel Xeon $X5450$ workstation with 3.0 GHz CPU and 16 GB memory. The code is implemented in MATLAB platform. Generally, the proposed method can perform L2R for one image within 30 minutes without any code optimization based on the top ranked N_O=100 online images for each label.

3.2. Baselines

We implement the proposed Label-to-Region Assignment vis Sparse representation (LAS) algorithm using the ℓ_1-Magic package [1], which first translates (5) into a linear programming problem and then employs the primal-dual algorithm to perform the optimization. We set the tolerance factor as 0.003 and the maximum number of primal-dual iterations as 50.

Two types of algorithms are used as baselines to evaluate the proposed bi-layer sparse coding formulation in the label to region assignment task. One is a SVM-based algorithm that first learns a model from the training samples for each label and then applies the obtained models to the testing samples extracted from the input image. Herein, the training samples indicate the patches selected by the proposed feature mining procedure. For each classifier, a patch is considered as a positive sample if it comes from the online images of the specific label, otherwise it is considered as a negative sample. In the training stage, we choose equal number of positive and negative samples and remove the excessive ones to balance the training of SVM. In testing, we first apply each classifier to image patches of different scales and then use the top ranked results to obtain the confidence of containing the specific label for each atomic patch in the input image. Results from all the classifiers are then fused to generate the N_C-dimensional label confidence vector, which is further processed by Algorithm 2 to obtain the labels of the atomic patches. Note that the training and testing procedures work on image patches of different scales and the ultimate goal is to obtain the semantic label annotation at the region-level. A binary SVM is implemented based on the SVM library [17]. A Gaussian Radial Basis Function kernel is used with the kernel parameter set as 1.

The other baseline algorithm uses the traditional K-NN method. In implementation, for each image patch of the input image, we select 50 nearest ones from the patches ensemble selected by the proposed feature mining procedure. Letting S_c denote the number of patches belonging to the c-th label, we can calculate the label confidence vector for specific image patch as $\frac{1}{50}[S_1, \ldots, S_{N_C}]$, and the top ranked ones in BOP representation are used to calculate the confidence vector of each atomic patch, which are further processed by Algorithm 2 to obtain the ultimate L2R results.

In order to evaluate the proposed iSIFT feature pool, we implement both LAS and the baselines by using two different feature descriptors: I) the standard dense SIFT feature [2] that collects one SIFT feature for each lattice of 10×10 pixels with a fixed scale factor of 16 pixels, and II) the iSIFT feature, where we generate new points for the triangles with size being no less than 10 pixels. Moreover, we also implement a simplification of the proposed solution to demonstrate the improvement brought by the proposed continuity-biased prior. The overall procedure is identical

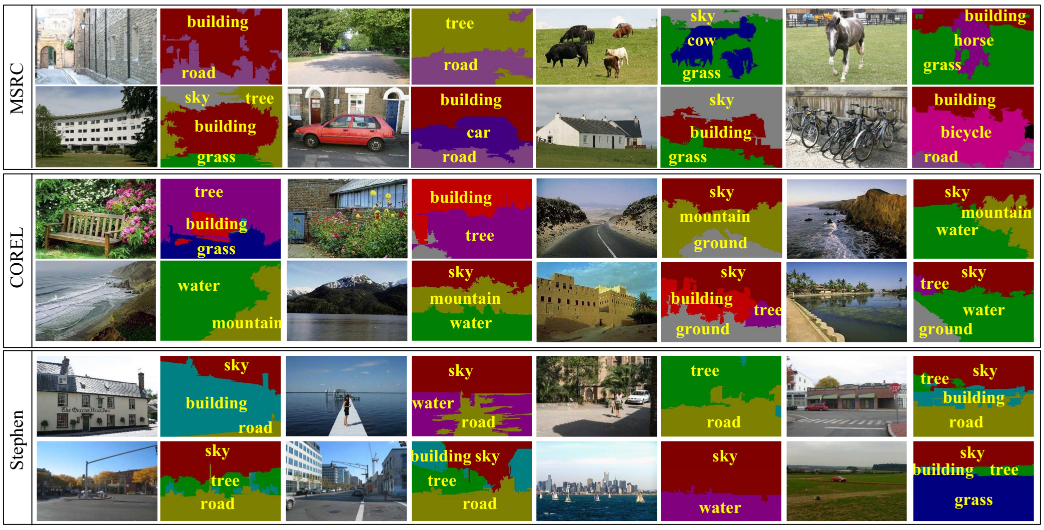

Figure 7. Example results on the MSRC, COREL and database by Stephen *et al.* in [20]. The original input images are shown in columns $2, 4, 6, 8$ and the corresponding parsed images associated with their region-level labels are shown in columns $3, 5, 7, 9$.

to the LAS, except that the system optimizes Eq. (3), instead of Eq. (5), in the 3rd step of Algorithm 1. For ease of representation, we denote the versions with different priors as: A) with only the sparsity priors, and B) with both the sparsity and continuity-biased priors. Thus, we obtain seven algorithms, i.e., 1) SVM-I; 2) SVM-II; 3) KNN-I; 4) KNN-II; 5) LAS-A-I, 6) LAS-A-II and 7) LAS-B-II.

The L2R assignment performance is evaluated in both qualitative and quantitative manners. The quantitative Label-to-Region assignment accuracy measures the percentage of pixels with agreement between the assigned label and ground truth.

3.3. Results and Discussions

Qualitative Evaluations. We show example results of L2R assignment in Figure 7 for the MSRC, COREL and database used in [20], respectively. The image search engine used here is BING. These results over various conditions well validate the effectiveness of our proposed solution. It is worth noting that our algorithm is scalable to large-scale applications and can be easily extended to perform online image parsing. The algorithm is amenable to fast implementation since the entire algorithm is suitable for parallel computation, and also pre-processing tricks can be utilized to further improve the computational efficiency. For example, one may first search all the possible labels to obtain the related online images and then perform BP algorithm in the feature mining step to build an off-line preliminary feature basis dictionary for each label. Also, the text-based semantic dictionary (e.g. WORDNET) can be used to discover the semantically related keyword or label groups to further improve the feature pruning procedure by downloading more related images.

Quantitative Evaluations. We report in Table 1 a comparison among the accuracies of different algorithms on these three datasets. For each input image, we use two different image search engines, BING (B) and GOOGLE (G). The detailed results for individual categories are shown in Figure 8. We can have following observations. 1) The proposed solution, namely LAS-B-II, achieves much higher accuracies on all the three databases as compared to both the KNN- and SVM- based methods. This clearly demonstrates the effectiveness of the sparse coding formulation for building the cross-image correspondence. 2) The algorithms SVM-II, KNN-II and LAS-A-II, which use iSIFT feature pool, outperform their counterparts, SVM-I, KNN-I and LAS-A-I, which use the standard densely sampled SIFT feature pool, respectively. It well demonstrates the advantages of the iSIFT feature pool over the dense SIFT features of fixed scale. 3) The comparison results of LAS-A-II and LAS-B-II show that the continuity-biased sparse coding can boost the performance of L2R assignment.

We do not further compare our solution with those algorithms for classifying and localizing objects in images [13, 26, 23, 19], because: a) our proposed solution works under the assumption that no region-level label annotation is provided for model training, which is however the general prerequisite for most typical algorithms; b) for each image label, those algorithms need to learn an individual detector, and thus are labor consuming and impractical for large-scale applications; and c) our solution works in an online fashion, meaning that image parsing is done without any additional training procedure.

Moreover, our solution distinguishes it from the closely related work in [25] in the following significant aspects. First, [25] parses a set of input images in batch by utilizing the cross-image contextual priors whereas our method can handle one single input image without any user-provided contextual knowledge. This makes the proposed algorithm

Table 1. Comparison of Label-to-Region assignment accuracies of different algorithms on MSRC, COREL and the database provided by Stephen *et al*. [20]. Herein, we use two different image search engines, BING (B) and GOOGLE (G).

Dataset	MSRC		COREL		Stephen[20]	
	B	G	B	G	B	G
SVM-I	0.33	0.29	0.31	0.33	0.38	0.36
SVM-II	0.42	0.36	0.38	0.43	0.42	0.44
KNN-I	0.48	0.45	0.46	0.44	0.46	0.48
KNN-II	0.51	0.51	0.52	0.52	0.50	0.52
LAS-A-I	0.57	0.55	0.56	0.53	0.55	0.58
LAS-A-II	0.62	0.61	0.55	0.59	0.62	0.62
LAS-B-II	**0.67**	**0.63**	**0.58**	**0.60**	**0.64**	**0.63**

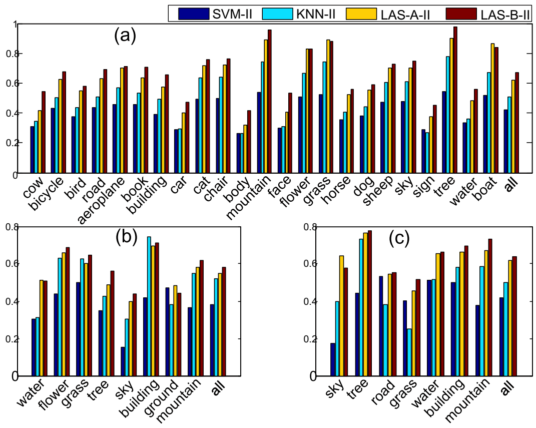

Figure 8. Detailed Label-to-Region accuracies on (a) MSRC, (b) COREL and (c) database by Stephen *et al*. [20]. The horizontal axis shows the name of each label and the vertical axis indicates the Label-to-Region assignment accuracies.

much more appealing for practical applications. Second, the label propagation algorithm provided in [25] requires that there are both differences and commonalities between the label annotations of different input images. Thus it cannot be directly applied to the COREL dataset as used in this work, since most of the images are provided with the same label annotations and there is less cross-image contextual information to rely on. In contrast, our method has much fewer limitations to the label annotations and thus is more attractive. Third, our method achieves a higher accuracy of 0.67 compared to the accuracy of 0.63 in [25] on the MSRC dataset. It is worth noting that our solution can also be used for simultaneous multi-label image annotation and parsing task.

4. Conclusions and Future Work

In this paper, we propose a nonparametric solution for automatically parsing a single input image with image-level label annotations into localized semantic regions by utilizing the auxiliary knowledge from the raw outputs of web image searches. Since the community-contributed images with rich tag information are becoming much easier to obtain, we expect that the keyword-query based image search can greatly benefit by applying our proposed technique to these tagged images.

The current solution can only process images with reasonably complete label annotation, and in the future we plan to relax this assumption and further investigate how to handle the images with partial annotation or noisy labels. Ultimately, we intend to extend the proposed framework to perform image parsing for a set of unlabeled images with semantically related regions.

5. Acknowledgement

This work is done for CSIDM Project No. CSIDM-200803 partially funded by a grant from the National Research Foundation (NRF) administered by the Media Development Authority (MDA) of Singapore. This work is also partially supported by the China 863 program (Grant No.2008AA01Z126).

References

[1] http://www.acm.caltech.edu/l1magic.

[2] A. Vedaldi and B. Fulkerson, VLFeat: An Open and Portable Library of Computer Vision Algorithms. 2008, http://www.vlfeat.org/

[3] C. Gu, J. Lim, P. Arbeláez, and J. Malik, Recognition using Regions. In *CVPR*, 2009.

[4] C. Galleguillos B. Babenko, A. Rabinovich, and S. Belongie, Weakly Supervised Object Localization with Stable Segmentations. In *ECCV*, 2008.

[5] D. Lowe, Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[6] L. Li, R. Socher, and F. Li, Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework. In *CVPR*, 2009.

[7] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. *Morgan Kaufmann Publishers Inc.*, 1988.

[8] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu, and S. Ma, Dual cross-media relevance model for image annotation. In *ACM MM*, 2007.

[9] J. Shotton, J. Winn, C. Rother, and A. Criminisi, Textonboost: Joint appearance, shape and context modeling for mulit-class object recognition and segmentation. In *IJCV*, 2009.

[10] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, Robust Face Recognition via Sparse Representation. *TPAMI*, 2009.

[11] J. Yuan, J. Li, and B. Zhang, Exploiting spatial context constraints for automatic image region annotation. In *ACM MM*, 2007.

[12] K. Murphy, The bayes net toolbox for matlab. *Computing Science and Statistics*, 2001.

[13] L. Cao and F. Li, Spatially coherent latent topic model for concurrent object segmentation and classification. In *ICCV*, 2007.

[14] L. Guibas and J. Stolfi, Primitives for the manipulation of general subdivisions and the computation of Voronoi. *TOG*, 1985

[15] M. Zhang and Z. Zhou, Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 2007.

[16] P. Felzenszwalb and D. Huttenlocher, Efficient graph-based image segmentation. *IJCV*, 2004.

[17] R. Fan, P. Chen, and C. Lin. Working set selection using the second order information for training svm. In *Journal of Machine Learning Research*, 2005.

[18] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, Learning object categories from Google's Image Search. In *ICCV*, 2005.

[19] R. Jin, J. Chai, and L. Si, Effective automatic image annotation via a coherent language model and active learning. In *ACM MM*, 2004.

[20] S. Gould, R. Fulton, and D. Koller, Decomposing a Scene into Geometric and Semantically Consistent Regions. In *ICCV*,2009.

[21] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, Descriptive visual words and visual phrases for image applications. In *ACM MM*, 2009.

[22] 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 *ACM CIVR*, 2009.

[23] V. Lavrenko, R. Manmatha, and J. Jeon, A model for learning the semantics of pictures. In *NIPS*, 2004.

[24] X. Wang, L. Zhang, X. Li, and W. Ma, Annotating Images by Mining Image Search Results. *TPAMI*, 2008

[25] X. Liu, B. Cheng, S. Yan, J. Tang, T. Chua, and H. Jin, Label to Region by Bi-Layer Sparsity Priors. In *ACM MM*, 2009.

[26] Y. Chen, Unsupervised learning of probabilistic object models (poms) for object classification, segmentation and recognition. In *CVPR*, 2008.

[27] Z. Wu, Q. Ke, M. Isard, and J. Sun, Building Features for Large Scale Partial-Duplicate Web Image Search. In *CVPR*, 2009.