Automatic Image Annotation with Weakly Labeled Dataset*

Wei Zhang[1], Yao Lu[1], Xiangyang Xue[1], Jianping Fan[2]
[1]School of Computer Science, Fudan University, Shanghai, China
[2]Department of Computer Science, UNC-Charlotte, NC28223, USA
{weizh, yaolu, xyxue}@fudan.edu.cn jfan@uncc.edu

ABSTRACT

It is very attractive to exploit weakly-labeled image dataset for multi-label annotation applications. In our paper the meaning of the terminology *weakly labeled* is threefold: i) only a small subset of the available images are labeled; ii) even for the labeled image, the given labels may be uncorrect or incomplete; iii) the given labels do not provide the exact object locations in the images. A novel method is developed to predict the multiple labels for images and to provide region-level labels for the objects. We cluster the image regions to learn several region-exemplars and predict the label vector for each image region as a locally weighted average of the label vectors on exemplars. By investigating the label confidence matrix for the region-exemplars from different perspectives (*column picture* and *row picture*), we sufficiently leverage the visual contexts, the semantic contexts, and the consistency between similarities in the visual feature space and semantic label space. Experimental results on real web images demonstrate the effectiveness of the proposed method.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation, Performance

Keywords

Multi-Label Image Annotation, Weakly-Labeled Dataset, Region-Level Labels

1. INTRODUCTION

With the massive explosion of web images, how to access these images efficiently is an important research task, and automatic image annotation has become more and more attractive. In real image annotation applications, multiple semantic concepts may occur simultaneously in an image; each individual label of one image is actually related to local regions rather than the whole image. It is attractive to learning the image classifiers from *weakly labeled* images available on the web. Herein the terminology *weakly labeled* can be investigated from three perspectives: i) Since manually annotating is time-consuming and labor-intensive, only a subset of images are labeled, and the number of available labeled images is much smaller than that of unlabeled ones; ii) In a collaborative image tagging system, people can tag the images according to their personal expertise and perception, and the label set for each labeled image may be uncorrect or incomplete; iii) The multiple labels are given loosely at the image level rather than at the object level (i.e., without providing the exact object locations in the images). Our goal is to precisely predict the multiple labels for images and to provide region-level labels for the objects, simultaneously.

In [2] a framework was proposed to improve the retrieval performance by refining noisy labels of a group of Flickr photos. [17] proposed a label refinement formulation considering the label characteristics from the points of view of low-rank, error sparsity, content consistency and label correlation. [15] and [14] both proposed graph-based semi-supervised learning frameworks, which can simultaneously explore the correlations among multiple labels and the label consistency over the graph. Those methods above inferred the correspondence between the images and their associated labels only at the image level rather than at the region level. [8] proposed a unified formulation to implement various tag analysis tasks including label-to-region assignment in a coherent way; however the correlations between labels are not exploited in [8].

In this paper a novel method is developed to simultaneously perform label prediction and label-to-region assignment based on weakly-labeled web image dataset. Each image is firstly segmented into several regions. We cluster the image regions to learn a small number of region-exemplars and predict the label vector for each image region as a locally weighted average of the label vectors on exemplars. By investigating the label confidence matrix for the region-exemplars from different perspectives (*column picture* and *row picture*), our method sufficiently leverages the visual contexts, the semantic contexts, and the associations between image-level and region-level labels. Different from [8], the correlations between the semantic concepts are effec-

*Area chair: Bernard Merialdo

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

tively captured in our method, which is of significance to the performance of multi-label image annotation. Experimental results on the real web images demonstrate the effectiveness of the proposed method.

2. ALGORITHM

Suppose that $\{(\mathbf{x}^1, \mathbf{y}^1), ..., (\mathbf{x}^l, \mathbf{y}^l)\}$ are l labeled images for training. Each image includes r_i regions: $\mathbf{x}^i = \{\mathbf{x}^i_1, ..., \mathbf{x}^i_{r_i}\}$, where the number of regions r_i may vary across images. Let $\mathcal{C} = \{c_1, ..., c_m\}$ be a semantic lexicon of m concepts, and let $\mathbf{y}^i = [\mathbf{y}^i_1, \ldots, \mathbf{y}^i_m]^\top \in \{0, 1\}^m$ be the initial $m-$dimensional label vector corresponding to the image $\mathbf{x}^i$, where $\mathbf{y}^i_s = 1(s = 1, \ldots, m)$ if the concept c_s is associated with the image $\mathbf{x}^i$, and 0 otherwise. Our goal is to predict the label vector for any new image and to predict the region-level label simultaneously.

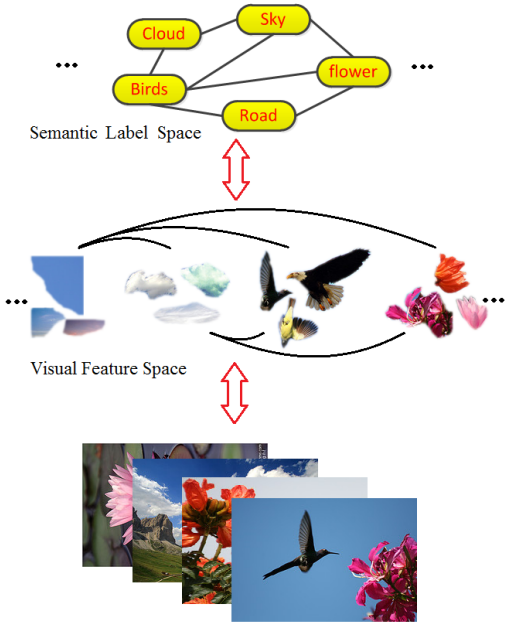

Figure 1: An overview of our method. Each image is firstly segmented into regions. All regions are clustered into several groups. In the visual feature space, region-exemplars are used to construct the visual context graph. In the semantic label space, concepts form the semantic context graph which captures the correlations between concepts.

Each image is firstly segmented into several regions. Let $\mathbf{h}^i_j \in [0, 1]^m (j = 1, ..., r_i)$ denote the label confidence vector of $\mathbf{x}^i_j$ (the $j-$th region in the $i-$th image), and the $s-$th component of $\mathbf{h}^i_j$ just measures the probability that the concept c_s is associated with the region $\mathbf{x}^i_j$. To address the scalability issue, all the image regions are clustered into several groups by clustering algorithms such as *Affinity Propagation* [6] based on the visual similarity. Suppose that n clusters are obtained, and one region-exemplar is learned for each cluster, then we get n region-exemplars $\{\dot{\mathbf{x}}_1, ..., \dot{\mathbf{x}}_n\}$. Inspired by [12], we approximately reconstruct each image region as a convex combination of its nearest exemplars:

$$\mathbf{x}^i_j \approx \sum_{q \in \langle \mathbf{x}^i_j \rangle} \pi_{ijq} \dot{\mathbf{x}}_q, \tag{1}$$

where $\langle \mathbf{x}^i_j \rangle$ denotes the index set of k closest exemplars for $\mathbf{x}^i_j$, and the combination coefficients π_{ijq} can be learned by quadratic programming (QP) as follows:

$$\min \frac{1}{2} \left\| \mathbf{x}^i_j - \sum_{q \in \langle \mathbf{x}^i_j \rangle} \pi_{ijq} \dot{\mathbf{x}}_q \right\|^2, s.t. \quad \pi_{ijq} \geq 0, \sum_{q \in \langle \mathbf{x}^i_j \rangle} \pi_{ijq} = 1 \tag{2}$$

Let $\dot{\mathbf{h}}_q \in [0, 1]^m$ denote the label confidence vector for the exemplar region $\dot{\mathbf{x}}_q, (q = 1, ..., n)$, which can be viewed as a point in the semantic label space. It it reasonable to preserve the neighborhood contexts when mapping image regions from the visual feature space to the semantic label space. Then, the label confidence vector for the image region $\mathbf{x}^i_j$ can be estimated as a locally weighted average of the labels on exemplars:

$$\mathbf{h}^i_j = \sum_{q \in \langle \mathbf{x}^i_j \rangle} \pi_{ijq} \dot{\mathbf{h}}_q, \tag{3}$$

Let $\pi_{ijq} = 0$ if $q \notin \langle \mathbf{x}^i_j \rangle$ and denote $\vec{\alpha}^i_j = [\pi_{ij1}, ..., \pi_{ijn}]^\top$, $\dot{H} = [\dot{\mathbf{h}}_1, ..., \dot{\mathbf{h}}_n] \in [0, 1]^{m \times n}$, then Eq. (3) can be expressed as:

$$\mathbf{h}^i_j = \dot{H} \vec{\alpha}^i_j, \tag{4}$$

Once the region-level label confidence vectors are learned, the image-level label confidence vectors can be estimated by $\sum_{j=1}^{r_i} \omega^i_j \mathbf{h}^i_j$ (herein ω^i_j is the voting weight and can be selected as the percentage of covered area of each region). To achieve the coherence between region-level labels and image-level labels, and thus realize the cross-level label propagation, we minimize the following loss function:

$$\min \sum_{i=1}^{l} \left\| \mathbf{y}^i - \sum_{j=1}^{r_i} \omega^i_j \mathbf{h}^i_j \right\|^2 = Tr\big((Y - \dot{H}A)^\top (Y - \dot{H}A)\big) \tag{5}$$

where $Y = [\mathbf{y}^1, ..., \mathbf{y}^l] \in \{0, 1\}^{m \times l}$, $\dot{H} = [\dot{\mathbf{h}}_1, ..., \dot{\mathbf{h}}_n] \in [0, 1]^{m \times n}$, and $A = [\sum_{j=1}^{r_1} \omega^1_j \vec{\alpha}^1_j, ..., \sum_{j=1}^{r_l} \omega^l_j \vec{\alpha}^l_j] \in R^{n \times l}$.

The label confidence matrix for the region-exemplars $\dot{H} = [\dot{\mathbf{h}}_1, ..., \dot{\mathbf{h}}_n] \in [0, 1]^{m \times n}$ can be investigated from different perspectives: i)*column picture*: Each column of $\dot{H}$ corresponds to the label distribution for one region-exemplar , and $\dot{\mathbf{h}}_q (q = 1, ..., n)$ can be viewed as the high-level feature vector of the $q-$th region-exemplar in the semantic label space, which differs from the low-level feature vector $\dot{\mathbf{x}}_q$ in the visual feature space; ii)*row picture*: Each row of $\dot{H}$ can be viewed as the voting scores from all the region-exemplars for each concept, and can also be viewed as the feature vector of the concept. By constructing different graphs from above two perspectives, we can sufficiently leverage visual context, semantic context, and the consistency between visual features and semantic concepts, to improve the performance of image annotation.

Let $G = (\{\dot{\mathbf{x}}_q\}_{q=1}^n, S)$ be the visual context graph with the vertex set corresponding to the region-exemplars $\{\dot{\mathbf{x}}_q\}_{q=1}^n$ and the adjacent matrix S measuring the visual similarities

between region pairs. S is an $n \times n$ sparse symmetric matrix and can be defined using visual features as follows:

$$S_{pq} = \begin{cases} exp\{-\rho \mathbf{d}(\dot{\mathbf{x}}_p, \dot{\mathbf{x}}_q)\} & \dot{\mathbf{x}}_p \in \langle \dot{\mathbf{x}}_q \rangle \vee \dot{\mathbf{x}}_q \in \langle \dot{\mathbf{x}}_p \rangle \\ 0 & otherwise \end{cases} \quad (6)$$

where $\langle \dot{\mathbf{x}}_q \rangle$ denotes the set of k closest exemplars for $\dot{\mathbf{x}}_q$, $\mathbf{d}(\dot{\mathbf{x}}_p, \dot{\mathbf{x}}_q)$ is the distance between exemplars $\dot{\mathbf{x}}_p$ and $\dot{\mathbf{x}}_q$, and ρ is the scaling parameter. The visual and semantic consistency for the exemplar regions can be achieved by solving the following problem:

$$\min \frac{1}{2} \sum_{pq} S_{pq} \big\| col(\dot{H}, p) - col(\dot{H}, q) \big\|^2 = Tr\big(\dot{H}(\widetilde{S} - S)\dot{H}^\top\big) \quad (7)$$

where $col(\dot{H}, p)$ denotes the $p-$th column of $\dot{H}$, i.e., $\dot{\mathbf{h}}_p$, and $\widetilde{S}$ is an $n \times n$ diagonal matrix with $\widetilde{S}_{qq} = \sum_{p=1}^{n} S_{pq}$.

On the other hand, let $G' = (\{c_s\}_{s=1}^m, W)$ be the semantic context graph with the vertex set corresponding to the concepts $\{c_s\}_{s=1}^m$ and the adjacent matrix W measuring the correlations between concepts. W is an $m \times m$ sparse symmetric matrix and can be defined as follows:

$$W_{st} = \begin{cases} exp\{-\sigma \mathbf{d}(c_s, c_t)\} & \mathbf{d}(c_s, c_t) < \epsilon \\ 0 & otherwise \end{cases} \quad (8)$$

where σ and ϵ are the parameters, and $\mathbf{d}(c_s, c_t)$ is the Normalized Google Distance (NGD) [4] between concepts c_s and c_t:

$$\mathbf{d}(c_s, c_t) = \frac{max\{logf(c_s), logf(c_t)\} - logf(c_s, c_t)}{logN - min\{logf(c_s), logf(c_t)\}} \quad (9)$$

where $f(c_s)$ is the numbers of the webpages returned by Google search engine when typing c_s as the search term, $f(c_s, c_t)$ is the number of webpages returned when typing c_s and c_t together as the search term, and N is the total number of the images in Google. The smaller the NGD is, the stronger the semantic relation is; so, the weight W_{st} measures the affinity between concepts c_s and c_t. At the same time, each row of $\dot{H}$ corresponds to the voting scores from all the region-exemplars for each concept and can be viewed as the feature vector of the concept. Strongly correlated concepts concepts should have similar voting scores, which can be achieved by solving the following problem:

$$\min \frac{1}{2} \sum_{st} W_{st} \big\| row(\dot{H}, s) - row(\dot{H}, t) \big\|^2 = Tr\big(\dot{H}^\top(\widetilde{W} - W)\dot{H}\big) \quad (10)$$

where $row(\dot{H}, s)$ denotes the $s-$th row of $\dot{H}$, and $\widetilde{W}$ is an $n \times n$ diagonal matrix with $\widetilde{W}_{tt} = \sum_{s=1}^{m} W_{st}$.

Therefore, by incorporating various contextual relations in a single framework as shown in Figure. 1, the proposed model for automatic image annotation takes the formulation as:

$$\begin{aligned} \min_{\dot{H}} \quad & Tr\big((Y - \dot{H}A)^\top (Y - \dot{H}A)\big) + \\ & \theta_1 Tr\big(\dot{H}(\widetilde{S} - S)\dot{H}^\top\big) + \theta_2 Tr\big(\dot{H}^\top(\widetilde{W} - W)\dot{H}\big) \end{aligned} \quad (11)$$

where θ_1 and θ_2 are the controlling parameters. Let the derivative of the above cost function with respect to $\dot{H}$ be zero, we have

$$\dot{H}(\theta_1(\widetilde{S} - S) + AA^\top) + \theta_2(\widetilde{W} - W)\dot{H} = YA^\top \quad (12)$$

which is essentially a Sylvester equation [15] widely used in control theory. Vectorizing the unknown matrix $\dot{H}$, Eq. (12) can be transformed to a linear system:

$$\begin{aligned} & [(\theta_1(\widetilde{S} - S) + AA^\top) \otimes I_m + I_n \otimes (\theta_2(\widetilde{W} - W))] vec(\dot{H}) \\ & = vec(YA^\top) \end{aligned} \quad (13)$$

where $vec(.)$ is the vectorization of the matrix, $\otimes$ is the Kronecker product, and I_m and I_n are $m \times m$ and $n \times n$ identity matrices, respectively. We can efficiently solve $vec(\dot{H})$ in Eq. (13) by a generalized minimal residual algorithm [13], and then obtain $\dot{H}$ from $vec(\dot{H})$. For any new image, we predict its region-level labels using $\dot{H}$ via Eq. (1) and Eq. (3), and then the image-level label vector is derived as well.

3. EXPERIMENTS

In this section, we evaluate the proposed method on the NUS-WIDE-SUB dataset [3, 8] which is a challenging collection of real-world web images from Flickr containing 18,325 images with 81 labels. We focus on evaluating our method when the number of labeled images is much smaller than that of unlabeled ones, and randomly split the dataset into the 10% subset for training and the 90% subset for testing. Since the user-provided tags is noisy, the training images are precisely re-annotated at first. We use Felzenswalb's graph cut algorithm [5] to segment each image into several regions. Four kinds of visual features are extracted for each region: 1) 14-dim color feature including mean RGB, HSV conversion, HUE histogram and SAT histogram; 2) 30-dim texture feature including LM-filter mean response [7] and LM-filter response histogram; 3) 8-dim geometric feature encoding the position and size information of the segment; 4) 500-dim BoW histogram. Based on the above visual features, the composite distance between image regions is computed using JEC[10]. Employing Affinity Propagation algorithm [6], we cluster the regions from the training images and obtain the region-exemplars.

Method	Training%	Precision	Recall
ML-LGC	50%	0.28	0.29
CNMF	50%	0.29	0.31
SMSE	50%	0.32	0.32
MISSL	50%	0.27	0.33
MEG	50%	**0.35**	0.37
Ours	**10%**	0.31	**0.41**

Table 1: Performance comparisons of different automatic annotation methods on NUS-WIDE-SUB.

We compare the proposed method with the state-of-the-art algorithms: 1) ML-LGC[16], 2) CNMF[9], 3)SMSE[1], 4)MISSL[11], 5)MEG[8]. In the label prediction task, we aim at predicting the labels of the testing images and use the provided 81 semantic concepts as the ground truth annotations for evaluation. We calculate the average precision and average recall to measure the performance. Table 1 gives the performance comparisons of the image annotation algorithms. The results of the the state-of-the-art are reported in [8]. From the result we can clearly see that our algorithm is better than or comparable with the state-of-the-art algorithms even though much less training examples are required than the state-of-the-art, which demonstrates the

Image	Before	After
	Road, sunset, sun, plane, airport	Plane, **sky**, sun, sunset, **clouds**, road, **vehicle**, airport
	Sky, lake, tree, water, grass, leaf, glacier	Tree, lake, **boat**, **mountain**, grass, sky, water
	House, tower, window, buildings, sky	House, buildings, sky, **tree**, **grass**, **garden**, **leaf**

Figure 2: The refined label rank lists are obtained according to the learned label confidence vector.

Figure 3: Region-level labeling results of our method for some images from NUS-WIDE-SUB dataset.

effectiveness of strategy to sufficiently leverage various contextual relations. Figure 2 gives the refined label rank lists obtained by the learned label confidence vector in comparison with the initial labels provided by the users. Moreover, the proposed algorithm is also suitable for the task of label-to-region assignment. Since the NUS-WIDE-SUB dataset has no ground truth for this task, we merely give some results in Figure 3. The results demonstrate that by learning the label confidences for the region-exemplars, our method can effectively estimate the label confidences for each image region, then the tasks of predicting both image-level and region-level labels are accomplished simultaneously.

4. CONCLUSIONS

In this paper a novel method is proposed to perform automatic image annotation with weakly-labeled image dataset. Clustering technique is employed to learn a small number of region-exemplars, and the labels for each image region are predicted as a locally weighted average of the labels on exemplars. By investigating the label confidence matrix for the region exemplars from *column* and *row* pictures, we sufficiently leverage the consistency of similarities between samples in the visual feature space and semantic label space, the correlations among the semantic concepts, and the associations between image-level and region-level labels, which are of significance to the performance of image annotation.

5. ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their helpful comments. This work was supported in part by the 973 Program (No.2010CB327906), the NSF of China (No.60903077 and No.60873178), and the STCSM's innovation program(No. 10511500703).

6. REFERENCES

[1] G. Chen, Y. Song, F. Wang, and C. Zhang. Semi-supervised multi-label learning by solving a sylvester equation. In *SDM*, 2008.

[2] L. Chen, D. Xu, I. W. Tsang, and J. Luo. Tag-based web photo retrieval improved by batch mode re-tagging. In *CVPR*, 2010.

[3] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *CIVR*, 2009.

[4] R. Cilibrasi and P. Vitany. The google similarity distance. In *TKDE*, 2007.

[5] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004.

[6] B. Frey and D. Dueck. Clustering by passing messages between data points. *Science, vol. 315, 2007.*

[7] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, 2001.

[8] D. Liu, S. Yan, Y. Rui, and H.-J. Zhang. Unified tag analysis with multi-edge graph. In *ACM MM*, 2010.

[9] Y. Liu, R. Jin, and L. Yang. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *AAAI*, 2006.

[10] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV*, 2008.

[11] R. Rahmani and S. Goldman. Missl: Multiple-instance semi-supervised learning. In *ICML*, 2006.

[12] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science, vol. 290, 2000.*

[13] Y. Saad and M. H. Schultz. Gmrs: A generalized minimal residual algorithm for solving nonsymmetrci linear systems. In *SIAM JSSC.*, 1986(7).

[14] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua. Inferring semantic concepts from community contributed images and noisy tags. In *ACM MM*, 2009.

[15] Z.-J. Zha, T. Mei, J. Wang, Z. Wang, and X.-S. Hua. Graph-based semi-supervised learning with multiple labels. *J. Vis. Commun. Image R., 2009.*

[16] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In *NIPS*, 2003.

[17] G. Zhu, S. Yan, and Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *ACM MM*, 2010.