2012 IEEE International Conference on Multimedia and Expo

Noisy Tag Alignment with Image Regions

Yang Liu, Jing Liu, Zechao Li, Hanqing Lu
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
Beijing, China
{*liuyang6, jliu, zcli, luhq*}*@nlpr.ia.ac.cn*

***Abstract*—With the permeation of Web 2.0, large-scale user contributed images with tags are easily available on social websites. How to align these social tags with image regions is a challenging task while no additional human intervention is considered, but a valuable one since the alignment can provide more detailed image semantic information and improve the accuracy of image retrieval. To this end, we propose a large margin discriminative model for automatically locating unaligned and possibly noisy image-level tags to the corresponding regions, and the model is optimized using concave-convex procedure (CCCP). In the model, each image is considered as a bag of segmented regions, associated with a set of candidate labeling vectors. Each labeling vector encodes a possible label arrangement for the regions of an image. To make the size of admissible labels tractable, we adopt an effective strategy based on the consistency between visual similarity and semantic correlation to generate a more compact set of labeling vectors. Extensive experiments on MSRC and SAIAPR TC-12 databases have been conducted to demonstrate the encouraging performance of our method comparing with other baseline methods.**

***Keywords*-Image Region Annotation; Partially-supervised Learning**

I. Introduction

There are explosive photo sharing websites with large-scale image collections available online, such as Flickr [1], Picasa [2], and Zooomr [3]. These Web 2.0 websites allow users not only share their photos, but tag and comment their interested ones. How to manage and index the huge image resources efficiently is a challenging problem. Automatic image annotation is the basement of image index, search technologies and other applications. Traditional image-level annotation methods focus on assigning labels to entire images and can hardly deal with the diversity and variation of web image content. Different from image-level annotation, image region annotation by learning an explicit correspondence between image regions and semantic labels within an image is more valuable because it can provide more detailed, reliable image annotation results and facilitate image retrieval based on tagging information.

In most public image collections, tags are provided to indicate the semantics of whole images rather than individual regions. Thus, compared to the traditional image annotation problem, image region annotation is more difficult due to the lack of training sets with region-level ground-truth. Furthermore, the web images usually confront with noisy labels which significantly limit the performance of traditional image region annotation methods.

In this paper, we formulate the task of image region annotation as a partially-supervised learning problem where instances and labels come in the form of bags with associated candidate labeling vector sets. Each labeling vector encodes a possible combination of image-level labels for corresponding regions and only one of the candidate labeling vectors is the correct one. Since the size of all possible labeling vectors is a value about numbers of regions and tags, naive generation of the candidate labeling set becomes intractable.

To overcome the above issues, we formulate the partially-supervised learning problem as a large margin discriminative model, and propose an efficient method to obtain a compact candidate labeling set. Apparently, only a subset of all possible labeling vectors should be maintained. Our method is based on the assumption of the correlation consistency between visual similarity and semantic relevance for image regions. This notion can also be understood as that, visual similar regions often reflect similar semantic themes, the visual similarity can be used to approximate the extent of two regions have same labels. Based on this view, we present an explicit formulation about the correlation consistency to obtain reliable correspondences between image regions and tags. The correspondences ambiguities between regions and labels are greatly reduced. Exploring the obtained candidate labeling vectors as a partially supervision, a large margin classifier is trained and solved with the concave-convex procedure (CCCP). In addition, we consider the fact that tagging a region with a relevant tag to the ground-truth is more favorable than with an irrelevant tag, and adopt a tag-correlation-based soft loss function to quantize the influences from the relevant extent between the estimated tag and the ground-truth for each region. Figure 1 shows the flowchart of the proposed method.

The main contributions of this paper are summarized as follows.

- We formulate the task of tag alignment with image

[1] http://www.flickr.com
[2] http://picasa.google.com
[3] http://zooomr.com

978-0-7695-4711-4/12 $26.00 © 2012 IEEE
DOI 10.1109/ICME.2012.143

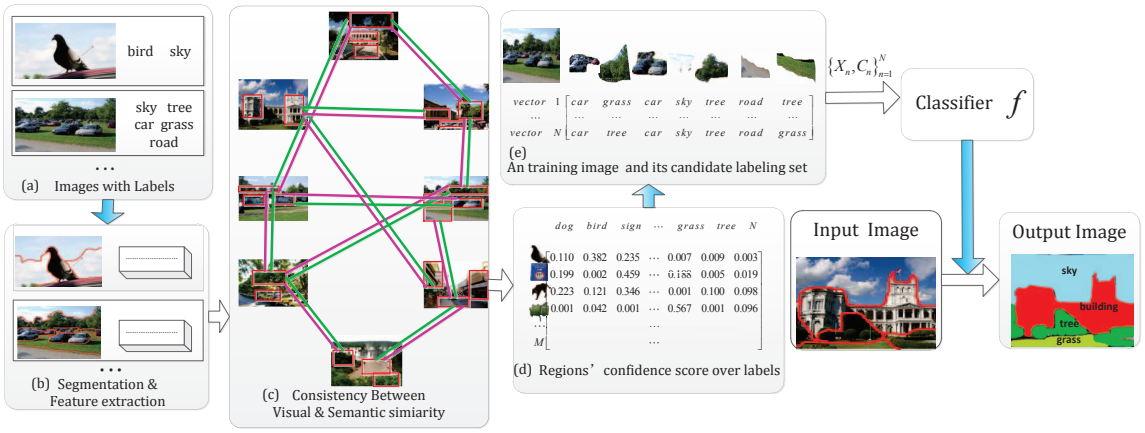

Figure 1. This is the flowchart of proposed algorithm. In figure (c) the edge marked with green represents visual similarity and semantic similarity is marked with pink.

regions as a partially-supervised learning problem.

- We present an automatic method to generate concise and compact candidate labeling sets.
- We adopt a tag-correlation-based soft loss function which is proved to be effective for image region annotation task.

The rest of this paper is organized as follows. Section 2 overviews some related work. Section 3 presents the partially-supervised learning framework. Section 4 elaborates an effective method to generate compact candidate labeling sets based on the correlation consistency. The experimental results are reported in Section 5, followed by the conclusions in Section 6.

II. Related work

In this section, some related works about image region annotation and partially-supervised learning methods are reviewed.

The classical methods about image region annotation can be roughly divided into two categories: supervised and partially-supervised. Several supervised solutions [1], [2], [3], [4] to this problem have been proposed using RF (Random Field) formulation and achieved good performance. The main difference between semi-supervised and partially-supervised is, semi-supervised learning means whether labeled or unlabeled samples are explored during the learning process, in which the labels are correct. Partially-supervised learning means that one sample can have several candidate labels but of which only one is the correct one. For partially-supervised methods, [5]. In [6], a bi-layer coding formulation was proposed for uncovering how an image or semantic region can be robustly reconstructed from the over-segmented image patches of an image set. The method in [7] is also well developed from sparse coding, it not only considers the intrinsic correlations among encoding regions and also integrates spatial correlations among basic regions.

For partially-supervised learning, in [8], the author formulates the learning problem as partially-supervised multiclass classification where each instance is labeled ambiguously with more than one label. Luo et al. [9] generalize these works to the cases where instances and possible labels come in the form of bags associated with candidate labeling sets, this work is the most relevant to ours. Since image region annotation is a more challenging task on either the scale of labeling instances and labels or the complex correlations among them, we make the following improvements : 1) an automatic method to generate the candidate labeling sets is proposed, instead of employing heuristic rules in [9]; 2) the image visual similarity and the tag relevance are exploited in our solution, while the independence assumption is given in [9]; 3) a more favorable loss function by considering the above correlations is presented in this paper which is more suitable for the image region annotation task.

III. Learning with Plausible Labels

A. Formulation

In our problem, the instances and candidate labels come in the form of bags with associated candidate labeling sets like $\{X_i, C_i\}_{i=1}^{N}$. X_i is a bag containing M_i instances $\{\mathbf{x}_{i,m}\}_{m=1}^{M_i}$, $\mathbf{x}_{i,m} \in R^d, \forall i = 1, ..., N, m = 1, ..., M_i$. Let $\mathcal{Y} = \{1, 2, ..., L\}$ denote the label space. $C_i = \{\mathbf{c}_{i,l}\}_{l=1}^{L_i}$ is the associated candidate labeling set of image X_i which includes L_i possible label vectors, each $\mathbf{c}_{i,l} \in \mathcal{Y}^{Mi}$ is a possible combination of M_i labels for the M_i instances in the i-th bag. The label vector can encode the relationships between instances and their labels explicitly. Considering an heuristic way, given an image containing two regions "sky" and "grass", prior knowledge tells us that the two regions cannot be labeled with the same label because of their distinctions on visual similarities. The candidate labeling

set C can be generated like this: $\mathbf{c}_{i,1}$ ={"sky","grass"}, $\mathbf{c}_{i,2}$ ={"grass","sky"}. However, this way to encode the relationships is too simple and unreliable. In subsection 4.1, we will introduce our strategy to generate candidate labeling set in details.

At first, we define he compatibility of image X and label vector $\mathbf{y}$ as follows:

$$\begin{aligned} f(X,\mathbf{y};\mathbf{w}) &= \sum_{m=1}^{M} \mathbf{w}\cdot\phi(\mathbf{x}_m)\otimes\varphi(y_m) \\ &+\alpha\cdot\sum_{i=1}^{M-1}\sum_{j=i+1}^{M} V(\mathbf{x}_i,\mathbf{x}_i)\cdot T(y_i=y_j) \end{aligned} \tag{1}$$

where $X = \{\mathbf{x}_1,...,\mathbf{x}_M\}$ and $\mathbf{y} = \{y_1,y_2,...,y_M\}$ is the label vector. ϕ and φ are the feature and label space mapping [10], and $\otimes$ is the Kronecker product. $\mathbf{w}$ is the model parameter. $V(\mathbf{x}_i,\mathbf{x}_j)$ indicates the visual similarity of region $\mathbf{x}_i$ and $\mathbf{x}_j$. T is an indicator function , where $T=1$ means $y_i=y_j$, while $T=0$ means it does not. α is a weight parameter that indicates the importance of the second component against the first term and we empirically set $\alpha=1$.

The first component measures the total compatibility of mapping all image regions to their corresponding labels. The second component incorporates the pairwise relationships between image regions, which encourages the semantic cohesion of similar regions. A gaussian function with a radius parameter σ is used to compute visual similarity :

$$V(\mathbf{x}_i,\mathbf{x}_j) = exp(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|_2^2}{\sigma^2}), \tag{2}$$

$\|\cdot\|_2$ denotes the l_2-norm. σ is set as the median value of all pairwise l_2-norm between the regions.

Loss function $\Delta(\widehat{\mathbf{y}},\mathbf{y})$ is introduced to measure the distance of the prediction label vector $\widehat{\mathbf{y}}$ and the true label vector $\mathbf{y}$ for the bag . Then the model parameter $\mathbf{w}$ is learned by minimizing the average loss on the training set $\{X_i,C_i\}_{i=1}^N$. In the case of 0-1 loss, the loss function simply statistics the number of wrongly classified instances in a bag:

$$l_\Delta(\widehat{\mathbf{y}},\mathbf{y}) = \sum_{m=1}^{M}\Delta(\widehat{y}_m,y_m), \tag{3}$$

$\Delta(\widehat{y}_m,y_m)$ is 1 when $\widehat{y}_m=y_m$ else is 0. Because the real label vector $\mathbf{y}$ is ambiguous while the candidate labeling set C is available , the loss function should be redefined as:

$$l_\Delta(\widehat{\mathbf{y}},C) = \min_{\mathbf{c}'\in C} l_\Delta(\widehat{\mathbf{y}},\mathbf{c}') \tag{4}$$

Direct minimizing this loss is very hard so [9] uses an upper bound function to replace the ambiguous loss:

$$\begin{aligned} & l_{max}(X,C;\mathbf{w}) \\ = \; & |\max_{\bar{\mathbf{c}}\notin C}(l_\Delta^A(\bar{\mathbf{c}},\mathbf{c})+f(X,\bar{\mathbf{c}};\mathbf{w})) - \max_{\mathbf{c}\in C} f(X,\mathbf{c};\mathbf{w})| \end{aligned} \tag{5}$$

During test period, the label vector for a bag is predicted according to the rule:

$$\widehat{\mathbf{y}} = \arg\max_{\mathbf{y}\in C} f(X,\mathbf{y};\mathbf{w}). \tag{6}$$

B. Learning Objective

We learn the model parameters $\mathbf{w}$ by solving the following optimization problem:

$$\min_{\mathbf{w}} \frac{\lambda}{2}\parallel\mathbf{w}\parallel_2^2 + \frac{1}{N}\sum_{i=1}^{N} l_{max}(X_i,C_i;\mathbf{w}), \tag{7}$$

Here, we empirically set $\lambda=\frac{1}{N}$ in the experiment.

An effective method to solve the optimization problem is the Concave-Convex Procedure(CCCP) [11] algorithm. The basic idea for CCCP algorithm is to substitute the concave part of the objective function with its first order Taylor expansion at the initial value at this iteration. At the r-th round, we use the following equation to replace the non-convex part $\max_{\mathbf{c}\in C_i} f(X_i,\mathbf{c};\mathbf{w})$ in the loss function:

$$\max_{\mathbf{c}\in C_i} f(X_i,\mathbf{c};\mathbf{w}^{(r)}) + (\mathbf{w}-\mathbf{w}^{(r)})\cdot\partial(\max_{\mathbf{c}\in C_i} f(X_i,\mathbf{c};\mathbf{w})) \tag{8}$$

later the stochastic subgradient descent algorithm is used to solve $\mathbf{w}^{(r)}$. More details about the solution please refer to [9].

C. Soft Loss function

As mentioned above, $\mathbf{w}$ is obtained by minimizing the average loss on the training data. In the case of 0-1 loss, the cumulative loss is exactly the number of misclassified instances in the bag. However, if the distribution of the class labels is nonuniform and interdependent, 0-1 loss is arbitrary to measure the real loss of misclassifying regions. Moreover, co-occurrence relationships among certain class labels should be considered to enhance the performance. We define a soft loss function to quantize the influences from the relevant extent between the estimated tag and the groundtruth for each region. Consider the label co-occurrence matrix U:

$$U(m,n) = \frac{corr(m,n)}{corr(m)+corr(n)-corr(m,n)}, \tag{9}$$

where $corr(m,n)$ is the number of images in the training set in which both label m and label n co-occur. $corr(m)$ is the number of images in the training set in which label m occurs. The higher $U(m,n)$ means label m and label n tend to appear together more frequently. The soft loss function is defined as following:

$$l_\Delta(\widehat{\mathbf{y}},\mathbf{y}) = \sum_{m=1}^{M}\Delta(\widehat{y}_m,y_m) = \sum_{m=1}^{M} e^{-U(\widehat{y_m},y_m)} \tag{10}$$

In the experiments, we adopt the soft loss function and compare it with the 0-1 loss function.

IV. Candidate Label Sets Generation

In this section, we will introduce how to generate candidate labeling sets, which is also the main contribution of this paper. Intuitively, only a subset of all possible label vectors should be exploited. Big scale of candidate labeling sets tend to include impossible label vectors which will degenerate training and cost too much time as well.

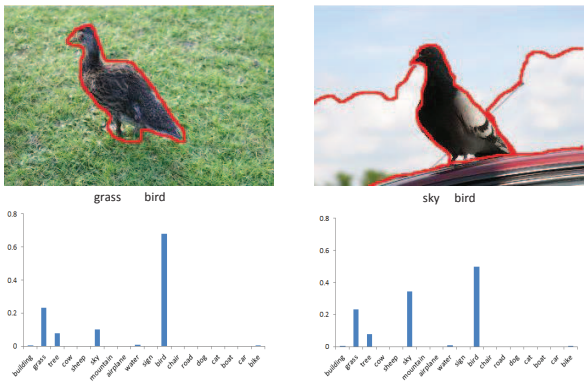

Figure 2. A simple example: visual similar regions may share similar confidence distribution over labels.

The key assumption is that similar regions usually tend to have large overlap in their semantic labels. Figure 2 illustrates this notion. Assume the region collections containing K regions $\{\mathbf{x}_1, ..., \mathbf{x}_K\}$. $\mathcal{Y} = \{1, 2, ..., L\}$ is the label space containing L labels. Denote R as a $K \times L$ matrix and R_{ij} means the confidence score of region $\mathbf{x}_i$ can have label j. Each row vector $\mathbf{r}_i$ of R can be seen as the confidence score distribution over labels of region $\mathbf{x}_i$.

Firstly, we still use Equation 2 to compute the $K \times K$ symmetric visual similarity matrix V, but the difference is, we only compute the region-pairs which one region is the k-nearest neighbors of another region. k is set as 100 for all the experiments.

Secondly, the semantic similarity of two regions can be measured with each pair dot product of $\mathbf{r}_i\mathbf{r}_j^T$. However, the correlation of labels should be leveraged in the approach. Here, Equation 9 is employed to represent the relationship among labels. The semantic similarity can be redefined as $\mathbf{r}_i U \mathbf{r}_j^T$.

Based on the assumptions of visual and semantic consistency, $V_{ij} \approx \mathbf{r}_i U \mathbf{r}_j^T$. We can get the following formulation:

$$\begin{aligned} &\min_R \sum_{i,j=1}^{K} (V_{ij} - \sum_{k,l=1}^{L} R_{ik} U_{kl} R_{jl})^2 \\ \Rightarrow &\min_R \parallel (V - RSR^T) \parallel_F^2 \\ &s.t. \quad R_{jl} \geq 0, \quad j = 1, 2, ..., K, \quad l = 1, 2, ..., L. \end{aligned} \tag{11}$$

We use the non-negative matrix factorization method [12] to solve this optimization.

After obtaining R, the candidate labeling set for each training image can be generated according to the confidence matrix. For each region, top T candidate labels are selected to do permutations and combinations and then generate the candidate labeling set. By this way, we can get the candidate labeling sets $\{C_i\}_{i=1}^N$ for the classifier f introduced in subsection 3.1.

V. Experiments

In this section, to validate the effectiveness of the proposed algorithm, we conduct extensive experiments on two public image datasets MSRC-350 [6] and SAIAPR TC-12 [13]. We compare our results with other existing baseline image region annotation algorithms.

A. Experimental Settings

1) Datasets: We use two public available datasets, MSRC-350 and SAIAPR TC-12 to evaluate our method. MSRC-350 is comprised of 350 images with 18 labels. SAIAPR TC-12 contains 40 subsets and we use the same subset as [7] which contains 251 images with 90 tags. Both of them are labeled with pixel-level groundtruth.

MSRC dataset does not provide the segmentation masks. Here, we use Ncut [14] algorithm to segment each image into several regions. For SAIAPR TC-12 dataset, the segmentation masks of images are provided. For both datasets, we extract Sift [15] features and use the Bag-of-Words(BOW) representations.

All the images of both datasets are used to train and we test on the same images respectively.

2) Comparing Schemes: We compare our proposed algorithm against the following state-of-the-art algorithms for image region annotation task.

- Bi-layer Sparse Coding proposed in [6].
- JSGSC proposed in [7].
- Multi-edge Graph proposed in [5].
- At last, we also implement the classical KNN algorithm to compare with our models. k is empirically set to 50 and 100.

We evaluate the image region annotation performance in two quantitative ways including pixel-level accuracy and region-level accuracy which respectively measures the percentage of pixels and regions with agreement between the assigned label and groundtruth. For MSRC dataset, the Ncut segmentation cannot generate the same segmentation with the manual groundtruth, so we assign each region a dominant label to be its groundtruth label.

B. Experiment Results

To evaluate the proposed image region annotation method, we conduct five experiments and illustrate the results in order.

The first experiment is designed to exploit the effects on the performance of different parameter settings of our model

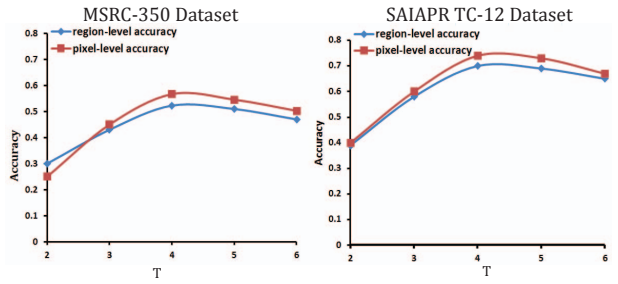

Figure 3. Effects of parameter T in our VSC candidate labeling sets generation method.

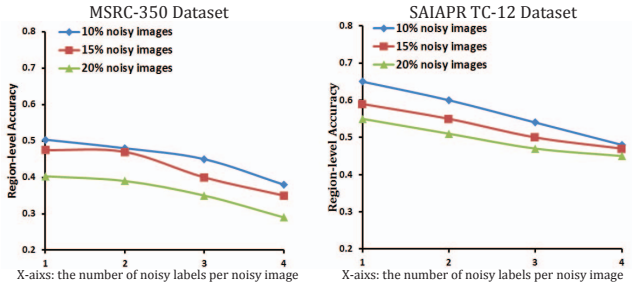

Figure 4. Effects of noisy labels of our algorithm on both datasets.

to get the best result. There are one adjustable parameters in our model: threshold T to decide how many candidate labels should be selected for each region to generate the candidate labeling sets. We choose T under two constraints, one is to include the groundtruth label and exclude the impossible labels of each instance as much as possible; another one is to control the scale of the set . We need to keep balance between the two norms. We set the same range $T \in \{2, 3, 4, 5, 6\}$ for both datasets. The results for effects of different parameter T are illustrated in Figure 3. We can find that the accuracies do not increase monotonically with T increases on both datasets. For both datasets, the best T is 4 which is in the middle range means that the most accurate candidate label vectors are included in the sets and impossible label vectors are excluded. After the peak point, with the increase of T, the accuracy is decrease slightly which means noisy label vectors are brought in.

In the second experiment, we verify the effectiveness of our VSC (Visual and Semantic Consistency) candidate labeling set generation method comparing with another heuristic strategy. The strategy is like this: for each region, we assign its groundtruth to itself and choose several labels randomly from the remaining labels set to do permutations and combinations then generate the candidate labeling sets. For fairness, in the first experiment we get the best performance when $T = 4$, so here we randomly choose 3 labels from remaining labels for each region. We name this baseline method as "heuristic method". Table I and Table II show the results using both candidate labeling sets generation methods. We can see from the tables: our candidate labeling set generation method achieves much higher performance than the heuristic method. The reason is, our method based on the consistency between visual similarity and semantic relevance can incorporate the prior knowledge of training data in an implicit manner and can reduce the correspondence ambiguities of regions and labels greatly.

Table I
THE PIXEL-LEVEL ACCURACY OF DIFFERENT CANDIDATE LABELING SETS GENERATION METHODS ON BOTH DATASETS.

method	MSRC-350	SAIAPR TC-12
$Heuristic$	0.58	0.38
VSC	**0.74**	**0.55**

Table II
THE REGION-LEVEL ACCURACY OF DIFFERENT CANDIDATE LABELING SETS GENERATION METHODS ON TWO PUBLIC DATASETS.

method	MSRC-350	SAIAPR TC-12
$Heuristic$	0.55	0.34
VSC	**0.70**	**0.53**

In the third experiment, we evaluate the performance of our method comparing with the state-of-art image region annotation methods. Table III and IV shows the results. Because the comparing baseline methods adopt different evaluations of performance so we directly report the best results in their papers. Several observations can be obtained. Firstly, proposed method outperforms all the other baseline methods except JSGSC on MSRC-350 dataset. It is worth noting that JSGSC uses much groundtruth spatial information while our method does not. Secondly, we can believe that our VSC method provides a good partially supervision to training.

Table III
THE PIXEL-LEVEL ACCURACY OF DIFFERENT ALGORITHMS ON TWO PUBLIC DATASETS. FOR KNN ALGORITHM, WE SET K=50 AND 100 EMPIRICALLY.

method	MSRC-350	SAIAPR TC-12
$kNN(k = 50)$	0.45	0.22
$kNN(k = 100)$	0.37	0.21
$Bi - layer$ [6]	0.63	-
$Multi - Edge\ Graph$ [5]	0.73	-
$ours$	**0.74**	**0.55**

In the fourth experiment, we discuss the effects of soft loss function comparing with 0-1 loss function. Table V and Table VI shows the results on both datasets. Only region-level accuracy is reported in this experiment. The region-level accuracies of both loss functions are similar. However, the soft loss improves the mean per class accuracy a little on both datasets which approves the tag-correlation-based loss is reasonable for the image region annotation task.

Table IV
THE REGION-LEVEL ACCURACY OF DIFFERENT ALGORITHMS ON TWO PUBLIC DATASETS. FOR KNN ALGORITHM, WE SET K=50 AND 100 EMPIRICALLY.

method	MSRC-350	SAIAPR TC-12
$kNN(k=50)$	0.42	0.19
$kNN(k=100)$	0.39	0.18
$Bi-layer$ [6]	0.64	0.39
$JSGSC$ [7]	**0.77**	0.49
$ours$	0.70	**0.53**

Table V
THE REGION-LEVEL ACCURACY AND MEAN PER CLASS ACCURACY OF SOFT LOSS AND 0-1 LOSS ON MSRC-350 DATASET.

method	overall	mean per-class
Δ_{soft}	**0.70**	**0.58**
$\Delta_{0/1}$	0.69	0.56

In the fifth experiment, we illustrate the results on artificial noisy datasets. The artificial training sets are conducted as follows: firstly, we randomly choose 10, 15, 20 percent of the training images as "noisy images" to form artificial noisy datasets. For each noisy image in the artificial dataset, except its own labels we respectively choose 1,2,3,4 labels from the remaining labels set as the additional noisy labels. Then we adopt the best parameter setting $T = 4$ as in the previous experiment has been verified. The region-level results on both datasets are shown in Figure 4. With the growing numbers of noisy images and noisy labels, the accuracy of our algorithm decreases at a tolerable speed. There are two reasons. Firstly, the noisy labels may always disobey the prior knowledge of visual and semantic consistency then gets a relatively low confidence score. Secondly, the framework of learning from candidate labeling sets itself has ability of tolerating noisy labels.

VI. CONCLUSION

In this paper, a novel image region annotation method based on partially-supervised learning framework is proposed. Based on the assumption of consistency between low-level visual features and high-level semantic relevances, the correspondence ambiguities between regions and labels are greatly reduced then accurate and appropriate size candidate labeling sets are created. We incorporate tag correlations into the soft loss function which is proved to be effective in boosting the performance. Extensive experiments show our method is advantageous over many state-of-the-art image region annotations methods.

VII. ACKNOWLEDGMENT

This work was supported by 973 Program (Project No. 2010CB327905) and the National Natural Science Foundation of China (Grant No. 60903146, 60835002).

Table VI
THE REGION-LEVEL ACCURACY AND MEAN PER CLASS ACCURACY OF SOFT LOSS AND 0-1 LOSS ON SAIAPR TC-12 DATASET.

method	overall	mean per-class
Δ_{soft}	**0.53**	**0.44**
$\Delta_{0/1}$	0.52	0.43

REFERENCES

[1] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *ICCV*, 2007.

[2] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma, "Annotation via multi-graph learning," in *Pattern Recognition*, 2009, pp. 218–228.

[3] 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 Multimedia*, 2007.

[4] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *IJCV*, vol. 81, pp. 2–23, 2009.

[5] D. Liu, S. Yan, Y. Rui, and H.-J. Zhang, "Unified tag analysis with multi-edge graph," in *ACM Multimedia*, 2010, pp. 25–34.

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

[7] Y. Yang, Y. Yang, Z. Huang, H. T. Shen, and F. Nie, "Tag localization with spatial correlations and joint group sparsity," in *CVPR*, 2011, pp. 881–888.

[8] T. Cour, B. Sapp, C. Jordan, and B. Taskar, "Learning from ambiguously labeled images," in *CVPR*, 2009, pp. 919–926.

[9] L. Jie and F. Orabona, "Learning from candidate labeling sets," in *NIPS*, 2010, pp. 1504–1512.

[10] I.Tsochantaridis, T.Joachims, T.Hofmann, and Y.Altun, "Large margin methods for structured and interdependent output variables." *JMLR*, vol. 6, pp. 1453–1484, 2005.

[11] A.Yuille and A.Rangarajan, "The concave-convex procedure," *Neural Computation*, vol. 15, pp. 915–936, 2003.

[12] H. S. Daniel D.Lee, "Algorithms for non-negative matrix factorization," in *NIPS*, 2000.

[13] H. J. Escalante, C. A. Hernandez, J. A. Gonzalez, and A. Loez-Loez., "The segmented and annotated iapr tc-12 benchmark." in *CVIU*, vol. 114, 2010, pp. 419–428.

[14] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE TPAMI*, vol. 22, no. 8, pp. 888 –905, 2000.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, pp. 91–110, 2004.