Correlative Multi-Label Multi-Instance Image Annotation

Xiangyang Xue[1], Wei Zhang[1]*, Jie Zhang[1], Bin Wu[1], Jianping Fan[2], Yao Lu[1]
[1]School of Computer Science, Fudan University, Shanghai, China
[2]Department of Computer Science, UNC-Charlotte, NC28223, USA
{ xyxue,weizh,072021164,wubin}@fudan.edu.cn jfan@uncc.edu yaolu@fudan.edu.cn

Abstract

In this paper, each image is viewed as a bag of local regions, as well as it is investigated globally. A novel method is developed for achieving multi-label multi-instance image annotation, where image-level (bag-level) labels and region-level (instance-level) labels are both obtained. The associations between semantic concepts and visual features are mined both at the image level and at the region level. Inter-label correlations are captured by a co-occurence matrix of concept pairs. The cross-level label coherence encodes the consistency between the labels at the image level and the labels at the region level. The associations between visual features and semantic concepts, the correlations among the multiple labels, and the cross-level label coherence are sufficiently leveraged to improve annotation performance. Structural max-margin technique is used to formulate the proposed model and multiple interrelated classifiers are learned jointly. To leverage the available image-level labeled samples for the model training, the region-level label identification on the training set is firstly accomplished by building the correspondences between the multiple bag-level labels and the image regions. JEC distance based kernels are employed to measure the similarities both between images and between regions. Experimental results on real image datasets MSRC and Corel demonstrate the effectiveness of our method.

1. Introduction

Automatic image annotation has become more and more attractive when digital images grow exponentially [17, 6]. Multiple semantic concepts (labels) may occur simultaneously in an image, e.g., $\{sheep\&grass\}$, $\{mountain\&sky\&water\}$, and so on. Many algorithms have been developed to enable multi-label learning recently [18, 4, 29, 9, 8, 19, 26]. On the other hand, each individual label of one image is actually related to local regions rather than the global image; each region can be viewed as an instance and the image is just a bag of instances. Based on this observation, some researchers also consider image classification as a multi-instance learning task [15, 11, 31, 28].

In the existing joint multi-label multi-instance learning framework, the correlations among multiple labels and the associations between the visual features and the semantic concepts are not exploited sufficiently to improve the annotation performance. For example, [15] degenerates the multi-label multi-instance problem to several multi-instance single-label problem, and the dependency between the image labels is not modeled. Although [28] considers the label-label correlations, these correlations are not related with the image features. It has been shown that the tendency of the semantic concepts to co-occur is usually not independent of the image visual features [8].

In the computer vision community, it is an interesting topic to assign labels to regions within an image. [13, 1] conduct object recognition by learning an explicit detection model for each label; however, they are not applicable in many real-world applications because it is difficult to collect the large scale labeled image regions per class. [10] performs image region labeling by Multiscale CRF which models spatial relations between labels; however, the label-label semantic correlation is not yet captured. [16] proposes a unified formulation to label-to-region assignment as well as automatic labeling; however, the label-label correlation is not effectively leveraged, either.

In this paper, a novel method for Correlative Multi-Label Multi-Instance image annotation is proposed. The input image is segmented and can be viewed as a bag of instances (regions). The global visual features of the entire image and the local features of the regions are extracted to capture coarse and fine patterns, respectively. For the training images, image-level labels are provided while region-level labels are unknown. To leverage the available image-level labeled samples for the model training, the region-level label identification on the training set is firstly accomplished by building the correspondences between the mul-

*Corresponding Author: weizh@fudan.edu.cn

978-1-4577-1102-2/11/$26.00 ©2011 IEEE

tiple image-level labels and the regions. For each test image, both image-level labels and region-level labels can be obtained in a single framework by realizing the cross-level label propagation. Inter-label correlations are captured by a co-occurence matrix of concept pairs. Structural max-margin technique is used to formulate the proposed model and multiple interrelated classifiers are learned jointly.

The proposed method takes an interesting formulation that tries to combine various contextual relations in a single framework. It can be investigated from different perspectives: i) the associations between semantic concepts and visual features (global and local); ii) the correlations among the multiple labels; iii) the cross-level label coherence between the labels at the image level and the labels at the region level.

The rest of this paper is organized as follows: In Section 2, we formulate the proposed model. Model learning and inference are given in Section 3. Experimental results on MSRC and Corel image datasets are shown in Section 4. Finally, we conclude this paper in Section 5.

2. The Proposed Model

Let I denote the global visual feature vector extracted from an entire image; suppose that the image is partitioned into m regions through the existing segmentation algorithms [2, 5, 22, 20], and let $\{R_r\}_{r=1}^{m}$ denote m local visual feature vectors extracted from the corresponding image regions, where the number of regions, m, might vary across different images. Let $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_c] \in \{1, -1\}^c$ denote $c-$dimensional concept label vector of an image, where $\mathbf{y}_l = 1(l = 1, \ldots, c)$ indicates that the image belongs to the $l-$th concept, and -1 otherwise. Let $\mathbf{h}_r = [\mathbf{h}_r(1), \ldots, \mathbf{h}_r(c)] \in \{1, -1\}^c$ denote the concept label vector of the $r-$th region in the image, where each entry $\mathbf{h}_r(l) \in \{1, -1\}$ likewise indicates the membership of this region on the $l-$th concept.

Suppose that n image-level labeled samples are available $\{(I^1, \{R_r^1\}_{r=1}^{m_1}, \mathbf{y}^1), \ldots, (I^n, \{R_r^n\}_{r=1}^{m_n}, \mathbf{y}^n)\}$, where the ith image includes m_i instances $(i = 1, \ldots, n)$, I^i denotes the global feature vector for the ith image, $\{R_r^i\}_{r=1}^{m_i}$ denotes a bag of regional feature vectors for the ith image, and $\mathbf{y}^i$ is the associated image-level multi-label vector. The region-level label vectors $\{\mathbf{h}_r^i\}_{r=1}^{m_i}$ are unknown. The task is to learn a discriminative model $f(I, \mathbf{y}, \{R_r\}, \{\mathbf{h}_r\})$ from the available image-level labeled samples. Then, for any new image, the associated image-level and region-level labels can be simultaneously inferred:

$$(\hat{\mathbf{y}}, \{\hat{\mathbf{h}}_r\}) = arg \max_{(\mathbf{y}, \{\mathbf{h}_r\})} f(I, \mathbf{y}, \{R_r\}, \{\mathbf{h}_r\}) \tag{1}$$

In real world applications, multiple labels do not appear independently but occur correlatively and usually interact with each other at semantic space [30]. For example, labels $sheep$ and $grass$ often co-occur, which can be considered as a pair of concepts with significant correlation; on the contrary, $sheep$ and $office$ seldom co-occur in the same image. The inter-label correlation matrix is constructed to characterize the interdependency between concepts and helps to learn the inter-related classifiers in the feature space. We can construct the inter-label correlation matrix by using the available image-level labeled samples. Suppose two labels l and t, and define empirical conditional probabilities $p(t|l) = \frac{\sum_{i=1}^{n}(\mathbf{y}_l^i+1)(\mathbf{y}_t^i+1)/4}{\sum_{i=1}^{n}(\mathbf{y}_l^i+1)/2}$ and $p(l|t) = \frac{\sum_{i=1}^{n}(\mathbf{y}_l^i+1)(\mathbf{y}_t^i+1)/4}{\sum_{i=1}^{n}(\mathbf{y}_t^i+1)/2}$. Denote the harmonic mean $a_{lt} = \frac{p(t|l)p(l|t)}{[p(t|l)+p(l|t)]/2}$, and define the inter-label correlation matrix as $A = [a_{lt}]_{c \times c}$.

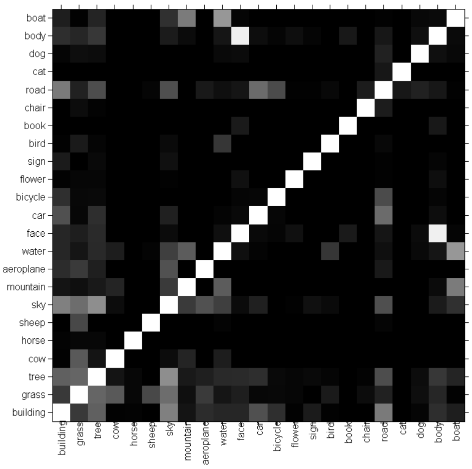

Figure 1. The inter-label correlation matrix based on the the harmonic mean of empirical conditional probabilities illustrates the interdependency between 23 concepts on the MSRC dataset. The brighter the block is, the stronger the correlation between labels exists.

As an illustration, Figure. 1 shows the inter-label correlation matrix illustrating the interdependency between 23 concepts on the MSRC (MicroSoft Research Cambridge) image dataset. The brighter the block is, the stronger the correlation between labels exists. The dark blocks indicate the concept pairs without correlations on the MSRC dataset. It is natural that the pairs of the same concepts correspond to the bright block. Among those pairs of the different concepts, we can find that the concept pair $face$ and $body$ have the most significant correlation. As for the concept $road$, those concepts $building$, $tree$, sky, car,$bicycle$ have strong correlations with it. It should be pointed out that there seems to be weak correlation between the concepts $bird$ and sky; the reason for such surprising observation is that in the MSRC dataset there are few images showing the scene $a\ bird\ flying\ in\ the\ sky$; on the contrary, there are quite a few images showing $a\ bird\ over\ water$ or $grass$.

Based on the inter-label correlation matrix, we formulate the Correlative Multi-Label Multi-Instance Model for image annotation using the structural max-margin technique such that various contextual relations are incorporated in a single framework:

$$\begin{aligned}&f(I,\mathbf{y},\{R_r\},\{\mathbf{h}_r\})\\=&\eta_1\sum_{l=1}^{c}\mathbf{y}_l(\mathbf{u}_l^\top\varphi(I)+b_l)+\eta_2\sum_{l=1}^{c}\sum_{r=1}^{m}\mathbf{h}_r(l)(\mathbf{v}_l^\top\phi(R_r)+b_l')\\&+\eta_3\sum_{l=1}^{c}\sum_{t\in\mathcal{N}_l}\mathbf{y}_l\mathbf{y}_t\mathbf{w}_{lt}^\top\varphi(I)-\eta_4\sum_{l=1}^{c}|\mathbf{y}_l-max_r\mathbf{h}_r(l)|\end{aligned} \tag{2}$$

where $\mathbf{u}_l$, $\mathbf{v}_l$ and $\mathbf{w}_{lt}$ are the parameter vectors to be learned, which are associated with the label l and the label pair (l,t) , respectively; b_l and b_l' are the bias parameters; $\varphi(I)$ and $\phi(R_r)$ are the (nonlinear) functions mapping the input global features of the entire image and the local features of the image region to the kernel spaces, respectively; $\eta_1,\eta_2,\eta_3,\eta_4(>0)$ are controlling parameters; $\mathcal{N}_l=\{t|t\neq l\wedge a_{lt}>T_0\}$ denotes the set of all concepts related with the concept l (T_0 is a predefined threshold).

The first part of Eq. (2) encodes the associations between the image-level labels and the global visual features; the second part of Eq. (2) models the associations between the region-level labels and the local visual features; the third part of Eq. (2) captures the inter-label correlations dependent on the image features; the last part of Eq. (2) measures the coherence between image-level labels and region-level labels. In multi-label multi-instance learning, for a specific label, one bag (image) is tagged positive if there is at least one instance (region) with the concerned label; otherwise the bag is tagged negative. We can minimize the loss function $\sum_{l=1}^{c}|\mathbf{y}_l-max_r\mathbf{h}_r(l)|$ to maximize the coherence between image-level labels and region-level labels, and thus realize the cross-level label propagation.

Figure. 2 illustrates the framework of the proposed correlative multi-label multi-instance model for image annotation: Each image is investigated globally, as well as it is viewed as a bag of local regions, and the associations between semantic concepts and (global and local) visual features are mined both at the image level and at the region level; By constructing the inter-label correlation matrix, the interdependency between concepts are modeled in the label space and the knowledge on the co-occurrence of labels in the same image are captured; The cross-level label coherence encodes the consistency between the labels at the image level and the labels at the region level.

3. Model Learning and Inference

Based on the proposed discriminative model $f(I,\mathbf{y},\{R_r\},\{\mathbf{h}_r\})$, we can estimate the image-level (bag-level) labels in addition to the region-level (instance-level) labels for each image as $(\hat{\mathbf{y}},\{\hat{\mathbf{h}}_r\})=arg\max_{(\mathbf{y},\{\mathbf{h}_r\})}f(I,\mathbf{y},\{R_r\},\{\mathbf{h}_r\})$. We learn the optimal parameter vector by minimizing the empirical risk on the training images $\{(I^1,\{R_r^1\}_{r=1}^{m_1},\mathbf{y}^1),\ldots,(I^n,\{R_r^n\}_{r=1}^{m_n},\mathbf{y}^n)\}$:

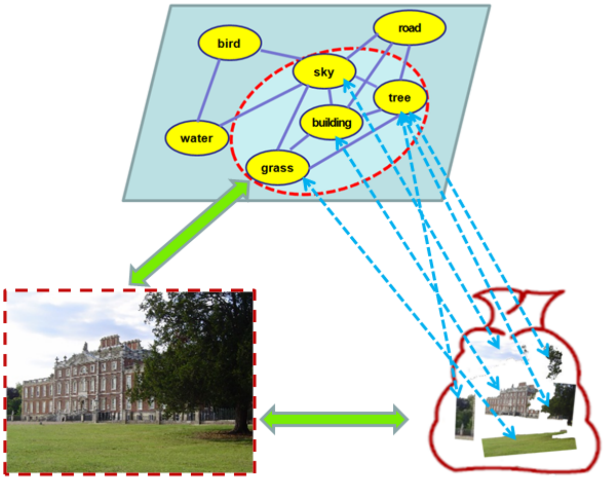

Figure 2. Our model for correlative multi-label multi-instance image annotation. Each image is described by global features as well as bag-of-regions. Various contextual relations are sufficiently leveraged in a single framework: concept-feature associations, inter-concept correlations, and cross-level label coherence.

$$\min\frac{1}{n}\sum_{i=1}^{n}\Delta(\mathbf{y}^i,\hat{\mathbf{y}}^i) \tag{3}$$

where $\Delta(\mathbf{y}^i,\hat{\mathbf{y}}^i)=\sum_{l=1}^{c}\mathbf{1}_{(\mathbf{y}_l^i\neq\hat{\mathbf{y}}_l^i)}$ counts the number of labels that are incorrectly predicted for the training image $\mathbf{x}^i$. Because

$$\begin{aligned}f(I^i,\hat{\mathbf{y}}^i,\{R_r^i\},\{\hat{\mathbf{h}}_r^i\})=&\max_{(\mathbf{y},\{\mathbf{h}_r\})}f(I^i,\mathbf{y},\{R_r^i\},\{\mathbf{h}_r\})\\\geq&\max_{\{\mathbf{h}_r'\}}f(I^i,\mathbf{y}^i,\{R_r^i\},\{\mathbf{h}_r'\})\end{aligned} \tag{4}$$

the upper bound of $\Delta(\mathbf{y}^i,\hat{\mathbf{y}}^i)$ can be derived as follows:

$$\begin{aligned}&\Delta(\mathbf{y}^i,\hat{\mathbf{y}}^i)\\\leq&\Delta(\mathbf{y}^i,\hat{\mathbf{y}}^i)+f(I^i,\hat{\mathbf{y}}^i,\{R_r^i\},\{\hat{\mathbf{h}}_r^i\})\\&-\max_{\{\mathbf{h}_r'\}}f(I^i,\mathbf{y}^i,\{R_r^i\},\{\mathbf{h}_r'\})\\\leq&\max_{(\mathbf{y},\{\mathbf{h}_r\})}\{\Delta(\mathbf{y}^i,\mathbf{y})+f(I^i,\mathbf{y},\{R_r^i\},\{\mathbf{h}_r\})\}\\&-\max_{\{\mathbf{h}_r'\}}f(I^i,\mathbf{y}^i,\{R_r^i\},\{\mathbf{h}_r'\})\end{aligned} \tag{5}$$

Therefore, given the image-level labeled training set, the proposed model can be learned by minimizing the following objective function:

$$\min_{\mathbf{u},\mathbf{v}} \quad \frac{1}{2}(\sum_{l=1}^{c}\|\mathbf{u}_l\|^2 + \sum_{l=1}^{c}\|\mathbf{v}_l\|^2 + \sum_{l=1}^{c}\sum_{t\in\mathcal{N}_l}\|\mathbf{w}_{lt}\|^2) \\ +\lambda\sum_{i=1}^{n}\Big(\max_{(\mathbf{y},\{\mathbf{h}_r\})}\{\Delta(\mathbf{y}^i,\mathbf{y}) + f(I^i,\mathbf{y},\{R_r^i\},\{\mathbf{h}_r\})\} \\ - \max_{\{\mathbf{h}'_r\}} f(I^i,\mathbf{y}^i,\{R_r^i\},\{\mathbf{h}'_r\})\Big) \tag{6}$$

where the first part is for regularization and λ is a trade-off parameter. The objective function can also be expressed as:

$$\min_{\mathbf{u},\mathbf{v},\xi} \frac{1}{2}(\sum_{l=1}^{c}\|\mathbf{u}_l\|^2 + \sum_{l=1}^{c}\|\mathbf{v}_l\|^2 + \sum_{l=1}^{c}\sum_{t\in\mathcal{N}_l}\|\mathbf{w}_{lt}\|^2) + \lambda\sum_{i=1}^{n}\xi^i \\ s.t. \quad \max_{\{\mathbf{h}'_r\}} f(I^i,\mathbf{y}^i,\{R_r^i\},\{\mathbf{h}'_r\}) - \max_{\{\mathbf{h}_r\}} f(I^i,\mathbf{y},\{R_r^i\},\{\mathbf{h}_r\}) \\ \geq \Delta(\mathbf{y}^i,\mathbf{y}) - \xi^i \quad \forall i\in\{1,\dots,n\}, \forall\mathbf{y}\in\{1,-1\}^c \tag{7}$$

where ξ^i is the slack variable.

Note that $[\max_{\{\mathbf{h}'_r\}} f(I^i,\mathbf{y}^i,\{R_r^i\},\{\mathbf{h}'_r\}) - \max_{\{\mathbf{h}_r\}} f(I^i,\mathbf{y},\{R_r^i\},\{\mathbf{h}_r\})]$ can be viewed as the margin between the ground truth labels and the prediction at the image level, while the region-level labels $\{\mathbf{h}_r\}$ are treated as latent variables. The objective function in (7) actually takes the form of structural SVM with latent variables [24, 27]. However, the algorithms in [24, 27] maintain a working set of active constraints, which leads to complicated optimization problems [26]. Approximately, based on the design of the proposed model, we can divide the optimization problem into inter-related subproblems and then learn the model more efficiently.

3.1. Learning $\mathbf{u}_l$ and $\{\mathbf{w}_{lt}\}_{t\in\mathcal{N}_l}$

Using the available image-level labeled training samples, we can first learn the parameter vectors $\mathbf{u}_l$ and $\{\mathbf{w}_{lt}\}_{t\in\mathcal{N}_l}$ as follows:

$$\min_{\mathbf{u},\xi}\frac{1}{2}(\sum_{l=1}^{c}\|\mathbf{u}_l\|^2 + \sum_{l=1}^{c}\sum_{t\in\mathcal{N}_l}\|\mathbf{w}_{lt}\|^2) + \lambda\sum_{i=1}^{n}\xi^i \\ s.t. \quad f_I(I^i,\mathbf{y}^i) - f_I(I^i,\mathbf{y}) \geq \Delta(\mathbf{y}^i,\mathbf{y}) - \xi^i \\ \xi^i \geq 0 \\ \forall i\in\{1,\dots,n\}, \forall\mathbf{y}\in\{1,-1\}^c \tag{8}$$

where $f_I(I,\mathbf{y}) = \eta_1\sum_{l=1}^{c}\mathbf{y}_l(\mathbf{u}_l^\top\varphi(I) + b_l) + \eta_3\sum_{l=1}^{c}\sum_{t\in\mathcal{N}_l}\mathbf{y}_l\mathbf{y}_t\mathbf{w}_{lt}^\top\varphi(I)$ is the image-level submodel. There are $n\times 2^c$ constraints and the optimization problem is too complex to be solved directly. However, based on the linear property of $f_I(I,\mathbf{y})$, we factor the image-level submodel formulation as $f_I(I,\mathbf{y}) = \sum_{l=1}^{c} f_I^l(I,\mathbf{y}_l,\mathbf{y}_{\mathcal{N}_l})$, of which each item with respect to label l is as follows:

$$f_I^l(I,\mathbf{y}_l,\mathbf{y}_{\mathcal{N}_l}) = \eta_1\mathbf{y}_l(\mathbf{u}_l^\top\varphi(I) + b_l) + \eta_3\sum_{t\in\mathcal{N}_l}\mathbf{y}_l\mathbf{y}_t\mathbf{w}_{lt}^\top\varphi(I) \tag{9}$$

Like [23, 26, 30], the optimization can be performed over a single label variable while the rest are fixed, and the learning procedure is approximately decoupled into c inter-related subproblems. For each $l\in\{1,\dots,c\}$,

$$\min_{\mathbf{u},\xi}\frac{1}{2}(\|\mathbf{u}_l\|^2 + \sum_{t\in\mathcal{N}_l}\|\mathbf{w}_{lt}\|^2) + \lambda_l\sum_{i=1}^{n}\xi_l^i \\ s.t. \quad f_I^l(I^i,\mathbf{y}_l^i,\mathbf{y}_{\mathcal{N}_l}^i) - f_I^l(I^i,\mathbf{y}_l,\mathbf{y}_{\mathcal{N}_l}^i) \geq \mathbf{1}_{(\mathbf{y}_l^i\neq\mathbf{y}_l)} - \xi_l^i \\ \xi_l^i \geq 0 \\ \forall i\in\{1,\dots,n\}, \forall\mathbf{y}_l\in\{1,-1\} \tag{10}$$

where $f_I^l(I^i,\mathbf{y}_l^i,\mathbf{y}_{\mathcal{N}_l}^i)$ is the partial model score based on the observational features and the *ground truth* labels, while $f_I^l(I^i,\mathbf{y}_l,\mathbf{y}_{\mathcal{N}_l}^i)$ is the partial model score based on the observational features and the *almost true* labels. Since $\mathbf{y}_l,\mathbf{y}_l^i\in\{1,-1\}$, there are only two cases: either $\mathbf{y}_l = \mathbf{y}_l^i$ or $\mathbf{y}_l = -\mathbf{y}_l^i$. If $\mathbf{y}_l = \mathbf{y}_l^i$, the constraints in (10) always hold; so, we can only focus on the case $\mathbf{y}_l = -\mathbf{y}_l^i$ and the constraints in (10) can be further written as:

$$f_I^l(I^i,\mathbf{y}_l^i,\mathbf{y}_{\mathcal{N}_l}^i) - f_I^l(I^i,-\mathbf{y}_l^i,\mathbf{y}_{\mathcal{N}_l}^i) \geq 1 - \xi_l^i \\ \xi^i \geq 0 \\ \forall i\in\{1,\dots,n\} \tag{11}$$

In the decoupled formulation, the model parameter vector can be learned with ease. Although the model parameter sub-vectors are learned label by label, the correlations between labels are still be taken into account due to the second part of Eq. (9) which encodes the inter-label dependency; Now, there are only n constraints in the optimization problem (10) s.t. (11) for each l, which is similar to two-class SVM. The dual of the optimization problem is as follows:

$$\max_{\alpha_l^i}\sum_{i=1}^{n}\alpha_l^i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_l^i\alpha_l^j\mathbf{y}_l^i\mathbf{y}_l^j K_{ij} \\ s.t. \quad \sum_{i=1}^{n}\alpha_l^i\mathbf{y}_l^i = 0, \lambda_l \geq \alpha_l^i \geq 0, \forall i\in\{1,\dots,n\} \tag{12}$$

where α_l^i denotes the dual variable, and

$$K_{ij} = 4(\eta_1^2 + \eta_3^2\sum_{t\in\mathcal{N}_l}\mathbf{y}_t^i\mathbf{y}_t^j)\varphi^\top(I^i)\varphi(I^j) \tag{13}$$

Using kernel method, we define:

$$\varphi^\top(I^i)\varphi(I^j) = exp\{-\rho\mathbf{d}(I^i,I^j)\} \tag{14}$$

where $\mathbf{d}(I^i, I^j)$ is the distance between images and ρ is the radius parameter of the Gaussian function.

Thus, the primal variable vectors $\mathbf{u}_l, \mathbf{w}_{lt}$ can be computed from the dual variables: $\mathbf{u}_l = 2\eta_1 \sum_{i=1}^{n} \alpha_l^i \mathbf{y}_l^i \varphi(I^i)$, $\mathbf{w}_{lt} = 2\eta_3 \sum_{i=1}^{n} \alpha_l^i \mathbf{y}_l^i \mathbf{y}_t^i \varphi(I^i)$.

3.2. Learning $\mathbf{v}_l$

Let $f_R(\{R_r\}, \{\mathbf{h}_r\}) = \sum_{l=1}^{c} \sum_{r=1}^{m} \mathbf{h}_r(l)(\mathbf{v}_l^\top \phi(R_r) + b_l')$, which is the region-level submodel of Eq. (2). Because the labels of the training samples are available only at the image level, it is of significance to identify the exact correspondences between multiple labels and the image regions such that the label for each region is automatically determined and the region-level submodel can be effectively learned. Like the previous work [21], we accomplish the region-level label identification by the clustering technique.

For each label, we employ the affinity propagation algorithm [7] to cluster the image regions (in the set of positive images and in set of the negative images, respectively) using local visual features. The region clusters derived from the set of positive images can further be divided into two kinds: the positive region clusters (those regions should be associated with the current label) and the negative region clusters. The positive region cluster tends to be of large size because at least one region is positive per image such that more regions may share common visual properties for the the current label and they are grouped into the same cluster (positive cluster). The negative region clusters may have smaller sizes because the negative regions are from different classes. At the same time, all the region clusters derived from the set of negative images are negative clusters. The positive region clusters derived from the set of positive images should be far away from the negative region clusters derived from the set of negative images. Meanwhile, the negative region clusters derived from the set of positive images might be close to some negative region clusters derived from the set of negative images.

Therefore, the differences between the positive region clusters and the negative region clusters derived from the set of positive images can be investigated from two perspectives: either their similarities to the negative region clusters derived from the set of negative images or their sizes. Thus we can identify the positive region clusters from the negative ones, and the current label is treated as the true label for all the image regions in the positive clusters. We can choose all the positive regions for training. To avoid the imbalance between positive and negative training samples, we do not choose all the negative regions for training, but just choose the subset of negative regions derived from the set of negative images. Now the parameter vectors $\mathbf{v}_l$ can be learned as follows:

$$\begin{aligned}
&\min_{\mathbf{v},\xi} \frac{1}{2} \sum_{l=1}^{c} \|\mathbf{v}_l\|^2 + \lambda \sum_{i=1}^{n} \xi^i \\
&s.t. f_R(\{R_r^i\}, \{\mathbf{h}_r^i\}) - f_R(\{R_r^i\}, \{\mathbf{h}_r\}) \geq \Delta(\{\mathbf{h}_r^i\}, \{\mathbf{h}_r\}) - \xi^i \\
&\xi^i \geq 0 \\
&\forall i \in \{1, \ldots, n\}, \forall \mathbf{h}_r \in \{1, -1\}^c
\end{aligned} \tag{15}$$

where $\Delta(\{\mathbf{h}_r^i\}, \{\mathbf{h}_r\}) = \sum_{l=1}^{c} \sum_{r=1}^{m_i} \mathbf{1}_{\left(\mathbf{h}_r^i(l) \neq \mathbf{h}_r(l)\right)}$. Similar to (8), the optimization can be approximately decoupled into l subproblems such that the parameter vectors are learned more effectively.

$$\begin{aligned}
&\min_{\mathbf{v},\xi} \frac{1}{2} \|\mathbf{v}_l\|^2 + \lambda \sum_{i=1}^{n} \sum_{r=1}^{m_i} \xi_r^i \\
&s.t. \quad \mathbf{h}_r^i(l)(\mathbf{v}_l^\top \phi(R_r^i) + b') \geq 1 - \xi_r^i, \quad \xi_r^i \geq 0 \\
&\forall i \in \{1, \ldots, n\}, \forall r \in \{1, \ldots, m_i\}
\end{aligned} \tag{16}$$

Likewise, we should compute $\phi^\top(R_r^i)\phi(R_p^j)$ in dual optimization problems. Again, using kernel method, we define:

$$\phi^\top(R_r^i)\phi(R_p^j) = exp\{-\varrho \mathbf{d}(R_r^i, R_p^j)\} \tag{17}$$

where $\mathbf{d}(R_r^i, R_p^j)$ is the distance between image regions based on the local visual features and ϱ is the radius parameter of the Gaussian function.

3.3. Inference

For any new image, the inference problem is to find the optimal label configuration $(\hat{\mathbf{y}}, \{\hat{\mathbf{h}}_r\}) = arg\max_{(\mathbf{y}, \{\mathbf{h}_r\})} f(I, \mathbf{y}, \{R_r\}, \{\mathbf{h}_r\})$. The size of multi-label space is exponential to the number of classes, and it is intractable to enumerate all possible label configurations to find the best one. Therefore we employ the iterative approach to approximate the optimal label configuration. First, we can initialize a multi-label configuration by two steps: i) For each image region, we estimate its label by the support vector machines derived from Eq. (16) such that the second item in Eq. (2) is maximized; ii) The image-level labels are initialized using the rule that one image is tagged positive if there is at least one region with the concerned label (otherwise, the image is tagged negative) such that the last item in Eq. (2), i.e., $\sum_{l=1}^{c} |\mathbf{y}_l - max_r \mathbf{h}_r(l)|$, is minimized. Then, based on the initial label configuration, we employ an approximate inference technique called $Iterated\ Conditional\ Modes$ (ICM) [25] to estimate the optimal image-level multi-label vector such that the sum of the first and third items in Eq. (2) is maximized: In each iteration, given $\mathbf{y}_{\mathcal{N}_l}$, we sequentially update $\mathbf{y}_l$ using the law:

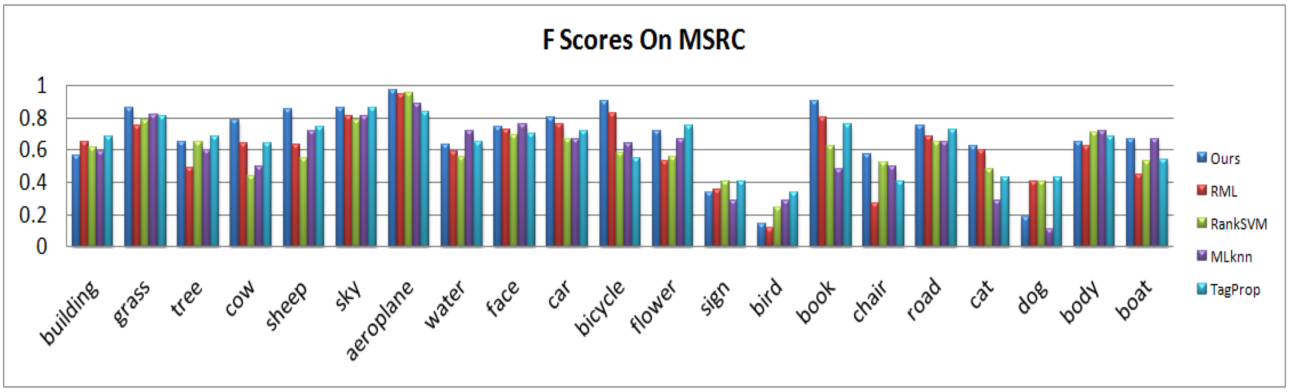

Figure 3. The results of our method in comparison with other related competitive algorithms in terms of F score for individual labels at the image-level on the MSRC dataset.

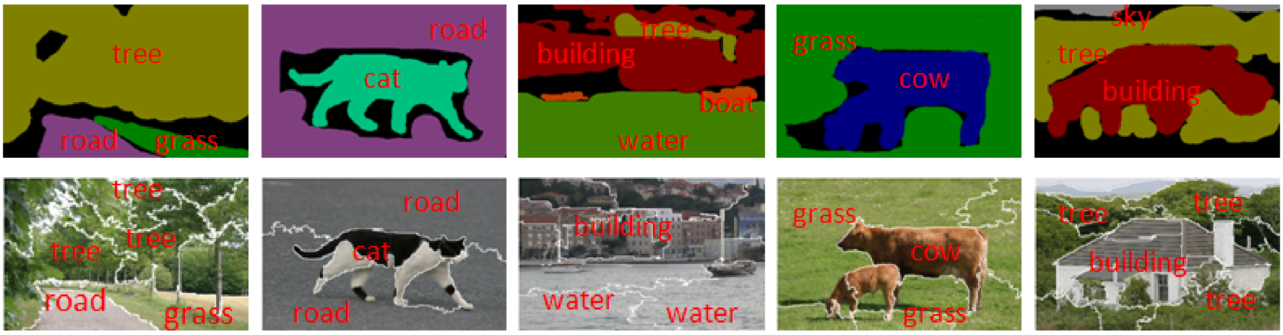

Figure 4. Region-level labeling results of our method for some exemplary images from MSRC. Top: the ground truth; Bottom: our results.

if $f_I^l(I, \mathbf{y}_l = 1, \mathbf{y}_{\mathcal{N}_l})$ is larger than $f_I^l(I, \mathbf{y}_l = -1, \mathbf{y}_{\mathcal{N}_l})$ then $\mathbf{y}_l = 1$; otherwise $\mathbf{y}_l = -1$. Furthermore, based on the derived image-level label vector, the region-level labels can even be refined by simultaneously maximizing the second item while minimizing the forth item in Eq. (2) via the *Cutting Plane* method [16, 12]. Alternatively, based on the derived image-level label vector, we can also refine the region-level labels by calculating the similarity between the image regions and the positive clusters corresponding to the current image-level multi-labels. Finally, the refined image-level labels can be further inferred from the refined region-level labels.

4. Experiments

In this section, we evaluate our method on MSRC and Corel [3] image datasets in comparisons with other related competitive algorithms:1) RML [18], 2) RankSVM [4], 3) MLknn [29], and 4) TagProp [9].

We first conduct experiments on MSRC (MicroSoft Research Cambridge) image dataset which is widely used in multi-label image annotation for performance comparison. It contains 591 images with totally 23 concepts. There are about 3 tags on average per image. We ignore the concepts *horse* and *mountain* since they have few positive samples. Thus there are totally 21 concepts. We randomly divide the dataset into two subsets: 70% for training and 30% for testing. Global and local features are extracted for each image. For each image, we first extract the global visual features: 12-dimensional CLD (Color Layout Descriptor), 64-dimensional SCD (Scalable Color Descriptor), 256-dimensional CSD (Color Structure Descriptor), and 80-dimensional EHD (Edge Histogram Descriptor). These four kinds of visual features are used to calculate the composite distance JEC (Joint Equal Contribution)[17] between the images in Eq. (14). Specifically, the distances for each kind of visual feature are first scaled to be bounded by 0 and 1, and then are averaged such that each kind of visual feature contributes equally towards the image similarity. On the other hand, all images are segmented into several regions and different kinds of local features are extracted for each region: 1) 14-dimensional color feature including mean RGB, HSV conversion, HUE histogram and SAT histogram; 2) 30-dimensional texture feature including LM-filter mean response [14] and LM-filter response histogram; and 3) 8-dimensional geometric feature encoding the position and size information of the segment. Likewise, the composite distance between image regions in Eq. (17) is computed using JEC[17]. Although pixel-level (region-level) ground truth is provided as well in MSRC dataset as shown in the top row of Figure.4, only image-level ground truth is employed for training in experiments. By clustering the image regions with Affinity Propagation [7], and analyzing the inter-cluster similarities together with the cluster sizes, the region-level label identification on the training set can be automatically accomplished, which helps to train the region-level submodel using the image-level labeled samples.

Figure.3 shows the results of our method (Ours) in comparison with other related competitive algorithms in terms of F score for individual labels at the image-level. F score is defined as the harmonic mean of precision and recall, i.e., $F = \frac{precision*recall}{(precision+recall)/2}$. As observed from the results, our method achieves better performance for most concepts, compared to the other related algorithms. Our method simultaneously extracts global and local visual features, and sufficiently leverages various contextual relations, which might be useful to improve the annotation performance. Our method annotates not only the entire image but also the regions within this image. Figure.4 gives some label-to-region assignment results for test images from the MSRC dataset produced by the proposed correlative multi-label multi-instance model. The other methods RML [18], RankSVM [4], MLknn [29], and TagProp [9] can not obtain the region-level labels, but the image-level labels only.

Corel data set [3] contains 5000 images and each image is labeled with 1-5 concepts and there are totally 374 concepts. We carry out the experiments on around 1000 images including ten concepts: $mountain$, sky, $clouds$, $tree$, $people$, $birds$, $buildings$, $bear$, $snow$, $rocks$. 70% are used for training and the rest 30% for testing. Again, global and local features are extracted for each image, and JEC distance is employed to measure the image-image similarity and the region-region similarity. Figure.5 shows the results of our method (Ours) in comparison with other related competitive algorithms in terms of F score for individual labels at the image-level. Figure.6 shows some label-to-region assignment results for test images from the Corel dataset at the region-level. From the experimental results, the following observations can be obtained: i) Our method performs best on five concepts $mountain$, $people$, $bear$, $snow$, $rocks$; ii) Our method achieves the comparable performance on three concepts $tree$, $birds$, $buildings$; iii)The performances on the concepts sky and $clouds$ seem to be not satisfying. Actually, the Corel data set is a weakly labeled dataset (i.e., the given "ground truth" labels of some images may be incomplete). As an example, the annotation result of our method on the bottom-right exemplary image (i.e., the last image) in Figure.6 includes sky but not $cloud$; however, the given image-level "ground truth" includes $clouds$ instead of sky. The incompleteness of the available "ground truth" may impact both the training efficacy and the testing evaluation.

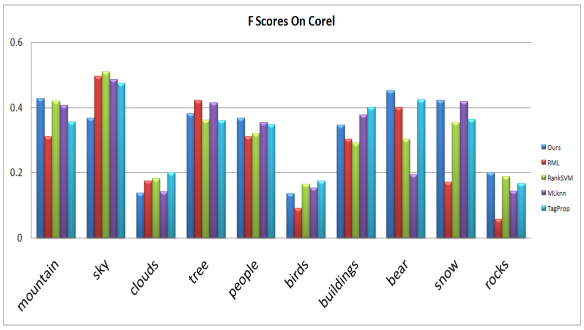

Figure 5. The results of our method in comparison with other related competitive algorithms in terms of F score for 10 labels at the image-level from the Corel dataset.

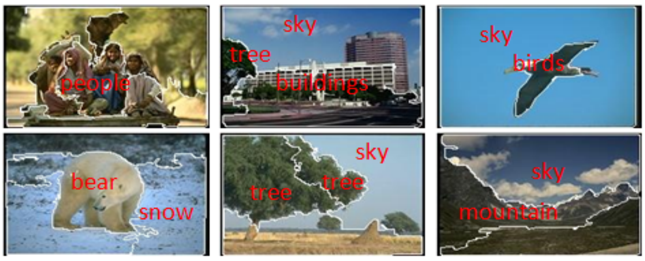

Figure 6. Region-level labeling results of our method for some exemplary images from the Corel dataset.

5. Conclusions

In this paper both image-level labels and region-level labels can be obtained in a single framework by capturing the feature-label associations, the inter-label correlations, and the cross-level label coherence. The associations between semantic concepts and visual features are mined both at the image level and at the region level. Inter-label correlations are captured by a co-occurence matrix of concept pairs. The cross-level label coherence encodes the consistency between the labels at the image level and the labels at the region level. Structural max-margin technique is used to formulate the proposed model. By decoupling the annotation task into inter-dependant subproblems, we learn multiple interrelated classifiers jointly. In our future work, we would investigate how to improve the annotation performance on the weakly labeled image datasets.

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), the STCSM's Innovation Program(No. 10511500703), and the Shanghai Leading Academic Discipline Project (No.B114).

References

[1] Y. Chen, L. Zhu, A. Yuille, and H.-J. Zhang. Unsupervised learning of probabilistic object models (poms) for object classification, segmentation, and recognition using knowledge propagation. *TPAMI 2009*. 1

[2] Y. Deng and b.s. Manjunath. Unsupervised segmentation of color-texture regions in images and video. *PAMI*, 23(8):800–810, 2001. 2

[3] P. Duygulu, K. Barnard, J. d. Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002. 6, 7

[4] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *NIPS*, 2002. 1, 6, 7

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

[6] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR*, 2004. 1

[7] B. Frey and D. Dueck. Clustering by passing messages between data points. *Science, vol.315, 2007*. 5, 7

[8] N. Ghamrawi and A. McCallum. Collective multi-label classification. In *CIKM*, 2005. 1

[9] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009. 1, 6, 7

[10] X. He, R. S. Zemel, and M. A. Carreira-Perpinan. Multiscale conditional random fields for image labeling. In *CVPR*, 2004. 1

[11] R. Jin, S. Wang, and Z.-H. Zhou. Learning a distance metric from multi-instance multi-label data. In *CVPR*, 2009. 1

[12] J. Kelley. The cutting-plane method for solving convex programs. *JSIAM 1960*. 6

[13] C. L. and F.-F. Li. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. *ICCV 2007*. 1

[14] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, 2001. 6

[15] Y.-X. Li, S. Ji, S. Kumar, J. Ye, and Z.-H. Zhou. Drosophila gene expression pattern annotation through multi-instance multi-label learning. In *IJCAI*, 2009. 1

[16] D. Liu, S. Yan, Y. Rui, and H.-J. Zhang. Unified tag analysis with multi-edge graph. *ACM MM 2010*. 1, 6

[17] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV*, 2008. 1, 6, 7

[18] J. Petterson and T. Caetano. Reverse multi-label learning. In *NIPS*, 2010. 1, 6, 7

[19] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative multi-label video annotation. In *ACM MM*, 2007. 1

[20] B. Russell, A. Efros, J. Sivic, W. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006. 2

[21] Y. Shen and J. Fan. Leveraging loosely-tagged images and inter-object correlations for tag recommendation. In *ACM MM*, 2010. 5

[22] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 22(8):888–905, 2002. 2

[23] D. Sontag, O. Meshi, T. Jaakkola, and A. Globerson. More data means less inference: A pseudo-max approach to structured learning. In *NIPS*, 2010. 4

[24] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, 2004. 4

[25] G. Winkler. Image analysis, random fields and dynamic monte carlo methods: A mathematical introduction. *Springer-Verlag, Berlin, Heidelberg*, 1995. 5

[26] Y. Xiang, X. Zhou, Z. Liu, T.-S. Chua, and C.-W. Ngo. Semantic context modeling with maximal margin conditional random fields for automatic image annotation. In *CVPR*, 2010. 1, 4

[27] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *ICML*, 2009. 4

[28] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. In *CVPR*, 2008. 1

[29] M. Zhang and Z. Zhou. Ml-knn: a lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007. 1, 6, 7

[30] W. Zhang, X. Xue, J. Fan, X. Huang, B. Wu, and M. Liu. Multi-kernel multi-label learning with max-margin concept network. In *IJCAI*, 2011. 2, 4

[31] Z.-H. Zhou and M.-L. Zhang. Multi-instance multi-label learning with application to scene classification. In *NIPS*, 2006. 1