Exploring Image Specific Structured Loss for Image Annotation with Incomplete Labelling

Xing Xu, Atsushi Shimada, and Rin-ichiro Taniguch

Department of Advanced Information Technology, Kyushu University, Japan

Abstract. In this paper, we address the problem of image annotation with *incomplete labelling*, where the multiple objects in each training image are not fully labeled. The conventional one-versus-all SVM (OVA-SVM) that performs fairly well on full labelling decays drastically under the incomplete setting. Recently, structured learning method termed OVA-SSVM is proposed to boost the performance of OVA-SVM by modeling the structured associations of labels and show efficiency under incomplete setting. The OVA-SSVM assumes that each training sample includes a single label and adopts an loss measure of classification style that as long as one of the predicted label is correct, the overall prediction should be considered correct. However, this may not be appropriate for the multi-label annotation task. In this paper, we extend the OVA-SSVM method to the multi-label situation and design a novel image specific structured loss measure to account for the dependencies between predicted labels relying on the image-label associations. Then we develop an efficient optimization algorithm to learn the model parameters. Finally, we present extensive empirical results on two benchmark datasets with various degree of incompletion, and show that proposed method outperforms OVA-SSVM and achieves competitive performance compared with other state-of-the-art methods which are also designed for the issue of incomplete labelling.

1 Introduction

Automatic image annotation is an important research problem, where each image is associated with a set of labels and the target is to learn a model that assigns multiple labels to an unlabeled new image to describe its visual content. In particular, the human annotations play an significant role in training an annotation model as they provide empirical knowledge of the image-label associations. Although the quality of human annotations is quite crucial, one can not expect to accurately obtain all the labels for a given image, since human labelers usually tag only prominent labels and typically miss out on several objects present in the image. Here we pose a practical issue of *incomplete labelling* that the training images are not completely tagged with all relevant labels from vocabulary. As shown in Figure 1, the images from two benchmark datasets have few human annotated labels and suffer from the problem of incomplete labelling.

The traditional annotation models such as generative models [3, 4] or nearest-neighbor based models [5–7] generally neglect the issue of incomplete labelling

bloom, leave clouds

Fig. 1. Example of incomplete labelling: two images are from benchmark datasets IAPRTC-12 [1] (left) and NUS-WIDE [2] (right). Potentially correct labels such as {*flower, plant, tree, trunk*}, {*sky, grass*} are missed from the ground truth of two images respectively.

and treat the human annotated dataset as completely labeled. For these models, given a labeled training image, labels that are not presented in the groundtruth of that image make limited contribution to the annotation model. When applying these models on incompletely labeled dataset, the annotation performance can hardly achieve optimal because of the insufficient annotations of the dataset. Therefore, in our work, we intend to develop an annotation method that is more efficient for the incompletely labeled dataset. It should be noted that our problem setting is different from the researches of tag completion [8–10] or tag recommendation [11, 12], where their goal is to complete partially tagged images offline or to recommend related tags to users online.

Regarding the methodology of annotation with incomplete labelling, one group of recent ongoing researches [13–16] directly modify conventional annotation prototypes such as multi-label ranking [13][16], binary SVM [14], and ridge regression [15], by incorporating additional consistency between visual and semantic cues in images to address the issue of incompletion. And the performance of these methods greatly depends on the assumption of consistency. Moreover, another group of works aim at boosting the conventional annotation models and adding new learning stage under the incomplete setting incrementally. The method utilized in the new learning stage could be multi-task learning [17], ensemble learning [18], and structured learning [19, 20]. Here we would like to stress that structured learning method is an efficient scheme to handle the difficulties of incomplete labelling. Firstly, it captures the interdependencies of labels from the structure in the output space. Secondly, the weak learning manner allows it to explore the potential usages of missing labels, and those missing labels can be captured by latent variables [21].

Specifically, in the celebrated work termed OVA-SSVM of [20], structured learning method is adopted to boost the performance of pre-trained OVA-SVM classifiers under incomplete setting, and it designs a structured loss function of image classification style to benefit the prediction of missing labels. And promising prediction results are obtained on ImageNet [22] dataset where each training image has a single label. However, the OVA-SSVM method may not be well ex-

tended to more practical circumstances where each training image has multiple labels, due to limitation of its structured loss used. Therefore, in this paper, we put effort to improve OVA-SVM in three folds: (1) We extend the OVA-SSVM so that the number of labels per training image can be flexible, which is more practical to the multi-label annotation problem. (2) We design a novel image specific structured loss function which is more appropriate than previous flat structured loss used in OVA-SSVM to account for the dependencies between predicted labels relying on the specific image. (3) We develop efficient optimization algorithm with lower complexity by exploiting the properties of proposed structured loss. Empirical evaluations on two annotation datasets with various degree of incompletion demonstrate that proposed annotation method can boost conventional OVA-SVM classifiers, perform better than previous structured learning method OVA-SSVM, and achieve competitive performance compared with state-of-the-art methods designed for incompletely labeled dataset.

2 Problem formulation

In this section, we will first introduce the conventional OVA-SVM used for image annotation task, then describe the OVA-SSVM method that uses structured learning to boost OVA-SVM classifiers under incomplete setting. Some notations used in the following sections are also defined in this section.

2.1 Conventional OVA-SVM

We are given an incompletely labeled dataset $\mathcal{T} = \{(x^1, Y^1), ..., (x^N, Y^N)\}$. Here $x^n \in \mathcal{X}$ represents the image feature vector, $Y^n \subseteq \mathcal{Y}$ is a set of labels, where $\mathcal{Y} = \{y_1, ..., y_C\}$ is the vocabulary of C labels. Note that Y^n is a subset of the ideally full set Ω^n of groundtruth labels for image x^n. Our goal is to learn an annotation model that, for an unseen image x, outputs an optimal set $\hat{Y}$ which includes K distinct labels.

A conventional annotation model consists of learning a series of binary OVA-SVM classifiers that distinguish a single label from all other. For a given label y_c, we denote the parameter vector of learnt OVA-SVM classifier as $\boldsymbol{w}_{OVA}^{y_c}$, then to predict a set of K labels $\hat{Y}$ for an unseen image x, the annotation model simply returns the labels with the K highest scores performing on classifiers of all labels:

$$\hat{Y} = arg \max_{Y \in \mathcal{Y}} \sum_{y_c \in Y} x \cdot \boldsymbol{w}_{OVA}^{y_c}, \tag{1}$$

where $Y \subseteq \mathcal{Y}$ represents any output set contain K labels. It is worth noting that the annotation model of OVA-SVM classifiers is suboptimal since that (1) the one-versus-all learning manner ignores the dependencies of labels, which implies that OVA-SVM optimizes the prediction of only a single output label, ignoring the "structure" altogether, (2) the performance of OVA-SVM classifiers drops drastically when incomplete labels for training image are provided.

2.2 OVA-SSVM

To overcome the disadvantages of conventional OVA-SVM and to exploit the structured associations in output label set Y, the structured learning method OVA-SSVM [20] considers that the training set consists of structured input-output pairs $\mathcal{T} \in (\mathcal{X} \times \mathcal{Y})^N$. The prediction rule of optimal output labels $\hat{Y}$ for an unseen image x is

$$\hat{Y} = arg\max_{Y \in \mathcal{Y}} \Phi(x, Y) \cdot \boldsymbol{w} = arg\max_{Y \in \mathcal{Y}} \sum_{y \in Y} \phi(x, y) \cdot \boldsymbol{w}, \tag{2}$$

where Φ is the joint feature vector that describes the relationship between input x and any structured output Y, ϕ is the joint feature vector for input x and single label y in Y, and $\boldsymbol{w}$ is the parameter vector to be learnt. In particular, given a set of pre-trained OVA-SVM classifiers $\{\boldsymbol{w}_{OVA}^{y_c}\}^{y_c \in \mathcal{Y}}$, the joint feature vector $\Phi(x, Y)$ in OVA-SSVM is defined as

$$\Phi(x, Y) = \sum_{y \in Y} x \circ \boldsymbol{w}_{OVA}^{y}, \tag{3}$$

where $x \circ \boldsymbol{w}_{OVA}^{y}$ represents the Hadamard product of x and $\boldsymbol{w}_{OVA}^{y}$. Then the annotation model in Eq. 2 can be formulated as

$$\hat{Y} = arg\max_{Y \in \mathcal{Y}} \sum_{y \in Y} \langle x \circ \boldsymbol{w}_{OVA}^{y}, \boldsymbol{w} \rangle. \tag{4}$$

We can learn from Eq.4 that OVA-SSVM incrementally learns a single parameter vector $\boldsymbol{w}$ that re-weights the parameters of existing OVA-SVM classifiers $\{\boldsymbol{w}_{OVA}^{y_c}\}^{y_c \in \mathcal{Y}}$ and incorporates the structure nature of output Y through the joint feature vector $\Phi(x, Y)$.

Moreover, for the incomplete setting of $\mathcal{T}$, the input-output relationship is not completely characterized by $(x, Y) \in \mathcal{X} \times \mathcal{Y}$. It is rational to introduce a set of unobserved latent variables, $Z = \{Z^1, ..., Z^N\}$, where Z^n represents the set of labels that appear in image x^n but were not annotated. The full set of labels for the image x^n is $\Omega^n = Y^n \cup Z^n$ (note that $Y^n \cap Z^n = \varnothing$). Now the joint feature vector $\Phi(x, \Omega)$ describes the relation among input x, output Y and latent variables Z, and it is defined as

$$\Phi(x, \Omega) = \sum_{y \in Y} x \circ \boldsymbol{w}_{OVA}^{y} + \sum_{z \in Z} x \circ \boldsymbol{w}_{OVA}^{z} \tag{5}$$

To train OVA-SSVM, the parameter vector $\boldsymbol{w}$ is determined by minimizing the regularized risk on the training set $\mathcal{T}$. Risk is measured through a user-provided structured loss function $\Delta(Y, Y^n)$ that quantifies how much the prediction Y differs from the given label set Y^n of image x^n. The resulting convex optimization problem is to minimize the objective function as

$$\min_{\boldsymbol{w}, \xi} \frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \frac{1}{N} \sum_{n=1}^{N} \xi_n \tag{6}$$
$$s.t \quad \boldsymbol{w} \cdot \Phi(x^n, \Omega^n) - \boldsymbol{w} \cdot \Phi(x^n, Y) \geq \Delta(Y, Y^n) - \xi_n, \quad \forall n, Y \in \mathcal{Y}.$$

The constraints of Eq.6 identify the prediction Y with a score $\boldsymbol{w} \cdot \Phi(x^n, Y)$ that is smaller than the score $\boldsymbol{w} \cdot \Phi(x^n, \Omega^n)$ of the "full" groundtruth Ω^n by a soft margin equals to the loss $\Delta(Y, Y^n)$ with the slack variable ξ_n. The optimization problem can be solved efficiently using a constraint generation strategy: we can generate the constraint by identifying the most violated (incorrect) prediction $\bar{Y}$ from Y for the current parameter vector $\boldsymbol{w}$ on x^n. This amounts to solving

$$\bar{Y} = arg \max_{\hat{Y} \in \mathcal{Y}} \{\Delta(Y, Y^n) + \boldsymbol{w} \cdot \Phi(x^n, Y)\}. \tag{7}$$

Given the definition of user-provided structured loss Δ, we can use $\bar{Y}$ of all $x^n \in \mathcal{X}$ to approximate a lower bound of the objective in Eq.6. Then we can compute the gradient of Eq.6, and alternately optimize the latent variables Z and the parameter vector $\boldsymbol{w}$. In the next section, we will introduce proposed image specific loss term which is elaborately designed for incomplete labeled training data, and derive the corresponding optimization algorithm in the structured learning framework.

3 Proposed structured loss under incomplete setting

3.1 Image specific structured loss

Since the given label set Y^n may not describe all the object in image x^n, an annotation model should not be penalized for predicting "incorrect" labels that actually describe those objects in x^n. To address this issue, a structured loss function Δ is designed in OVA-SSVM. Given a set of predicted output labels Y for x^n, the OVA-SSVM method would not give penalty if one of the predicted labels $y \in Y$ is similar to any of the groundtruth labels $y^n \in Y^n$. The loss function is defined as

$$\Delta(Y, Y^n) = \min_{y \in Y} \min_{y^n \in Y^n} d(y, y^n), \tag{8}$$

where $d(y, y^n)$ is the error term measuring the difference between label y and y^n. In practice, $d(y, y^n)$ could be a flat error measure: $d(y, y^n) = 0$ if $y = y^n$, and 1 otherwise. And $d(y, y^n)$ could also be a hierarchical error measure: $d(y, y^n)$ is the shortest path distance between y^n and y in a taxonomic vocabulary tree.

Actually, there are several limitations of the structured loss of Eq.8 for the incomplete setting. Firstly, to predict output labels Y, ensuring that only one of the predicted labels is similar to the groundtruth is not enough. In other words, it is expected that each of the predicted labels is similar to any (even all) of the groundtruth labels. Secondly, the error measure of $d(y, y^n)$ is either coarse to quantify the difference of labels (i.e. flat error measure), or rigorous to require the prior construction of taxonomic tree (i.e. hierarchical error measure). Thirdly, the error measure indicates that the variances of labels are based on the global statistics of training data, whereas for the incomplete setting, it is not sufficient to model the relatedness of missing labels and groundtruth labels. In

Figure 2, we demonstrate two examples of label prediction using *flat* structured loss in OVA-SSVM. It can be observed that, although flat structured loss (in fifth column) is generated to be zero (since the predicted labels *bloom*, *man* match the incomplete groundtruth Y^n), the predicted result (in third column) is inferior which contains several incorrect labels, e.g. {*fruit, forest*}, {*woman, bottle, forest*}. Thus, it implies that numerically minimizing the structured loss of Eq.8 could not guarantee all predicted labels to be similar to groundtruth labels.

Image x^n	Incomplete labels Y^n	Predicted labels Y		Structured loss Δ	
		OVA-SSVM (Flat)	Proposed	Flat	Image specific
	bloom, leave	bloom, flower, fruit, forest, branch	bloom (0), leave (0), trunk (0.531), flower (0.552), plant (0.765)	0	0.3696
	man, one, rock	man, woman, front, bottle, forest	man (0), rock (0), tee-shirt (0.647), hand (0.685), waterfall (0.689)	0	0.4042

Fig. 2. Examples of label prediction using flat loss (OVA-SSVM (Flat)) *vs.* image specific structured loss (proposed method). These two images are selected from IAPRTC-12 dataset. Note that in the fourth column, the image specific loss of each predicted label is also provided. And the loss values Δ in last two columns are calculated according to Eq.8 and Eq.9 respectively.

To address the limitations of Eq.8, we assume that each of the predicted labels is related to *all* of the groundtruth labels, and we desire the structured loss term to capture the variances of labels relying on the specific image content. Our proposed image specific loss function is formulated as

$$\Delta(Y, Y^n; x^n) = \frac{1}{|Y|}\frac{1}{|Y^n|}\sum_{y\in Y}\sum_{y^n\in Y^n} d(y, y^n; x^n). \tag{9}$$

Here the error measure $d(y, y^n; x^n)$ is image specific, representing the difference of label y and y^n particularly on image x^n. In addition, the structured loss $\Delta(Y, Y^n; x^n)$ ensures that each of the predicted labels in Y to be related to all the groundtruth labels in Y^n. Since the incomplete label set Y^n is small, here we restrict the structured loss of Eq.9 to moderately consider the dependencies between each of the predicted labels and all labels in Y^n.

Inspired by the works [10, 14], we cast measuring $d(y, y^n; x^n)$ to comparing the relatedness of image x^n to labels y and y^n. In particular, for a given label y_c, let $\mathcal{X}_c^+$ be the set of images that are annotated with label y_c, and the remaining images as $\mathcal{X}_c^- = \mathcal{X} \backslash \mathcal{X}_c^+$. For image x^n in $\mathcal{X}_c^+$, we define the relatedness of image x^n to label y_c as $R(x^n, y_c) = 1$ since x^n is annotated with y_c. And for image x^n belongs to $\mathcal{X}_c^-$ of y_c, we determine the relatedness score of $R(x^n, y_c)$ considering three factors: visual similarity, semantic similarity and image-label association in the visual neighborhood. Specifically, $R(x^n, y_c)$ consists of

- Visual similarity based relatedness score $R_V(x^n, y_c)$: We compute the visual distance $dist(\cdot)$ (scaled to range $[0, 1]$) of x^n with its nearest neighbor $x^* \in \mathcal{X}^+$, and define $R_V(x^n, y_c) = 1 - dist(x^n, x^*)$.
- Semantic similarity based relatedness score $R_S(x^n, y_c)$: We first compute the correlation score between pairwise labels y_i and y_j, $\forall y_i, y_j \in \mathcal{Y}$ as: $co\_occur(y_i, y_j) = \frac{f_{i,j}}{f_i + f_j - f_{i,j}}$, where f_i and f_j are the count of occurrence of labels y_i and y_j, and $f_{i,j}$ is the count of co-occurrence of labels y_i and y_j. Let Y^n be the label set of image x^n, we define $R_S(x^n, y_c) = \max_{y \in Y^n} co\_occur(y_c, y)$.
- Reverse nearest neighbors based relatedness score $R_N(x^n, y_c)$: For a fixed value of $M(= 5)$, let p_m be the number of images in $\mathcal{X}_c^+$ that have x^n as their m^{th} nearest neighbor. Then we define $R_N(x^n, y_c) = \sum_{m=1}^{M} \frac{p_m}{m} / \sum_{m=1}^{M} p_m + \varepsilon$, where $\varepsilon > 0$ is a small number to avoid division by zero.

Finally, $R(x^n, y_c)$ is defined as the average of these three scores, similar as in [14]:

$$R(x^n, y_c) = average(R_V(x^n, y_c) + R_S(x^n, y_c) + R_N(x^n, y_c)). \tag{10}$$

Now we can calculate the error measure $d(y, y^n; x^n)$ by comparing the relatedness scores of image x^n to labels y and y^n as

$$d(y, y^n; x^n) = R(x^n, y^n) - R(x^n, y) = 1 - R(x^n, y). \tag{11}$$

Recalling that $y^n \in Y^n$ is the groundtruth label of x^n, thus it has highest relatedness score (equals to 1). It can be learnt that the calculation of Eq.11 is directly determined by the relatedness score $R(x^n, y)$ of label y to image x^n. And if the predicted label y has larger relatedness score to x^n, it would have small difference with all the groundtruth labels. This is consistent with the proposed structured loss of Eq.9, which now can be efficiently measured by the relatedness of predicted labels to the specific image.

Compared with the flat/hierarchical structured loss, our proposed structured loss of Eq.9 has several advantages. Firstly, as shown in Figure 2, although the loss values (in last column) are numerically larger than "zero" of flat structured loss (in fifth column), the predicted labels is more similar to the provided incomplete labels. This is because proposed structured loss moderately considers the predicting labels based on their relatedness to specific image content, and the relatedness measure is elaborately designed and more appropriate than the simple 0-1 measure. Secondly, the proposed structured loss is more flexible to the number of groundtruth labels as it accumulatively measures each of the predicted labels to all the groundtruth, while the flat structured loss focuses on the

most dominant one in the predicted label to a single label of the groundtruth labels. Thirdly, the relatedness measure can be directly and precisely computed from labeled training images, while to construct the hierarchical measure, usually prior knowledge of taxonomy or large quantities of training data with full labelling is required.

3.2 Optimization method

Given the proposed structured loss function of Eq.9, we can generate the most violated constraint of prediction $\bar{Y}$ for image x^n according to Eq.7 as the form

$$\begin{aligned}\bar{Y} &= arg\max_{Y\in\mathcal{Y}}\{\frac{1}{|Y|}\frac{1}{|Y^n|}\sum_{y\in Y}\sum_{y^n\in Y^n} d(y,y^n;x^n)+\sum_{y\in Y}\boldsymbol{w}\cdot\phi(x^n,y)\}\\ &= arg\max_{Y\in\mathcal{Y}}\{\frac{1}{|Y|}\sum_{y\in Y}(1-R(x^n,y))+\sum_{y\in Y}\boldsymbol{w}\cdot\phi(x^n,y)\},\end{aligned} \tag{12}$$

where the calculation of structured loss $\Delta(Y,Y^n;x^n)$ is converted to compute the relatedness scores of predicted label set Y to image x^n, as described in Section 3.1. We can obtain the solution of Y of Eq.12 by simply sorting the term $\frac{1}{|Y|}(1-R(x^n,y_c))+\boldsymbol{w}\cdot\phi(x^n,y_c)$ for each label $y_c\in\mathcal{Y}$, and then choose the top K labels for $\bar{Y}$. Solving Eq.12 greedily takes $\mathcal{O}(C\log C)$, thus it is faster than the constraints generation method in OVA-SSVM which takes $\mathcal{O}(C^2\log C)$. After we have generated the most violated constraint $\bar{Y}$ for each image, the lower bound of the objective function in Eq.6 can be derived as

$$J(\boldsymbol{w})=\frac{\lambda}{2}\|\boldsymbol{w}\|^2+\frac{1}{N}\sum_{n=1}^{N}\left[\Delta(\bar{Y}-Y^n)+\boldsymbol{w}\cdot\Phi(x^n,\bar{Y})-\boldsymbol{w}\cdot\Phi(x^n,\Omega^n)\right], \tag{13}$$

and the gradient of $J(\boldsymbol{w})$ with respect to $\boldsymbol{w}$ is

$$\nabla_{\boldsymbol{w}}J(\boldsymbol{w})=\lambda\boldsymbol{w}+\frac{1}{N}\sum_{n=1}^{N}\left[\boldsymbol{w}\cdot\Phi(x^n,\bar{Y})-\boldsymbol{w}\cdot\Phi(x^n,\Omega^n)\right]. \tag{14}$$

It can be observed in Eq.13 and Eq.14 that calculating $J(\boldsymbol{w})$ and its gradient involves in computing the joint feature vector $\Phi(x^n,\Omega^n)$ on "full" label set Ω^n of each image. And $\Phi(x^n,\Omega^n)$ can be efficiently computed according to Eq.5 with latent variable Z^n. To learn the parameter vector $\boldsymbol{w}$ with latent variable Z^n, we follow the previous alternating optimization technique proposed in [19, 20]. Specifically, we alternate between optimizing the parameter vector $\boldsymbol{w}^t$ by initializing the latent variable Z^n for each image in the t^{th} iteration, and re-estimate the latent variable Z^n for the $(t+1)^{th}$ iteration given the learnt parameter vector $\boldsymbol{w}^t$. The pseudocode for solving the alternating optimization problem is depicted in Algorithm 1.

Algorithm 1 Alternating optimization of proposed method

Input: Incompletely labeled training data $\mathcal{T} = \{(x^n, Y^n)\}_{n=1}^N$, pre-trained binary classifiers $\{\boldsymbol{w}_{OVA}^{y_c}\}_{c=1}^C$
Output: Parameter vector $\boldsymbol{w}$
1: Initialize $\boldsymbol{w}_0 = \mathbf{1}$ for iteration $t = 0$
2: **repeat**
3: Set $t = t + 1$
4: **for** $n = 1, ..., N$ **do**
5: Assign latent variable $Z_t^n = \{arg\max_{Y \in \mathcal{Y}} \boldsymbol{w}_{t-1} \cdot \Phi(x^n, Y)\} \backslash Y^n$ for x^n (preserving $K - |Y^n|$ missing labels)
6: **end for**
7: **for** $n = 1, ..., N$ **do**
8: Generate the most violated constraint $\bar{Y}_t$ for x^n according to Eq.12
9: **end for**
10: Compute objective $J_t(\boldsymbol{w})$ and gradient $\nabla_{\boldsymbol{w}} J_t(\boldsymbol{w})$ according to Eq.13 and Eq.14
11: Minimize loss of Eq.6 to calculate $\boldsymbol{w}_t$
12: **until** Loss in Eq.6 is converged

4 Experimental evaluation

In this section, we evaluation the effectiveness of proposed method through comparing it with the previous OVA-SSVM and other state-of-the-art annotation methods under incomplete setting.

4.1 Experimental setup

Datasets and Features. Our evaluation experiments are conducted on two publicly available benchmark datasets: IAPRTC-12 [1] and NUS-WIDE [2]. These two datasets are very challenging with significant diversity among the images that are obtained from the social web. Table 1 shows the general statistics of these two datasets, and it is worth noting that they cover both conditions of large vocabulary size and large number of images. In our experiments, for IAPRTC-12 dataset, we use the same multiple features as those in [5, 6, 14, 15]. These multiple features consist of global and local features. The global features include histograms in RGB, HSV and LAB color space, and the GIST features; and the local features include the SIFT and hue descriptors obtained densely from multi-scale grid, and from Harris-Laplacian interest points. For NUS-WIDE dataset, besides global GIST features, we also extract five types of SIFT based local features (C-SIFT, Opponent-SIFT, RGB-SIFT, RG-SIFT) using the public colorDescriptor tools [23]. The SIFT based features are computed without orientation invariance and the grid has a step size of three. The codebook for each SIFT based feature is generated from 7,000 randomly selected images, and quantized to 4,000 corresponding k-means clusters. For both datasets, we first separately perform L2 normalization for each type of feature, and then concatenate them to an fused feature vector (37,152-dimension for IAPRTC-12 and 20,512-dimension for NUS-WIDE) to represent each image.

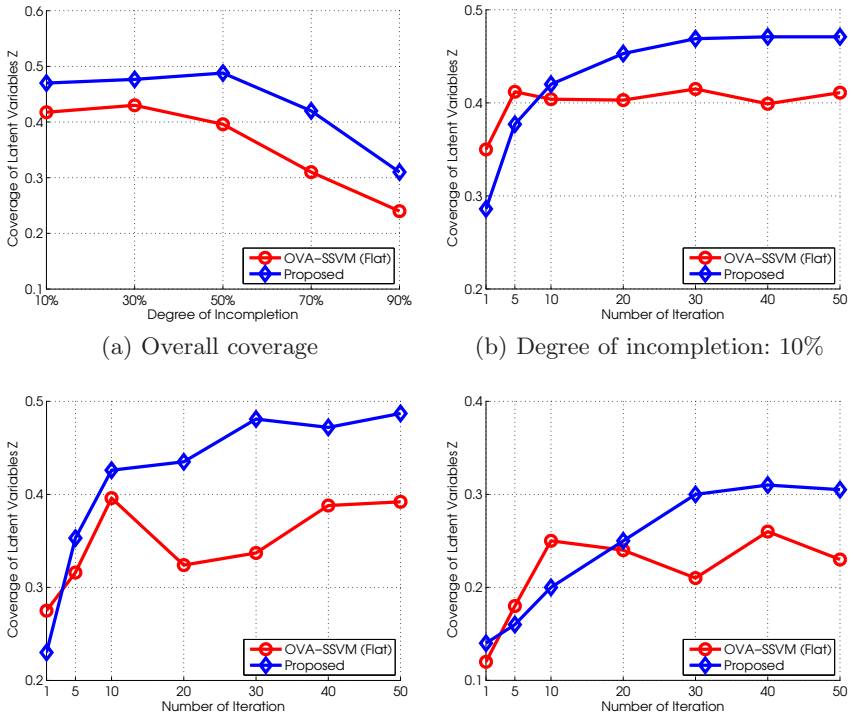
Table 1. General statistics for the two datasets used for evaluation. The items in the second row are listed in the format "training/test", and items in the third and fourth rows are given in the format "mean/minimum/maximum".

	IAPRTC-12	NUS-WIDE
Total labels	291	81
No. of images	17,665/1,962	138,563/92,484
Labels per image	5.7/1/23	1.8/1/20
Images per label	34/153/4,999	2,512/333/16,425

Incomplete setting. We consider the original IAPRTC-12 as fully labeled dataset since the average number of labels per image is more than 5 (5.7 in Table 1), which could be sufficient to describe multiple objects in an image. To simulate the incomplete setting, we randomly delete partial labels for each image, and the deletion process stands by the principle $\min(1, \lceil M \times (1 - ratio) \rceil)$ ensure that each image preserve at least one label. Here M denotes the number of original labels of an image, $\lceil \cdot \rceil$ denotes the ceiling function which gives the smallest integer not smaller than the given value, and $ratio$ represents the degree of incompletion. In our experiments, we set $ratio = \{10\%, 30\%, 50\%, 70\%, 90\%\}$, and it indicates that the larger the ratio is, the higher the degree of incompletion would be. For NUS-WIDE, as the average number of labels per image is less than 2 (1.8 in Table 1), which could be insufficient compared with the situation of IAPRTC-12, we treat the NUS-WIDE as incomplete labeled dataset, and directly utilize the original annotations for incomplete setting.

Binary classifiers. As proposed method needs pre-trained binary classifier for each class as a starting point for structured learning, we follow previous works [14, 20] and learn OVA-SVM classifiers for initialization. In particular, we train a linear OVA-SVM classifier for each label using Pegasos [24] algorithm and calibrate the raw confidence scores from the SVM classifiers to probabilities with Platt [25] algorithm. Finally, we obtain linear OVA-SVM classifiers with compatible probability scores and use them as initial input to proposed method.

Evaluation metrics. Given an unlabeled test image, we first compute the score for each label using the learnt model, and then select five top-scoring ($K = 5$, $|Y| = 5$) labels according to Eq.4. And we use two standard criteria to evaluate the performance: (1) average precision per label P, (2) average recall per label R. Note that the P and R scores are obtained by first computing precision and recall for each label and then averaging. In addition, as the number of labels in NUS-WIDE dataset is considerably small, we add another two criteria: *Hamming loss* and *Average AUC*, which take the performance of overall prediction and ranking into account. For all the adopted evaluation metrics except *Hamming loss*, larger numerical value indicates better performance.



(a) Overall coverage (b) Degree of incompletion: 10%

(c) Degree of incompletion: 50% (d) Degree of incompletion: 90%

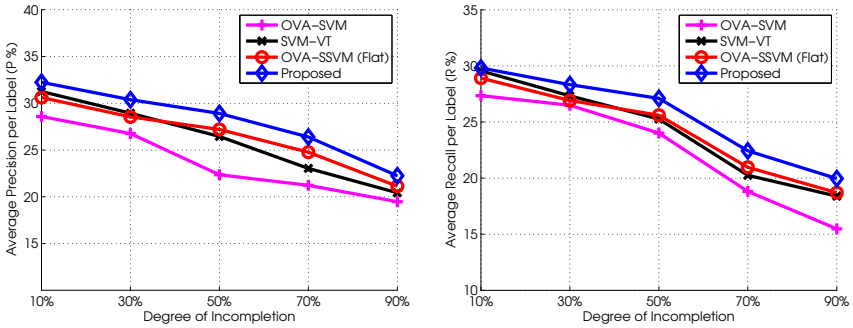
Fig. 3. Evaluation of coverage of latent variables with various degree of incompletion.

4.2 Evaluation on IAPRTC-12 dataset

Assessment of assigning latent variables. We first consider proposed method using image specific structured loss and OVA-SSVM (Flat) method using flat structured loss, to compare the efficiency of structured learning with latent variables. Specifically, we explore how closely the assigned latent variable Z^n matches those labels $\Omega^n \backslash Y^n$ deleted from the originally full annotations of image x^n when training as in Algorithm 1. We use a measure termed $Coverage = \frac{1}{N}\sum_{n=1}^{N} \frac{|Z^n \cap (\Omega^n \backslash Y^n)|}{|Z^n|}$ to represent the averaged intersection between Z^n and $\Omega^n \backslash Y^n$ for all training image $x^n \in \mathcal{X}$. Note that higher coverage indicates better assignments of latent variables.

Figure 3(a) shows the overall coverage of latent variable to the deleted labels in the full annotations with different degree of incompletion. It can be observed that (1) the coverage of latent variable of both methods increases when the degree of incompletion becomes lower, and this is reasonable because the more labels we have, the better we can predict the missing labels; (2) our proposed method consistently obtains higher coverage for missing labels than OVA-SSVM (Flat) which simply uses flat structured loss, as the image specific structured loss used

in our method is more efficient to exploit various contextual information of labels and images under the incomplete setting. Furthermore, in Figure 3(b)-(d), we explicitly demonstrate the changing of coverage of the latent variables through the iterations (as described in Algorithm 1) under different degree of incompletion: 10%, 50%, 90%. We can learn that proposed image specific structured loss is appropriate to ensure our method to perform robustly, while OVA-SSVM (Flat) seems to be unstable through the iterations and results in inferior coverage. Especially, when the degree of incompletion is pretty high (e.g. 50%, 90%), the coverage of proposed method is significant better than OVA-SSVM (Flat), which solidly verifies the superiority of proposed method under the incomplete setting.



(a) Comparison of P (b) Comparison of R

Fig. 4. Comparison of annotation performance with various degree of incompletion.

Overall comparison with various degree of incompletion. To make comprehensive comparison, we first explore the labeling results from binary classifiers: OVA-SVM and SVM-VT [14] (OVA-SVM combined with proposed image specific label relatedness as depicted in Section 3.1, without structured learning), then boost the binary classifiers by structured learning via flat structured loss (OVA-SSVM (Flat)) and image specific structured loss (proposed method).

Figure 4 shows the annotation results of four methods in terms of P and R with various degree of incompletion. Firstly, it can be seen that as the degree of incompletion decreases, the performance of all methods becomes better, since we have more labels for training. Secondly, our method can boost the performance of binary classifiers OVA-SVM and SVM-VT under incomplete setting, which verifies the efficiency of the incrementally structured learning. Thirdly, regarding the structured learning stage, proposed method performs remarkably better than OVA-SSVM (Flat) which uses the flat structured loss, especially when the degree of incompletion is considerably high (50% $\sim$ 90%). The reason behind this is

that we use more appropriate structured loss which efficiently accounts for the dependencies between the predicted labels under the incomplete setting.

4.3 Evaluation on NUS-WIDE dataset

Regarding experiments on incompletely labeled NUS-WIDE dataset, besides the four methods compared above, we also consider state-of-the-art annotation methods with assumptions of full labelling and incomplete labelling. Methods for full labelling include JEC [5], Tagprop [6], and M3L [26]. Methods for incomplete labelling consist of SVM-VT [14], MLR-GL [13], Fasttag [15] and LEML [16]. To make fair comparison, we use codes provided by the authors and follow the instructions in corresponding papers to tune model parameters.

Table 2. Annotation performance comparison among different methods on NUS-WIDE dataset. Previous and our best results are highlighted in bold.

Method	P (%)	R (%)	Average AUC	Hamming loss
JEC [5]	11.9	16.6	0.557	0.083
Tagprop [6]	13.2	23.8	0.707	0.074
OVA-SVM	12.3	22.8	0.782	0.079
M3L [26]	16.1	23.2	0.791	0.071
SVM-VT [14]	16.7	24.3	0.806	0.069
MLR-GL [13]	14.2	23.5	0.722	0.078
Fasttag [15]	**18.4**	21.3	**0.834**	0.067
LEML [16]	17.5	24.6	0.798	0.076
OVA-SSVM (Flat) [20]	16.9	24.1	0.772	0.070
Proposed	17.7	**25.6**	0.819	**0.064**

Table 2 shows the annotation performance of different methods. And we can make the following observations: (1) The proposed method consistently boosts the binary SVM classifiers (OVA-SVM and SVM-VT) and also obtain better performance than OVA-SSVM (Flat). (2) The annotation methods including proposed method designed for incomplete labelling are generally superior to conventional annotation methods with full labelling, which again addresses the significance of tackling the issue of incompletion of practical annotation data. (3) The proposed method performs comparable or better than even the recently proposed methods with incomplete labelling, which corroborates the efficiency of structured learning on capturing the semantic correlations of labels when labels are incomplete.

Figure 5 gives qualitative samples of the annotation results of the proposed method on the two datasets. In particular, for IAPRTC-12 dataset, we preserve the original training images without the deletion process to evaluate the generalization of proposed method. From the samples we can see that, although the number of groundtruth labels are few, our method can still make correct

prediction to them. In addition, our method can also reflect semantic connectedness among the predicted labels, e.g. {*field, landscape*}, {*gravel, road, stone*}, {*beach, sand, ocean*}, etc. This further demonstrates the effectiveness of proposed method using structured learning.

Fig. 5. Samples of annotation results of the proposed method on IAPRTC-12 (the upper row) and NUS-WIDE (the lower row). The red labels are the groundtruth and black ones are top five labels predicted using proposed method.

5 Conclusion and future work

In this paper, to tackle the issue of incomplete labelling, we leverage the structured learning method to boost the performance of conventional OVA-SVM classifiers, and we formulate an image specific structured loss function which is more appropriate to explore the dependencies of predicted multiple labels. We further develop an efficient optimization algorithm with lower computational complexity to learn model parameters. Experimental evaluation verifies that the proposed annotation method is efficient to handle the issue of incomplete labeling and performs superior than several existing methods. In the future, we are planning to extend our method to the scenario where even some of the incomplete labels are incorrectly assigned to the training samples. This in turn would facilitate the annotation model to be robust against the defection of training data.

Acknowledgement. This work was partly supported by Grant-in-Aid for Scientific Research (B), Grant Number 24300074. We thank reviewers for the precious comments.

References

1. Grubinger, M.: Analysis and Evaluation of Visual Information Systems Performance. PhD thesis, Victoria University (2007)
2. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: a real-world web image database from national university of singapore. In: Proceedings of the ACM international conference on image and video retrieval. (2009) 48
3. Xiang, Y., Zhou, X., Chua, T.S., Ngo, C.W.: A revisit of generative model for automatic image annotation using markov random fields. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2009) 1153–1160
4. Feng, S., Manmatha, R., Lavrenko, V.: Multiple bernoulli relevance models for image and video annotation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Volume 2. (2004) 1002–1009
5. Makadia, A., Pavlovic, V., Kumar, S.: A new baseline for image annotation. In: 10th European Conference on Computer Vision (ECCV). (2008) 316–329
6. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In: IEEE 12th International Conference on Computer Vision (ICCV). (2009) 309 –316
7. Verma, Y., Jawahar, C.: Image annotation using metric learning in semantic neighbourhoods. In: 12th European Conference on Computer Vision (ECCV). (2012) 836–849
8. Wu, L., Jin, R., Jain, A.: Tag completion for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **35** (2013) 716–727
9. Lin, Z., Ding, G., Hu, M., Wang, J., Ye, X.: Image tag completion via image-specific and tag-specific linear sparse reconstructions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2013) 1618–1625
10. Xu, X., Shimada, A., Taniguchi, R.i.: Tag completion with defective tag assignments via image-tag re-weighting. In: IEEE International Conference on Multimedia and Expo (ICME). (2014) 1–6
11. Sigurbjörnsson, B., van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: Proceedings of the 17th international conference on World Wide Web (WWW). (2008) 327–336
12. Agrawal, R., Gupta, A., Prabhu, Y., Varma, M.: Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In: Proceedings of the 22nd international conference on World Wide Web (WWW). (2013) 13–24
13. Bucak, S.S., Jin, R., Jain, A.K.: Multi-label learning with incomplete class assignments. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2011) 2801–2808
14. Verma, Y., Jawahar, C.V.: Exploring svm for image annotation in presence of confusing labels. In: British Machine Vision Conference (BMVC). (2013)
15. Chen, M., Zheng, A., Weinberger, K.: Fast image tagging. In: Proceedings of the 30th International Conference on Machine Learning (ICML). (2013) 1274–1282
16. Yu, H.F., Jain, P., Kar, P., Dhillon, I.S.: Large-scale multi-label learning with missing labels. In: Proceedings of the 30th International Conference on Machine Learning (ICML). (2013)
17. Binder, A., Samek, W., Müller, K.R., Kawanabe, M.: Enhanced representation and multi-task learning for image annotation. Computer Vision and Image Understanding (CVIU) (2013) 466–478

18. Dimitrovski, I., Kocev, D., Loskovska, S., Deroski, S.: Detection of visual concepts and annotation of images using ensembles of trees for hierarchical multi-label classification. In: Recognizing Patterns in Signals, Speech, Images and Videos. (2010) 152–161
19. Lou, X., Hamprecht, F.A.: Structured learning from partial annotations. In: Proceedings of the 29th International Conference on Machine Learning (ICML). (2012) 1519–1526
20. McAuley, J.J., Ramisa, A., Caetano, T.S.: Optimization of robust loss functions for weakly-labeled image taxonomies. International Jurnal of Computer Vision (IJCV) **104** (2013) 343–361
21. Yu, C.N.J., Joachims, T.: Learning structural svms with latent variables. In: Proceedings of the 26th International Conference on Machine Learning (ICML). (2009) 1169–1176
22. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2009) 248–255
23. Van De Sande, K.E., Gevers, T., Snoek, C.G.: Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **32** (2010) 1582–1596
24. Shalev-Shwartz, S., Singer, Y., , Srebro, N.: Primal estimated sub-gradient solver for svm. In: Proceedings of the 24th International Conference on Machine Learning (ICML). (2007)
25. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers. (1999) 61–74
26. Hariharan, B., Zelnik-manor, L., Vishwanathan, S.V.N., Varma, M.: Large scale max-margin multi-label classification with priors. In: Proceedings of the 27th International Conference on Machine Learning (ICML). (2010)