Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Multi-task multi-label multiple instance learning

Yi SHEN, Jian-ping FAN

(*Department of Computer Science, University of North Carolina at Charlotte 28223, USA*)

E-mail: {yshen9, jfan}@uncc.edu

Received Sept. 14, 2010; Revision accepted Sept. 30, 2010; Crosschecked Sept. 14, 2010

Abstract: For automatic object detection tasks, large amounts of training images are usually labeled to achieve more reliable training of the object classifiers; this is cost-expensive since it requires hiring professionals to label large-scale training images. When a large number of object classes come into view, the issue of obtaining a large enough amount of the labeled training images becomes more critical. There are three potential solutions to reduce the burden for image labeling: (1) allowing people to provide the object labels loosely at the image level rather than at the object level (e.g., loosely-tagged images without identifying the exact object locations in the images); (2) harnessing large-scale collaboratively-tagged images that are available on the Internet; and, (3) developing new machine learning algorithms that can directly leverage large-scale collaboratively- or loosely-tagged images for achieving more effective training of a large number of object classifiers. Based on these observations, a multi-task multi-label multiple instance learning (MTML-MIL) algorithm is developed in this paper by leveraging both inter-object correlations and large-scale loosely-labeled images for object classifier training. By seamlessly integrating multi-task learning, multi-label learning, and multiple instance learning, our MTML-MIL algorithm can achieve more accurate training of a large number of inter-related object classifiers (where an object network is constructed for determining the inter-related learning tasks directly in the feature space rather than in the label space). Our experimental results have shown that our MTML-MIL algorithm can achieve higher detection accuracy rates for automatic object detection.

Key words: Object network, Loosely tagged images, Multi-task learning, Multi-label learning, Multiple instance learning

doi:10.1631/jzus.C1001005 **Document code:** A **CLC number:** TP391.4

1 Introduction

For many machine learning tasks, large amounts of training samples are labeled to achieve more reliable classifier training. For automatic object detection tasks, large amounts of training images should be labeled for object classifier training, but identifying the object regions accurately from the images and providing the object labels precisely at the region level could be extremely labor-intensive. On the other hand, people may be willing to provide the object labels loosely at the image level rather than at the region level (e.g., loosely-tagged images without identifying the exact object locations in the images). Rather than requiring people to provide the object labels accurately at the region level, it is more attractive to develop new machine learning algorithms that are able to leverage large-scale loosely-tagged images (i.e., object labels are provided loosely at the image level without identifying the exact locations of the objects in the images) for object classifier training. With these machine learning tools, we can easily harness large-scale social images such as Flickr images, for training a large number of object classifiers, and can effectively tackle the challenging issue of automatic object detection.

Large-scale loosely-tagged images, which are available on the Internet, can have multiple advantages: (1) They can represent various visual properties of the object classes more sufficiently. (2) They can be obtained easily by leveraging the collaborative efforts of a large number of Internet users. Our fundamental belief is that a large group of Internet users with diverse backgrounds can do better job than a small team of professionals, as illustrated by Wikipedia. (3) Both their tags and their visual properties are diverse, thus giving a real-world point of departure for object detection. Therefore, it is very attractive to develop new machine learning frameworks that can leverage the loosely-tagged images for object classifier training.

By treating each image as a bag of instances (image regions), multiple instance learning (MIL) is a well-accepted candidate that can be used to leverage loosely-tagged images for object classifier training, and many wonderful techniques have been developed in the last few decades (Maron and Ratan, 1998; Zhang *et al.*, 2002; Chen *et al.*, 2006; Zhu and Zhang, 2006; Vijayanarasimhan and Grauman, 2008; Zha *et al.*, 2008). All these existing MIL works can be categorized into two groups: (1) given the labels at the bag level (i.e., image level), automatically determining the instance labels and transforming the MIL problem into a traditional supervised learning problem at the instance level; and (2) treating each bag of instances as a single training sample and training the relevant classifiers directly at the bag level. Note that both are attempting to transform the MIL task into a traditional supervised learning task.

For the first group of MIL tools (i.e., determining the instance labels automatically by using the bag labels), most existing works have made some hidden assumptions on the distributions of the positive instances in the positive bags. These hidden assumptions may be incorrect in many real-world practices. For the second group of MIL tools (i.e., learning the bag-level classifiers directly), the bag-level classifiers may not be reliable for automatic object detection because the image instances in the same positive bags could be very diverse (i.e., they may belong to different object classes), and it is unfair to treat each bag of instances (which may belong to different object classes) as a single training sample (one uniform training sample for one certain object class), especially when the image instances in the positive bags are diverse. Thus, it is very attractive to develop new algorithms for assigning multiple labels which are given at the image level to the most relevant image instances automatically.

Another critical issue for the automatic object detection task is that there are a large number of object classes and some of them are dependent. Such inter-object correlations may further bring two critical issues: (1) the computational cost for training a large number of object classifiers grows non-linearly as the number of object classes increases; and (2) the relationships among these object classes cannot be ignored because completely ignoring the inter-object correlations may seriously affect the discrimination power of the object classifiers.

To address the loose tags and inter-object correlations issues jointly, a multi-task multi-label multiple instance learning (MTML-MIL) algorithm is developed to leverage both large-scale loosely-tagged images and the inter-object correlations for achieving more effective training of a large number of inter-related object classifiers. Our MTML-MIL algorithm contains three key components (Fig. 1): (1) automatic tag-instance correspondence identification by determining the instance labels when multiple labels are loosely given at the image level; (2) object network construction for determining the inter-related learning tasks directly in the feature space rather than in the label space; (3) multi-task structured support vector machine (SVM) by incorporating the object network, structured SVM (Tsochantaridis *et al.*, 2005; Joachims *et al.*, 2009), and multi-task learning (Torralba *et al.*, 2004; Evgeniou *et al.*, 2005; Jiang *et al.*, 2007; Fan *et al.*, 2008a) to model the inter-task relatedness more precisely and leverage the inter-object correlations for training a large number of inter-related object classifiers jointly.

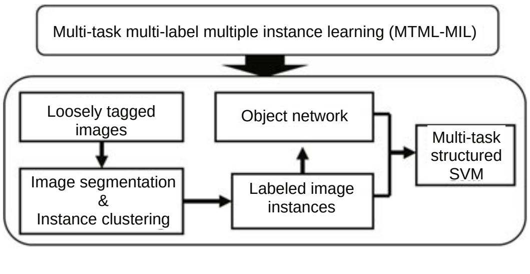
Multi-task multi-label multiple instance learning (MTML-MIL)
Loosely tagged images
Image segmentation & Instance clustering
Object network
Labeled image instances
Multi-task structured SVM

Fig. 1 Flowchart of our MTML-MIL algorithm

2 Related work

Some pioneering work has been done on multiple instance learning (Maron and Ratan, 1998; Zhang *et al.*, 2002; Chen *et al.*, 2006; Vijayanarasimhan and Grauman, 2008). Chen *et al.* (2006) developed an interesting approach called MILES (multiple instance learning via embedded instance selection) to enable region-based image annotation when the labels are available only at the image level. Vijayanarasimhan and Grauman (2008) developed a multi-label multiple instance learning approach to achieve more effective learning from the loosely-labeled images. Maron and Ratan (1998) and Zhang *et al.* (2002) incorporated multiple instance learning (MIL) techniques to learn the object detectors from the loosely-labeled images.

In order to incorporate multi-label images for classifier training, some pioneering work has been done by dividing multi-label learning into a set of binary classification problems or transforming multi-label learning into a label ranking problem (Boutell *et al.*, 2004; Zhu and Zhang, 2006). Boutell *et al.* (2004) addressed the issue of multi-label image annotation by learning a set of binary classifiers. Zhu and Zhang (2006) and Zha *et al.* (2008) integrated multiple instance learning with multi-label learning for scene classification by exploiting the inter-label correlations in the label space. Because classifier training is performed in the feature space rather than in the label space, it is very attractive to develop new algorithms that can directly model the inter-object correlations in the feature space.

Multi-task learning has widely been studied by exploiting the correlations between multiple learning tasks (Torralba *et al.*, 2004; Evgeniou *et al.*, 2005; Kumar and Herbert, 2006; Jiang *et al.*, 2007; Yang *et al.*, 2007; Fan *et al.*, 2008a). Torralba *et al.* (2004) developed a novel JointBoost algorithm to support multi-task learning. Jiang *et al.* (2007) extended the JointBoost algorithm for multi-class concept detection by sharing common kernels. The boosting algorithm can be very sensitive to data noise; thus, it cannot directly be used to leverage the loosely-tagged images with large tag uncertainty for classifier training. Kumar and Herbert (2006) proposed discriminative random fields (DRF) to exploit the inter-patch correlations for object detection. Recently, Yang *et al.* (2007) extended the DRF technique for image/video concept detection. Fan *et al.* (2007; 2008a) constructed concept ontology for identifying the inter-related learning tasks in the concept space and achieved hierarchical training of a large number of inter-related concept classifiers (Fan *et al.*, 2008b).

The statistical rules, such as object co-occurrence context, have been derived from large-scale image collections for supporting context-driven object detection and some pioneering work has been done recently (Liu *et al.*, 2006; Qi *et al.*, 2007; Zha *et al.*, 2008; Tang *et al.*, 2009). Liu *et al.* (2006), Qi *et al.* (2007), and Tang *et al.* (2009) exploited the correlations between the image/video concepts to enhance automatic image/video annotation, and some interesting statistical models have been developed to leverage such inter-concept context for concept classifier training.

3 Multiple instance learning

Multiple instance learning is defined as follows: for each given label of interest, its positive bags refer to those sets of instances that are associated with the given label, in which at least one instance in each bag is responsible to the given label, while its negative bags refer to those sets of instances that are not associated with the given label, and none of these instances are responsible to the given label.

3.1 Instance clustering

In our current implementation, a new scheme is developed for ambiguous image representation by using 'bags of instances': (1) each loosely-tagged image is first partitioned into a set of image regions by using JSEG (Deng and Manjunath, 1999) and multiple segmentations are integrated to obtain more meaningful image regions (image instances) for object detection (Russell *et al.*, 2006); (2) each image region is treated as one instance; and, (3) multi-modal visual features are extracted from each image instance to characterize its various visual properties more sufficiently. These visual features include color histograms, edge histograms, Tamura textures, and region shape.

Our mixture-of-kernels algorithm is further used to integrate multi-modal visual features and their base kernels for instance similarity characterization (Fan *et al.*, 2008a). For two image instances u and v, their visual similarity context is defined as

$$\kappa(u,v)=\sum_{l=1}^{\tau}\alpha_l\kappa_l(u,v),\qquad \sum_{l=1}^{\tau}\alpha_l=1,$$

where τ is the number of feature subsets (i.e., the number of base kernels), $\alpha_l \geq 0$ is the importance factor for the lth base kernel $\kappa_l(u, v)$ and can be obtained automatically.

Because the instance labels are loosely given at the bag level, we need to develop new algorithms for assigning the bag labels to the most relevant instances automatically. To obtain the exact correspondences between the image instances and the bag labels, an instance clustering algorithm is first used to partition the image instances in the positive bags into multiple clusters. To achieve automatic instance clustering, an instance similarity graph is first established where each node represents one image instance and the weights on the edges are used to characterize the visual similarity context between the relevant image instances. Then the affinity propagation algorithm (Frey and Dueck, 2007) is used to partition the instance similarity graph. By passing messages between the nodes, all these image instances in the positive bags are grouped into multiple clusters according to their visual similarity context.

We further define 'relevant clusters' as the instance groups that are responsible to the given label, and the number of relevant clusters and the number of their image instances could be arbitrary. Thus, we still need another step to identify the relevant clusters.

3.2 Relevant cluster identification

When large amounts of positive bags are available, it makes sense that the relevant clusters may have larger sizes. However, this assumption is not always true when the positive instances possess only a small part of the positive bags. When the loosely-tagged images are completely tagged (i.e., each image instance is atomic and has one and only one most relevant label), an irrelevant cluster for one given tag (label) should be a relevant cluster for the other label. For those loosely-tagged images with complete tags, and for each particular label, its relevant cluster should be far away from its negative instances in the irrelevant clusters, and the irrelevant clusters can appear in both the positive and the negative bags.

Given an instance cluster G_i (which could be either a relevant cluster or an irrelevant cluster) in the positive bags Ω and an instance cluster G_j in the negative bags $\bar{\Omega}$, their inter-cluster visual similarity context is defined as

$$\delta(G_i, G_j) = \frac{1}{2|G_i||G_j|} \sum_{u \in G_i} \sum_{v \in G_j} \left[\hat{\kappa}(u, v) + \bar{\kappa}(u, v)\right], \tag{1}$$

where $|G_i|$ and $|G_j|$ are the total numbers of the image instances for the clusters G_i and G_j, respectively, $\kappa(u, v)$ is the pairwise similarity context between image instance u from G_i and image instance v from G_j, $\hat{\kappa}(u, v)$ and $\bar{\kappa}(u, v)$ are the pairwise visual similarity context between image instance u from G_i and image instance v from G_j by using the kernel weights for the instance clusters G_i and G_j, respectively. All these kernel weights are automatically provided during the instance clustering process.

The best-matched cluster pair (G_i, G_k) between the positive and the negative bags is determined by

$$\bar{\delta}(G_i, G_k) = \max\left\{\delta(G_i, G_j) | G_i \in \Omega, G_j \in \bar{\Omega}\right\}. \tag{2}$$

This process for best-matched cluster pair determination is continued until all these instance clusters (which could be either relevant clusters or irrelevant clusters) in the positive bags have found the best-matched negative cluster in the negative bags.

The instance clusters in the positive bags are then partitioned into two groups according to their pairwise inter-cluster visual similarity context with the irrelevant clusters in the negative bags: positive groups versus negative groups. The relevant clusters in the positive bags should be far away from the negative clusters in the negative bags (i.e., with small values of $\bar{\delta}(\cdot, \cdot)$) (Maron and Ratan, 1998; Fan *et al.*, 2010). In contrast, the irrelevant clusters in the positive bags may be close to the negative clusters in the negative bags (i.e., with large values of $\bar{\delta}(\cdot, \cdot)$), and they can be assigned into the negative groups.

Due to the problem of incomplete tagging, i.e., some image components may not be tagged, some irrelevant clusters in the positive bags may not have strong correlations with the negative clusters in the negative bags (i.e., they may belong to different object classes). Thus, those irrelevant clusters may also be far from the irrelevant clusters in the negative bags. However, those irrelevant clusters in the positive bags should be small in size as compared with the relevant clusters. Thus, they can further be separated from the relevant clusters according to their size differences. The relevant clusters in the positive bags may have significant differences with the irrelevant clusters in the positive bags, either on their correlations with the irrelevant clusters in the negative bags or on their sizes. Thus, it is easy to separate the relevant clusters from the irrelevant clusters in the positive bags. When the relevant clusters are identified from the irrelevant clusters in the positive bags, the given tag is treated as the ground-truth label for all their image instances in the relevant clusters. The F score is used to evaluate the performance of our tag-instance correspondence identification algorithm and some experimental results are given in Fig. 2. By performing inter-cluster correlation analysis, our tag-instance correspondence identification algorithm can support more effective multiple instance learning by determining the instance labels more precisely.

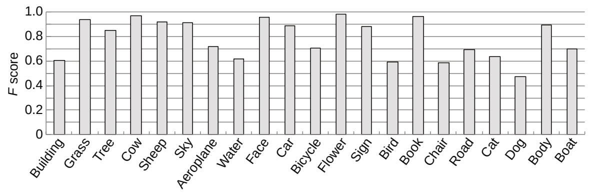

Fig. 2 The *F* scores for our multiple instance learning (instance label identification) algorithm on the MSRC image set using ground truth segmentation

4 Object network

After the instance labels are determined automatically, we can generate large-scale image instances for each tag (label) of interest and each label is used to interpret the semantics of one certain object class. Thus, these image instances can further be used to determine the inter-object correlations and construct an object network. Our object network consists of two key components: object classes and their inter-object correlations.

Given two object classes C_i and C_j, their inter-object visual similarity context $\gamma(C_i, C_j)$ is determined by

$$\gamma(C_i, C_j) = \frac{1}{2|C_i||C_j|}\sum_{u\in C_i}\sum_{v\in C_j}[\hat{\kappa}(u,v) + \bar{\kappa}(u,v)], \tag{3}$$

where $|C_i|$ and $|C_j|$ are the total numbers of the image instances for the object classes C_i and C_j respectively, $\hat{\kappa}(u,v)$ and $\bar{\kappa}(u,v)$ are the kernel-based similarity context between two image instances u and v by using the kernel weights for the object classes C_i and C_j, respectively. All these kernel weights are automatically provided during the instance clustering process.

The co-occurrence correlation $\rho(C_i, C_j)$ between two object classes C_i and C_j is defined as

$$\rho(C_i, C_j) = -P(C_i, C_j)\log\frac{P(C_i, C_j)}{P(C_i) + P(C_j)}, \tag{4}$$

where $P(C_i, C_j)$ is the co-occurrence probability for two object classes C_i and C_j in our image collections, and $P(C_i)$ and $P(C_j)$ are the occurrence probabilities for C_i and C_j, respectively.

Given two object classes C_i and C_j, their visual similarity context $\gamma(C_i, C_j)$ and their co-occurrence correlation $\rho(C_i, C_j)$ are first normalized into the same interval. The inter-object correlation $\phi(C_i, C_j)$ between C_i and C_j is finally defined as

$$\phi(C_i, C_j) = \eta \cdot \bar{\gamma}(C_i, C_j) + (1 - \eta) \cdot \bar{\rho}(C_i, C_j), \tag{5}$$

where η is the weighting factor and it is determined through cross-validation, and $\bar{\gamma}(C_i, C_j)$ and $\bar{\rho}(C_i, C_j)$ are the normalized visual similarity context and co-occurrence correlation, respectively. The weighting factor is set as $\eta = 0.6$ in our current implementation because the visual similarity context is more important than the co-occurrence correlations for inter-object correlation characterization.

Some experimental results for the inter-object correlations $\phi(\cdot,\cdot)$ are given in Table 1. Part of our object network for our image sets is shown in Fig. 3. Our object network may have multiple advantages: (1) It can characterize the inter-object correlations explicitly and provide a good environment to identify the inter-related learning tasks directly in the feature space. (2) It can provide a good environment to integrate the training instances from multiple inter-related object classes for training their inter-related classifiers jointly and can bring more powerful inference schemes to enhance their discrimination and adaptation power significantly.

5 Multi-task structured learning

When a large number of object classes come into view, direct modeling of the inter-object correlations over the whole graph (object network) becomes computationally intractable. In this work, a multi-task

Table 1 Inter-object correlations

Object pair	ϕ	Object pair	ϕ	Object pair	ϕ	Object pair	ϕ
road-building	0.72	road-sign	0.59	boat-aeroplane	0.73	mountain-road	0.62
sign-building	0.63	sheep-horse	0.66	body-dog	0.83	body-cat	0.79
cat-grass	0.03	tree-cat	0.72	book-grass	0.06	sky-book	0.02
boat-sky	0.06	aeroplane-mountain	0.43	mountain-water	0.75	body-sky	0.06
grass-building	0.04	horse-cow	0.31	cat-dog	0.85	car-bicycle	0.53
bird-aeroplane	0.54	car-aeroplane	0.55	boat-grass	0.02	dog-cow	0.51

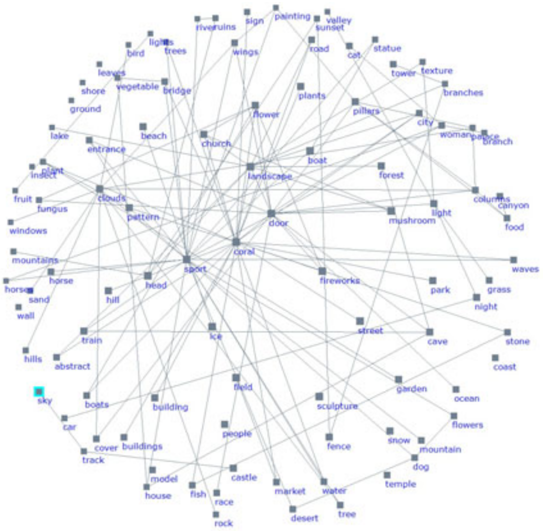

Fig. 3 Part of our object network. Each object class is linked with multiple most relevant object classes with larger values of inter-object correlation

structured SVM scheme is developed by incorporating the first-order nearest neighbors (i.e., clique for each object class on the object network), multi-task learning, and structured SVM to leverage the inter-object correlations to achieve more accurate training of a large number of inter-related object classifiers.

For a given object class ‘bridge’, its first-order nearest neighbors on the object network are shown in Fig. 4. One can observe that the first-order nearest neighbors on the object network are strongly correlated and their training instances may share some common visual properties in the feature space. It is not appropriate to train the classifiers for these inter-related object classes independently. To leverage the inter-object correlations for training the inter-related classifiers jointly, multi-task learning is used in our structured learning framework. The idea behind multi-task learning is that if multiple inter-related learning tasks share a common prediction component, this common prediction component can be estimated more accurately by considering these inter-related learning tasks together (Evgeniou *et al.*, 2005; Fan *et al.*, 2008a). Structured SVM (Tsochantaridis *et al.*, 2005; Joachims *et al.*, 2009) has been proposed to exploit the inter-label correlations in the label space for supporting structure prediction.

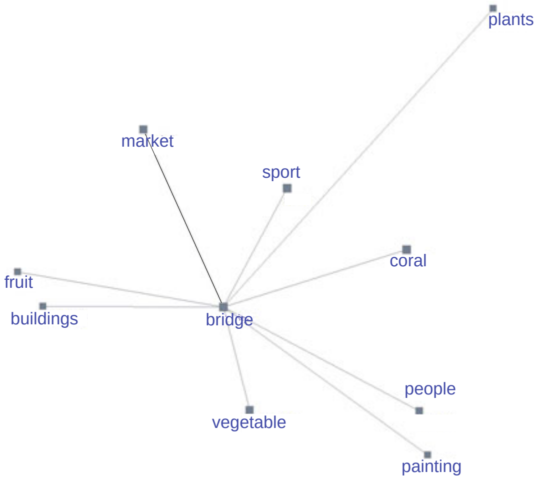

Fig. 4 The inter-related object classes for the object class ‘bridge’

In this work, a multi-task structured SVM scheme is developed by incorporating the object network, multi-task learning, and structured SVM to enhance the discrimination power of a large number of inter-related object classifiers: (1) The object network is used to identify the inter-related learning tasks directly in the feature space, e.g., training multiple inter-related object classifiers jointly; (2) The inter-task relatedness is characterized explicitly by using the strengths of the inter-object correlations $\phi(\cdot,\cdot)$, and a common prediction component is used to model the inter-task relatedness, which is shared among these inter-related object classifiers; (3) The structured SVM is integrated with multi-task learning to model the inter-task relatedness more precisely and estimate the common prediction component more accurately. By seamlessly integrating multi-task learning with structured SVM, our multi-task structured SVM algorithm can exploit the inter-object correlations explicitly in the input space (i.e., the common space for classifier training and object detection). Thus, it can provide a new approach for inter-related classifier training and address the issue of multiple tags more effectively.

In our multi-task structured SVM scheme, a common regularization term W_0 of the SVM classifier is used to model the inter-task relatedness among multiple SVM classifiers for the inter-related object classes. Given an object class C_j, its classifier is defined as

$$f_{C_j}(x) = \sum_{C_t \in \mathcal{T}} \gamma_t (W_0 + V_t)^{\mathrm{T}} \varPhi_t(x), \tag{6}$$

where W_0 is the common regularization term shared among the classifiers for multiple inter-related object classes centered at C_j (Fig. 4), V_t is the individual regularization term for the classifier between the given object class C_j and its neighbor C_t, γ_t is the weight related to how the object class C_t contributes to the classification of the object class C_j, and $\varPhi_t(x)$ is a sign indicating whether x could be mapped to some kernel space.

W_0 can be estimated more reliably by minimizing their joint objective function J for T inter-related learning tasks:

$$\begin{aligned} J &= \frac{1}{2}\left(\|W_0\|^2 + \sum_{t=1}^{T} \lambda_t \|V_t\|^2\right) \\ &\quad + c_0 \sum_{t=1}^{T} \sum_{i=1}^{n_j} \xi_{ti} + \sum_{t=1}^{T} c_t \sum_{i=1}^{n_t} \eta_{ti}, \end{aligned} \tag{7}$$

where $\xi_{ti} \geq 0$, $\eta_{ti} \geq 0$, and n_j and n_t are the total numbers of training instances for the object classes C_j and C_t, respectively.

To solve the joint optimization problem, we use the Lagrangian Principle. We add a dual set of variables, one for each constraint, and obtain the Lagrangian L of the optimization problem:

$$\begin{aligned} L &= J - \sum_{t=1}^{T} \sum_{i}^{n_j} \beta_{ti} \left(\langle W_0 + V_t, \varPhi_t(x_{ji}) \rangle + \xi_{ti} - 1 \right) \\ &\quad + \sum_{t=1}^{T} \sum_{i=1}^{n_t} \overline{\beta}_{ti} \left(\langle W_0 + V_t, \varPhi_t(x_{ti}) \rangle - \eta_{ti} + 1 \right) \\ &\quad - \sum_{t=1}^{T} \sum_{i=1}^{n_j} \sigma_{ti} \xi_{ti} - \sum_{t=1}^{T} \sum_{i=1}^{n_t} \overline{\sigma}_{ti} \eta_{ti}. \end{aligned}$$

We now seek a saddle point of the Lagrangian L. For example, the partial difference of L satisfies

$$\begin{aligned} \frac{\partial L}{\partial W_0} &= W_0 - \sum_{t=1}^{T} \sum_{i=1}^{n_j} \beta_{ti} \varPhi_t(x_{ji}) \\ &\quad + \sum_{t=1}^{T} \sum_{i=1}^{n_t} \overline{\beta}_{ti} \varPhi_t(x_{ti}), \\ \frac{\partial L}{\partial V_t} &= \lambda_t V_t - \sum_{i=1}^{n_j} \beta_{ti} \varPhi_t(x_{ji}) + \sum_{i=1}^{n_t} \overline{\beta}_{ti} \varPhi_t(x_{ti}), \\ \frac{\partial L}{\partial \xi_{ti}} &= c_0 - \beta_{ti} - \sigma_{ti}, \frac{\partial L}{\partial \eta_{ti}} = c_t - \overline{\beta}_{ti} - \bar{\sigma}_{ti}, \end{aligned}$$

and we obtain

$$\begin{aligned} W_0 &= \sum_{t=1}^{T} \sum_{i=1}^{n_j} \beta_{ti} \varPhi_t(x_{ji}) - \sum_{t=1}^{T} \sum_{i=1}^{n_t} \overline{\beta}_{ti} \varPhi_t(x_{ti}), \\ V_t &= \frac{1}{\lambda_t} \left(\sum_{i=1}^{n_j} \beta_{ti} \varPhi_t(x_{ji}) - \sum_{i=1}^{n_t} \overline{\beta}_{ti} \varPhi_t(x_{ti}) \right), \\ c_0 &= \beta_{ti} + \sigma_{ti}, \quad c_t = \overline{\beta}_{ti} + \overline{\sigma}_{ti}. \end{aligned}$$

The dual form of the problem is then simplified as

$$\begin{aligned} L &= \sum_{t=1}^{T} \sum_{i=1}^{n_j} \beta_{ti} + \sum_{t=1}^{T} \sum_{i=1}^{n_t} \overline{\beta}_{ti} \\ &\quad - \frac{1}{2} \left(\|W_0\|^2 + \sum_{t=1}^{T} \lambda_t \|V_t\|^2 \right). \end{aligned}$$

Given the training image instances for T inter-related object classes on the object network, the margin maximization process for joint classifier training is then transformed into a quadratic problem:

$$\begin{aligned}\max_{\beta_{ti},\bar{\beta}_{ti}} L \;=\; & \sum_{t=1}^{T}\sum_{i=1}^{n_j}\beta_{ti}+\sum_{t=1}^{T}\sum_{i=1}^{n_t}\overline{\beta}_{ti} \\ & -\frac{1}{2}\Bigg[\sum_{t,s=1}^{T}\sum_{i=1}^{n_j}\sum_{l=1}^{n_l}\beta_{ti}\beta_{sl}K_{ts}(x_{ji},x_{jl}) \\ & -\sum_{t,s=1}^{T}\sum_{i=1}^{n_j}\sum_{l=1}^{n_s}\beta_{ti}\overline{\beta}_{sl}K_{ts}(x_{ji},x_{sl}) \\ & -\sum_{t,s=1}^{T}\sum_{i=1}^{n_t}\sum_{k=1}^{n_j}\overline{\beta}_{ti}\beta_{sk}K_{ts}(x_{ti},x_{kj}) \\ & +\sum_{t,s=1}^{T}\sum_{i=1}^{n_t}\sum_{k=1}^{n_s}\overline{\beta}_{ti}\overline{\beta}_{sk}K_{ts}(x_{ti},x_{sk}) \\ & +\sum_{t=1}^{T}\frac{1}{\lambda_t}\Big(\sum_{i,l=1}^{n_j}\beta_{ti}\beta_{tl}K_t(x_{ji},x_{jl}) \\ & -2\sum_{i=1}^{n_j}\sum_{l=1}^{n_t}\beta_{ti}\overline{\beta}_{tl}K_t(x_{ji},x_{jl}) \\ & +\sum_{i,l=1}^{n_t}\overline{\beta}_{ti}\overline{\beta}_{tl}K_t(x_{ti},x_{tl})\Big)\Bigg]\end{aligned}$$

subject to: $\forall i,\ \ \forall t,\ \ \beta_{ti}\geq 0,\ \ \overline{\beta}_{ti}\geq 0.$

To deal with the structured prediction problem, it is very attractive to construct a joint kernel function that is better suited to joint-space support estimation. In this work, a tensor product is incorporated to define the joint kernel $\kappa((x_i,y_i),(x_j,y_j))$ as

$$\kappa((x_i,y_i),(x_j,y_j))=\kappa(x_i,x_j)\kappa_s(y_i,y_j), \tag{8}$$

where $\kappa(x_i,x_j)$ is the kernel for the similarity between x_i and x_j, and $\kappa_s(y_i,y_j)$ is the semantic kernel to characterize the semantic similarity context between the labels y_i and y_j of two object classes (i.e., inter-class correlation on the label space).

By learning from a joint training instance set $\Omega=\{(x_{it},y_{it})|i=1,2,...,n;\ t=1,2,...,T\}$ for T inter-related object classes on the object network, the classifier for the given object class C_j can be determined as

$$\begin{aligned}& f_{C_j}(x)= \\ & \sum_{h,t=1}^{T}\gamma_t\kappa_s(t,h)\left(\sum_{i=1}^{n_j}\beta_{hi}\kappa(x_{ji},x)-\sum_{i=1}^{n_h}\overline{\beta}_{hi}\kappa(x_{hi},x)\right) \\ & \quad +\sum_{t=1}^{T}\frac{\gamma_t}{\lambda_t}\left(\sum_{i=1}^{n_j}\beta_{ti}\kappa(x_{ji},x)-\sum_{i=1}^{n_t}\overline{\beta}_{ti}\kappa(x_{ti},x)\right).\end{aligned}$$

One can observe that our classifiers for inter-related object classes consist of two components: (1) individual prediction component, and (2) common prediction component.

By learning two different sets of the weights β and $\bar{\beta}$ for the training instances simultaneously, our multi-task structured SVM algorithm can automatically establish two independent decision boundaries for both the common prediction component (shared among the inter-related discriminant functions) and the individual prediction component of the discriminant function for each particular object class. The training instances, which are used to construct the common prediction component for multiple inter-related object classifiers (i.e., support vectors with large values of $\bar{\beta}$), are less important for the individual prediction components for these inter-related object classifiers (i.e., with smaller values or even zero values of β).

By integrating the training instances for multiple inter-related object classes to learn a common prediction component, and separating it from their individual prediction components, our multi-task structured SVM algorithm can significantly enhance the discrimination power and the generalization ability of the inter-related classifiers. When all these inter-related classifiers are available, they are used for detecting the objects from the images. Some object detection results are given in Fig. 5.

6 Algorithm evaluation

Our experiments were performed on two sets of loosely-tagged images: (1) 3 814 MSRC image instances (image regions) (http://research.microsoft.com/) and 30k Corel images (Fan *et al.*, 2004); (2) one million loosely-tagged images which are collected from Flickr (http://www.flickr.com; Fan *et al.*, 2010). Because MSRC images and Corel images are easy to obtain, we use them as our test image sets, so that other researchers can easily assess the real performances of our algorithms. Moreover, Flickr allows us to collect large-scale and realistic loosely-tagged images. It is very attractive to use such realistic loosely-tagged images for developing new algorithms that can tackle the issue of tag uncertainty and learn the object classifiers reliably. Thus, the Flickr image set was used as the training image set in our

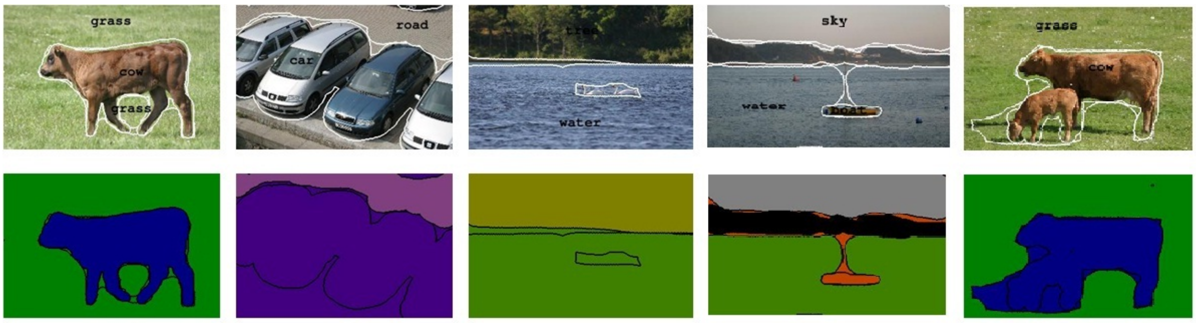

Fig. 5 Some object detection results using our MTML-MIL algorithm

experiments.

To assess the effectiveness of our proposed algorithms, the algorithm evaluation work focused on comparing the performance differences between various approaches for object classifier training: (1) our MTML-MIL algorithm versus the structured SVM algorithm by exploiting the inter-label correlations in the output space (Tsochantaridis *et al.*, 2005; Joachims *et al.*, 2009); (2) our MTML-MIL algorithm versus the MILES without exploiting the inter-label correlations explicitly (Chen *et al.*, 2006), and (3) our MTML-MIL algorithm versus the multi-label MIL (MLMIL) technique developed by Zha *et al.* (2008). In this work, AUC (area under the receiver operating characteristic curve) was adopted to evaluate the classification performance (Hanley and Mcneil, 1982), describing the probability that a randomly chosen positive image is ranked higher than a randomly chosen negative image.

Using the same set of multi-modal visual features for image content representation, we compared the performance differences between two approaches to integrating loosely-tagged images for object classifier training: the MILES approach (Chen *et al.*, 2006) versus our MTML-MIL algorithm (Figs. 6–8). Our MTML-MIL algorithm significantly improved the accuracy for detecting the inter-related object classes. The significant improvement on the detection accuracy benefits from two components:

1. The object classifiers for the inter-related object classes are trained jointly by leveraging their inter-object correlations for object classifier training. Thus, our MTML-MIL algorithm can address the issue of multiple tags more effectively. Our MTML-MIL algorithm can address the issue of visual ambiguity more effectively by learning the inter-related classifiers for the inter-related object classes jointly. It can also enhance the discrimination and adaptation power of the inter-related object classifiers significantly by learning from the training instances for other inter-related object classes on the object network. Incorporating the training instances from other inter-related object classes for classifier training will significantly enhance the generalization ability of their classifiers, especially when the available training instances for the given object class may not be representative for large amounts of unseen test images. In contrast, MILES does not consider the inter-object (inter-label) correlations explicitly, which may result in lower accuracy rates for detecting some inter-related object classes.

2. Through an instance clustering and inter-cluster correlation analysis process, our MTML-MIL algorithm can address the issue of loose tags more effectively, which is crucial for leveraging the loosely-tagged images for object classifier training.

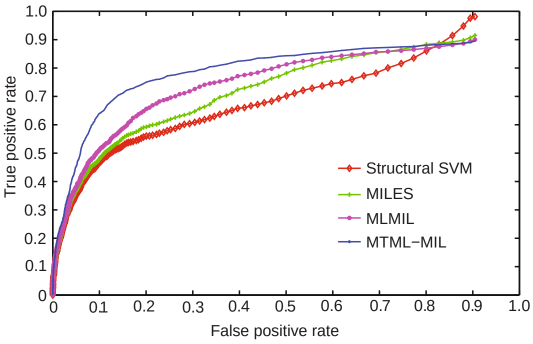

Fig. 6 ROC curves for performance comparison between our MTML-MIL algorithm and other most relevant algorithms

Two existing approaches, the structured SVM algorithm (Tsochantaridis *et al.*, 2005; Joachims *et al.*, 2009) and the MLMIL algorithm (Zha *et al.*, 2008), have considered the inter-label (inter-object) correlations for classifier training. Our MTML-MIL algorithm is somewhat similar in spirit to these two approaches for object classifier training, but significantly different in multiple important aspects. Compared with both the structured SVM algorithm and the MLMIL algorithm, our MTML-MIL algorithm has multiple advantages: (1) It can explicitly model the inter-object correlations and the inter-task relatedness in the inter-related object classifiers (i.e., common regularization component W_0), which may provide a good environment to leverage the inter-object correlations and the inter-task relatedness for inter-related classifier training and enhance their discrimination power significantly; (2) It can save the cost for object detection by exploiting the inter-object correlations in the feature space. In contrast, both the structured SVM algorithm and the MLMIL algorithm model the inter-object correlations in the output (label) space rather than in the input (feature) space. In average, our MTML-MIL algorithm has better performance than the structured SVM algorithm and the MLMIL algorithm (Table 2).

Table 2 Average AUC (area under the ROC curve) scores on MSRC

Algorithm	Average AUC score
Structural SVM	0.6952
MILES	0.7304
MLMIL	0.7539
Our MTML-MIL	0.7965

By generalizing the multi-class SVM algorithm, the structured SVM algorithm focuses on supporting structural output prediction for a large number of SVM object classifiers, e.g., modeling the inter-object context in the output space and exploiting the inter-object context in the label space rather than in the feature space. In contrast, our MTML-MIL algorithm can directly model the inter-task relatedness in the feature (or input) space for classifier training and testing, and explicitly exploit the inter-object correlations to achieve a more effective training of a large number of inter-related object classifiers. As shown in Figs. 6–8, our MTML-MIL algorithm had a better performance (higher AUC rates) as compared with the structured SVM algorithm.

We also compared the performances of our MTML-MIL algorithm and the MLMIL technique. By explicitly modeling the inter-task relatedness in the feature space, our MTML-MIL algorithm can provide a good environment to leverage the inter-object correlations and the inter-task relatedness for inter-related object classifier training, which may result in higher discrimination powers for a large number of inter-related object classifiers. Compared with the MLMIL algorithm, our MTML-MIL algorithm can achieve very competitive performance (Figs. 6–8). For some object classes, the MLMIL algorithm obtained a little bit higher accuracy rates, but our MTML-MIL algorithm achieved on average a better accuracy rate (Table 2) for the MSRC image set with 21 object classes (http://research.microsoft.com/; Zha *et al.*, 2008).

Besides MSRC images, we also compared the performances of our MTML-MIL algorithm and the normal SVM and MLMIL on 30k Corel images for 98 concept categories (Fan *et al.*, 2004), where the original images were associated with totally 5k tags and 98 frequent object tags were chosen to be our distinguished concept categories. On average, each Corel image may have more than 10 image regions (instances). Thus, there were 300k instances, which are too large to handle effectively (i.e., to pre-compute and store the kernel-based similarity matrix for all these 300k instances) by a single PC. When the size of image instances reaches one million, some existing techniques for kernel-based image clustering and SVM classifier training (such as LIBSVM (Fan *et al.*, 2005)) may take years to run out an $O(m^3)$ algorithm on a single PC. Thus, it is very attractive to develop a distributed computing framework to enable kernel-based image clustering and SVM classifier training.

To address the issue of computational complexity reduction more effectively, two approaches were used: (1) only the first-order neighbors and their image instances were integrated for inter-related classifier training; and (2) a cascade learning framework was incorporated for training the SVM classifiers in a distributed way (Graf *et al.*, 2004). The idea of the cascade learning framework was to equally divide the image instances and iteratively aggregate the final SVM classifiers. Given an object class on the object correlation network, all the positive image instances for the object class and its first-order neighbors are integrated for joint classifier training. To enhance the discrimination power of a large number of inter-related object classifiers, the cascade framework was

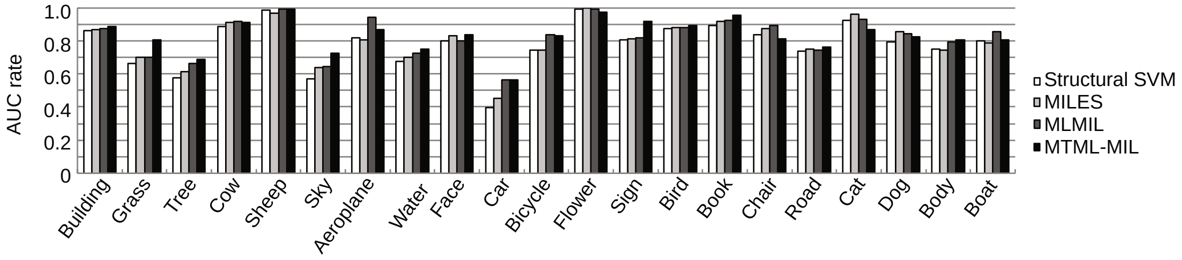

Fig. 7 Performance comparison on AUC (area under the ROC curve) rates between our MTML-MIL algorithm and other most relevant algorithms

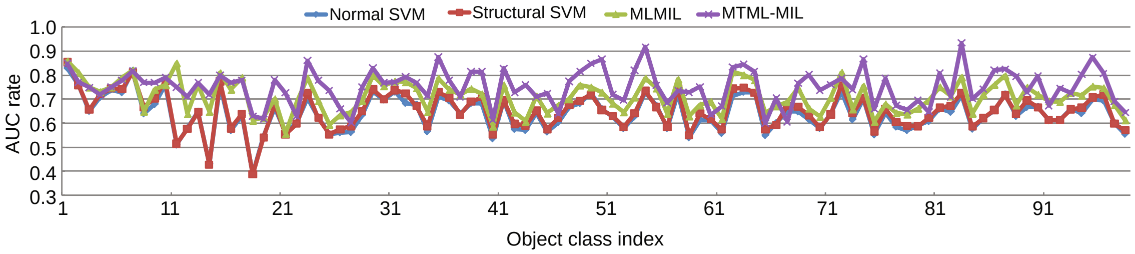

Fig. 8 Performance comparison on 30k Corel images with 98 object classes

used to sample the positive instances from other object classes that are not the first-order neighbors of the given object class.

The original version of the MILES algorithm focuses on solving a 1-norm SVM problem over all the training instances. It is too time-consuming to leverage the MILES algorithm for training the classifiers for a large number of inter-related object classes. Moreover, it is not easy to adapt the MILES algorithm to the cascade learning framework. Thus, it is hard to obtain the performance of the MILES algorithm on a large number of inter-related object classes using a large Corel image set for classifier training. Based on this observation, we did not compare the performance difference between our MTML-MIL algorithm and the MILES algorithm on the 30k Corel image set with 98 object classes. As shown in Fig. 8 and Table 3, our proposed algorithm had a significant improvement over the normal SVM (Fan *et al.*, 2005) and obtained on average a little higher AUC score than MLMIL (Zha *et al.*, 2008).

Table 3 Average AUC (area under the ROC curve) scores on Corel images

Algorithm	Average AUC score
Normal SVM	0.6438
Structural SVM	0.6512
MLMIL	0.7236
Our MTML-MIL	0.7549

7 Conclusions

For an automatic object detection task, a multi-task multi-label multiple instance learning (MTML-MIL) framework is developed to leverage both the inter-object correlations and large-scale loosely-tagged images for achieving a more effective training of a large number of inter-related object classifiers. By identifying the correspondences between multiple tags at the image level and bag of instances (multiple image regions) automatically, our MTML-MIL algorithm can automatically transform the bag labels into the instance labels for harnessing large-scale loosely-tagged images for object classifier training. By incorporating the object network for determining the inter-related learning tasks directly in the feature space, rather than in the label space, our MTML-MIL algorithm can seamlessly integrate structured SVM and multi-task learning to model inter-related object classifiers and enhance their discrimination power significantly. Experimental results on a large number of object classes have provided very positive results.

References

Boutell, M.R., Luo, J., Shen, X., Brown, C.M., 2004. Learning multi-label scene classification. *Pattern Recogn.*, **37**(9):1757-1771. [doi:10.1016/j.patcog.2004.03.009]

Chen, Y., Bi, J., Wang, J.Z., 2006. MILES: multiple instance learning via embedded instance selection. *IEEE Trans. PAMI*, **28**(12):1931-1947. [doi:10.1109/TPAMI.2006.248]

Deng, Y., Manjunath, B.S., 1999. Color Image Segmentation. IEEE CVPR, p.2446-2451. [doi:10.1109/CVPR.1999.784719]

Evgeniou, T., Micchelli, C.A., Pontil, M., 2005. Learning multiple tasks with kernel methods. *J. Mach. Learn. Res.*, **6**:615-637.

Fan, J., Gao, Y., Luo, H., 2004. Multi-Level Annotation of Natural Scenes Using Dominant Image Components and Semantic Image Concepts. ACM Multimedia, p.540-547. [doi:10.1145/1027527.1027660]

Fan, J., Luo, H., Gao, Y., Jain, R., 2007. Incorporating concept ontology for hierarchical video classification, annotation and visualization. *IEEE Trans. Multimedia*, **9**(5):939-957. [doi:10.1109/TMM.2007.900143]

Fan, J., Gao, Y., Luo, H., 2008a. Integrating concept ontology and multi-task learning to achieve more effective classifier training for multi-level image annotation. *IEEE Trans. Image Process.*, **17**(3):407-426. [doi:10.1109/TIP.2008.916999]

Fan, J., Gao, Y., Luo, H., Jain, R., 2008b. Mining multi-level image semantics via hierarchical classification *IEEE Trans. Multimedia*, **10**(1):167-187. [doi:10.1109/TMM.2007.911775]

Fan, J., Shen, Y., Zhou, N., Gao, Y., 2010. Harvesting Large-Scale Weakly-Tagged Image Databases from the Web. IEEE CVPR, p.802-809. [doi:10.1109/CVPR.2010.5540135]

Fan, R., Chen, P., Lin, C.J., 2005. Working set selection using the second order information for training SVM. *J. Mach. Learn. Res.*, **6**:1889-1918.

Frey, B.J., Dueck, D., 2007. Clustering by passing messages between data points. *Science*, **315**(5814):972-976. [doi:10.1126/science.1136800]

Graf, H.P., Cosatto, E., Bottou, L., Durdanovic, I., Vapnik, V., 2004. Parallel Support Vector Machines: the Cascade SVM. NIPS, p.1-8.

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**(1):29-36.

Jiang, W., Chang, S.F., Loui, A., 2007. Context-Based Concept Fusion with Boosted Conditional Random Fields. IEEE ICASSP, p.949-952. [doi:10.1109/ICASSP.2007.366066]

Joachims, T., Finley, T., Yu, C., 2009. Cutting-plane training of structural SVMs. *Mach. Learn.*, **77**(1):27-59. [doi:10.1007/s10994-009-5108-8]

Kumar, S., Herbert, M., 2006. Discriminative random fields. *Int. J. Comput. Vis.*, **68**(2):179-201. [doi:10.1007/s11263-006-7007-9]

Liu, J., Li, M., Ma, W.Y., Liu, Q., Lu, H., 2006. An Adaptive Graph Model for Automatic Image Annotation. ACM Multimedia Workshop on MIR, p.61-70. [doi:10.1145/1178677.1178689]

Maron, O., Ratan, A.L., 1998. Multiple-Instance Learning for Natural Scene Classification. ICML, p.341-349.

Qi, G.J., Hua, X.S., Rui, Y., Tang, J., Mei, T., Zhang, H.J., 2007. Correlative Multi-Label Video Annotation. ACM Multimedia, p.17-26. [doi:10.1145/1291233.1291245]

Russell, B., Efros, A., Sivic, J., Freeman, W., Zisserman, A., 2006. Using Multiple Segmentations to Discover Objects and Their Extent in Image Collections. IEEE CVPR, p.1605-1614. [doi:10.1109/CVPR.2006.326]

Tang, J., Hua, X., Wang, M., Gu, Z., Qi, G., Wu, X., 2009. Correlative linear neighborhood propagation for video annotation. *IEEE Trans. SMC*, **39**(2):409-416. [doi:10.1109/TSMCB.2008.2006045]

Torralba, A., Murphy, K.P., Freeman, W.T., 2004. Sharing Features: Efficient Boosting Procedures for Multiclass Object Detection. IEEE CVPR, p.762-769. [doi:10.1109/CVPR.2004.1315241]

Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y., 2005. Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, **6**:1453-1484.

Vijayanarasimhan, S., Grauman, K., 2008. Keywords to Visual Categories: Multiple-Instance Learning for Weakly Supervised Object Categorization. IEEE CVPR, p.1-8. [doi:10.1109/CVPR.2008.4587632]

Yang, J., Liu, Y., Ping, E.X., Hauptmann, A.G., 2007. Harmonium Models for Semantic Video Representation and Classification. SIAM Conf. on Data Mining, p.1-12.

Zha, Z., Hua, X.S., Mei, T., Wang, J., Qi, G.J., Wang, Z., 2008. Joint Multi-Label Multi-Instance Learning for Image Classification. IEEE CVPR, p.1-8. [doi:10.1109/CVPR.2008.4587384]

Zhang, Q., Yu, W., Goldman, S.A., Fritts, J.E., 2002. Content-Based Image Retrieval Using Multiple-Instance Learning. ICML, p.682-689.

Zhu, Z.H., Zhang, M.L., 2006. Multi-Instance Multi-Label Learning with Application to Scene Classification. NIPS, p.1609-1616.