Two-Dimensional Multilabel Active Learning with an Efficient Online Adaptation Model for Image Classification

Guo-Jun Qi, Xian-Sheng Hua, *Member*, *IEEE*, Yong Rui, *Senior Member*, *IEEE*, Jinhui Tang, *Student Member*, *IEEE*, and Hong-Jiang Zhang, *Fellow*, *IEEE*

Abstract—Conventional active learning dynamically constructs the training set only along the sample dimension. While this is the right strategy in binary classification, it is suboptimal for multilabel image classification. We argue that for each selected sample, only some effective labels need to be annotated while others can be inferred by exploring the label correlations. The reason is that the contributions of different labels to minimizing the classification error are different due to the inherent label correlations. To this end, we propose to select sample-label pairs, rather than only samples, to minimize a multilabel Bayesian classification error bound. We call it two-dimensional active learning because it considers both the sample dimension and the label dimension. Furthermore, as the number of training samples increases rapidly over time due to active learning, it becomes intractable for the offline learner to retrain a new model on the whole training set. So we develop an efficient online learner to adapt the existing model with the new one by minimizing their model distance under a set of multilabel constraints. The effectiveness and efficiency of the proposed method are evaluated on two benchmark data sets and a realistic image collection from a real-world image sharing Web site—Corbis.

Index Terms—Active learning, online adaption, multilabel classification, image annotation.

1 Introduction

The goal of image classification is to assign a set of labels to images based on their semantic content. In most existing approaches, image classification has been formulated as either multiclass or multilabel problem. As multiclass problem, each image can be labeled by one and only one class. An example under such a classification setting is the Caltech 101 [1] annotation in which each image in this data set is classified as only one object category. However, in most real-world problems, multiple labels can be assigned to an image. For example, in many online image sharing Web sites (e.g., Flickr, Picasa, and Yahoo! Gallery), most of the images have more than one tags manually labeled by users. This classification setting results in a multilabel problem, which is more complex and challenging compared to a multiclass problem. In this paper, we will focus on image classification under this multilabel setting. Specifically, we will use active learning as the tool, and extend it from a one-dimensional sample-centric approach to a two-dimensional joint sample-label-centric approach for multilabel image classification. We further propose an online multilabel classification algorithm, which can incrementally update the classification model once new image samples are selected by the proposed active learning strategy. Such an online algorithm can avoid retraining the multilabel model so that it is computationally efficient to adapt the classifier to capture the semantic changes of the online image content.

In traditional classification scenarios [2], [3], [4], a batch of training images is first *statically* annotated by a set of semantic labels and then they are used to train a classifier. However, in many online applications (e.g., the image sharing Web sites), users can *dynamically* upload new images, which have significant difference from the existing image collections due to the changes of the user-focuses and semantic "concept drift" in the low-level feature space. Moreover, these images can often be annotated by multiple labels simultaneously and it poses more challenges to handle these multilabel image sets. To deal with this online setting, traditional approaches are restricted by the following two problems:

- To adapt the existing classifier to the "concept" drift over time, we must manually collect the multilabel ground truth from the newly acquired images. As is well known, it is labor intensive and subject to annotation errors, especially when these image sets

- *G.-J. Qi is with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Everitt Laboratory, MC-702, 1406 W. Green St., Urbana, IL 61801-2918. E-mail: qi4@illinois.edu.*
- *X.-S. Hua is with the Internet Media Group, Microsoft Research Asia, 5F, Sigma Center, 49 Zhichun Road, Haidian District, Beijing 100190, P.R. China. E-mail: xshua@microsoft.com.*
- *Y. Rui is with the Microsoft China R&D Group, 2F, Sigma Center, Microsoft Advanced Technology Center, 49 Zhichun Road, Haidian District, Beijing 100190, P.R. China. E-mail: yongrui@microsoft.com.*
- *J. Tang is with the School of Computing, National University of Singapore, 21 Lower Kent Ridge Road, Singapore 119076. E-mail: tangjh@comp.nus.edu.sg.*
- *H.-J. Zhang is with the Microsoft Advanced Technology Center, 3F, Sigma Center, 49 Zhichun Road, Haidian District, Beijing 100190, P.R. China. E-mail: hjzhang@microsoft.com.*

Manuscript received 16 Mar. 2008; revised 19 June 2008; accepted 19 Aug. 2008; published online 21 Aug. 2008.
Recommended for acceptance by K. Murphy.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2008-03-0154.
Digital Object Identifier no. 10.1109/TPAMI.2008.218.

are large and need to annotate multiple labels for each image. In most cases, it is unnecessary to completely label all the new images and all their associated labels due to the fact that there exist redundancies between the different images. Therefore, we can design a strategy to utilize these redundancies to improve adaptation efficiency of the online models with only a small number of elaborately selected samples.

- Once a set of images is collected together with their labeling ground truth, a direct solution to obtaining a new classifier is to retrain the classification model with all the historical training set plus the newly acquired images. However, the intensive computational cost has restricted many sophisticated models to be retrained in practice. So an efficient algorithm is desired to incrementally adapt the image models with the new images.

To handle the above issues, we propose an online two-dimensional active learning algorithm for multilabel image classification together with an efficient online adaptation model.

Active learning is one of the most widely used approaches in image classification, as it can significantly reduce the effort in labeling training samples [5], [6], [7], [8]. Specifically, active learning approaches iteratively annotate a set of elaborately selected samples so that the expected classification error is minimized in each iteration. As a result, the total number of training samples that need to be labeled is smaller than nonactive learning approaches. The core of any active learning approach is the sample selection strategy. In the past decade, a number of active learning approaches were developed using different sample selection strategies [9], [10], [6]. For example, the authors of [11], [2] have explored reduction in uncertainty as the sample selection criterion and competitive performances have been achieved. Most of these approaches focus on the binary classification. However, in many real-world applications [12], [13], [3], a sample is usually associated with multiple labels rather than a single one. Under such a multilabel setting, each sample will be annotated as either "positive" or "negative" for each and every label. As a result, active learning with multilabel samples is much more challenging than that with binary-label ones, especially when the number of labels is large.

A direct way to tackle active learning under multilabel setting is to translate it into a set of binary problems, i.e., each category/label is independently handled by a binary active learning algorithm. For example, in [12], [14], two research groups have proposed such a binary-based active learning algorithm for multilabel classification problem. However, these approaches do not take into account the inherent relationship among multiple labels. In this paper, we propose a novel active learning strategy that iteratively selects sample-label pairs to minimize the expected classification error. Specifically, in each iteration, human annotators are only required to annotate/confirm a selected part of labels of selected samples while the remaining unlabeled part can be inferred according to the label correlations. We call this strategy Two-Dimensional Active Learning (2DAL) because it considers not only the samples to be labeled along the sample dimension but also the labels associated with these samples along the label dimension. An intuitive explanation of this strategy is that there exist both sample and label redundancies for multilabel samples. Therefore, annotating a set of selected sample-label pairs provides enough information for training the classifiers since the information in these pairs can be propagated to the rest along both the sample "dimension" and the label "dimension." Such a strategy can significantly reduce the required human labor. For example, "field" and "mountain" tend to occur simultaneously in an image. Therefore, it is reasonable to select only one label (e.g., "mountain") for annotation since the uncertainty of the other label can be remarkably decreased after annotating this one. Another example is "mountain" and "urban." In contrast to "field" and "mountain," these two labels often do not occur simultaneously. Thus, annotating one of them will probably eliminate the existence of the other one.

For the online applications, the second important issue is about an efficient online model adaptation algorithm. As more and more new sample-label pairs are added into training set during the 2DAL procedure, the multilabel model must be updated accordingly. The most straightforward approach is to retrain the model on the whole training set. However, such an offline approach will become impractical when more and more samples come into the training set over time. On the other hand, as the semantic meanings of the image concepts are constantly changing due to the evolution of user focuses (e.g., users are constantly changing their attentions due to the evolution of the fashion and news) and photography techniques (e.g., film photography 10 years before versus digital photographing today), a balance scheme should be incorporated into model adaptation algorithm so that it can trade-off between the old knowledge preserved in the existing model and the new information contained in newly acquired images. Thus, retraining the model with all old and new samples equally weighted in a batch-mode manner is not a proper scheme, especially when the number of the new samples is much smaller than the number of the old ones. It probably underestimates the effect of new images.

In contrast to the naive retraining approach, we propose a novel online adaptation algorithm for multilabel image classification. It can directly update the existing model with the new samples rather than with the whole training images. Thus, it is much more efficient for model adaptation during the 2DAL procedure than the traditional retraining approach. Furthermore, instead of equally using the old and new samples, this adaptation algorithm balances between preserving the old knowledge and complying with the new information. It can better leverage the new samples to capture evolution of concept semantics over time in many online applications. In particular, we formulate such an online adaptation algorithm by optimizing a variational problem, which minimizes the distance between the new and old models under a set of multilabel constraints. Compared to the widely used fully Bayesian approach [1] that requires to construct a set of intractable conjugate distributions, the model can be efficiently updated.

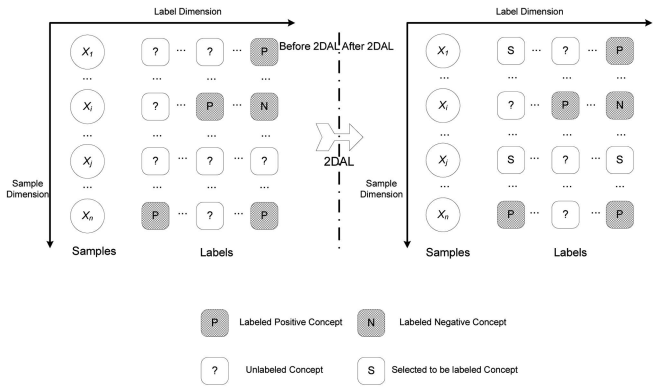

Fig. 1. The proposed 2DAL strategy.

In brief, in this paper, we summarize our contributions as follows:

- We propose a novel two-dimensional active learning strategy for multilabel image classification. It actively selects sample-label pairs to minimize the expected Bayesian classification error bound. This strategy utilizes the rich label correlations so that the entire annotation labor can be dramatically reduced.
- In each active learning iteration, an efficient online adaptation algorithm is developed to update the multilabel model without the need of retraining with all historical training samples. It can also balance between the knowledge preserved in the old model and the information contained in newly acquired samples, as well as capture the semantic evolution of the image concept.

2 Two-Dimensional Active Learning Strategy

In this section, we detail the underlying idea of the proposed 2DAL strategy in multilabel setting.

2.1 Description of 2DAL Framework

Fig. 1 illustrates the proposed 2DAL strategy. Different from the typical binary active learning formulation that selects the most informative samples for annotation, we jointly select both the samples and labels simultaneously. The underlying assumption is annotating a portion of well-selected labels that provides sufficient information for learning the classifier. As shown in Fig. 1, this strategy trades-off between the annotation labor and the learning performance along two dimensions, i.e., the sample and label dimensions. In contrast, traditional active learning algorithms can be seen as a one-dimensional active selection approach along only sample dimension. More specifically, along label dimension, all of the labels correlatively interact. Therefore, once labels are partially annotated, the remaining unlabeled concepts can be inferred based on label correlations. Theoretically, the label correlations have a connection with the expected Bayesian Error Bound (see the following lemma and theorem in Section 2.2), and thus, these label correlations can help reduce the prediction errors in the testing set during the active learning procedure. This approach saves much labor compared to fully annotating multiple labels especially when the number of labels is huge.

It is worth noting that as illustrated in Fig. 1, during 2DAL process, samples may have incomplete labels since the set is only partially labeled. This is different from traditional active learning algorithm. In Section 4.2, we will address how to learn the classification model from incomplete labels.

2.2 Multilabel Bayesian Error Bound

2DAL learner requests annotations on the basis of sample-label pairs, which, once incorporated into the training set, are expected to result in the lowest classification error. Here, we will first derive a *multilabel Bayesian Error Bound* when a sample-label pair is selected under multilabel setting. 2DAL will iteratively select the ones to minimize this bound.

We begin by defining some notations. For each sample $\boldsymbol{x}$, it has m labels y_i $(1 \leq i \leq m)$, each of which indicates whether its corresponding concept occurs. As stated before, in each 2DAL iteration, some of these labels have already been annotated while others not. Let $U(\boldsymbol{x}) \triangleq \{i|(\boldsymbol{x}, y_i)$ is unlabeled sample-label pair$\}$ denote the set of indexes of unlabeled part and $L(\boldsymbol{x}) \triangleq \{i|(\boldsymbol{x}, y_i)$ is labeled sample-label pair$\}$ denote the labeled part for $\boldsymbol{x}$. Note that $L(\boldsymbol{x})$ can be an empty set $\emptyset$, which indicates that no label has been annotated for $\boldsymbol{x}$. Let $P(\boldsymbol{y}|\boldsymbol{x})$ be the conditional distribution over samples, where $\boldsymbol{y} = \{0,1\}^m$ is the complete label vector and $P(\boldsymbol{x})$ is the sample distribution.

First, we establish a Bayesian error bound for classifying one unlabeled y_i once y_s is selected for annotating. This error bound originates from the equivocation bound given in [15], and we extend it to multilabel setting so that it can handle sample-label pairs.

Lemma 1. *Consider a sample* $\mathbf{x}$ *and its index set of labeled part* $U(\mathbf{x})$ *and unlabeled part* $L(\mathbf{x})$. *Once an unlabeled* y_s *is selected to request annotation (but not yet know its label), the Bayesian classification error* $\mathcal{E}(y_i|y_s, y_{L(\mathbf{x})}, \mathbf{x})$ *for an unlabeled* $y_i, i \in U(\mathbf{x})$ *is bounded as*

$$\begin{aligned} \frac{1}{2}H\big(y_i|y_s; y_{L(\mathbf{x})}, \mathbf{x}\big) - \epsilon \leq \mathcal{E}\big(y_i|y_s; y_{L(\mathbf{x})}, \mathbf{x}\big) \\ \leq \frac{1}{2}H\big(y_i|y_s; y_{L(\mathbf{x})}, \mathbf{x}\big), \end{aligned} \tag{1}$$

where

$$\begin{aligned} H\big(y_i|y_s; y_{L(\mathbf{x})}, \mathbf{x}\big) = \sum_{t,r\in\{0,1\}} \{-P\big(y_i = t, y_s = r|y_{L(\mathbf{x})}, \mathbf{x}\big) \\ \times \log P\big(y_i = t|y_s = r; y_{L(\mathbf{x})}, \mathbf{x}\big)\} \end{aligned}$$

is the conditional entropy of y_i *given the selected part* y_s *(both* y_i *and* y_s *are random variables because they have not been labeled) and the known labeled part* $y_{L(\mathbf{x})}$; $\epsilon = \frac{1}{2}\log\frac{5}{4}$ *is a constant.*

This lemma will be proven in Appendix A.

Remark 1. It is worth noting that this bound is irrelevant to the true label of the selected y_s. In fact, before the annotator gives the label of y_s, the true value of y_s is unknown. However, no matter what y_s holds, one or zero, this bound always holds.

Based on this lemma, we can obtain the following theorem that bounds the multilabel error.

Theorem 1 (Multilabel Bayesian classification error bound). *Under the condition of Lemma 1, the Bayesian classification error bound $\mathcal{E}(\boldsymbol{y}|y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x})$ for sample $\boldsymbol{x}$ is*

$$\begin{aligned}\mathcal{E}\big(\boldsymbol{y}|y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x}\big) &\triangleq \frac{1}{m}\sum_{i=1}^{m}\mathcal{E}\big(y_i|y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x}\big)\\ &\le \frac{1}{2m}\sum_{i=1}^{m}\big\{H\big(y_i|y_{L(\boldsymbol{x})}, \boldsymbol{x}\big) - MI\big(y_i; y_s|y_{L(\boldsymbol{x})}, \boldsymbol{x}\big)\big\},\end{aligned} \tag{2}$$

where $MI(y_i; y_s|y_{L(\boldsymbol{x})}, \boldsymbol{x})$ is the mutual information between the random variables y_i and y_s given the known labeled part $y_{L(\boldsymbol{x})}$.

Proof.

$$\begin{aligned}&\mathcal{E}\big(\boldsymbol{y}|y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x}\big)\\ &\overset{(1)}{=} \frac{1}{m}\sum_{i=1}^{m}\mathcal{E}\big(y_i|y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x}\big)\\ &\overset{(2)}{\le} \frac{1}{2m}\sum_{i=1}^{m}H\big(y_i|y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x}\big)\\ &\overset{(3)}{=} \frac{1}{2m}\sum_{i=1}^{m}\big\{H\big(y_i|y_{L(\boldsymbol{x})}, \boldsymbol{x}\big) - MI\big(y_i; y_s|y_{L(\boldsymbol{x})}, \boldsymbol{x}\big)\big\},\end{aligned} \tag{3}$$

where (1) is the definition of multilabel classification error, (2) directly follows Lemma 1, and (3) makes use of the relationship between mutual information and entropy: $MI(X;Y) = H(X) - H(X|Y)$. □

With the above theorem, we will derive the 2DAL selection strategy in the following section.

2.3 Pool-Based Two-Dimensional Multilabel Active Learning

We are concerned with *pool-based active learning*, i.e., a large pool $\mathcal{P}$ is available to the learner sampled from $P(\boldsymbol{x})$ and then the proposed 2DAL elaborately selects the sample-label pairs from this pool to reduce the expected classification error. We first write the expected Bayesian classification error over all samples in $\mathcal{P}$ before selecting a sample-label pair $(\boldsymbol{x}_s, y_s)$

$$\mathcal{E}^b(\mathcal{P}) = \frac{1}{|\mathcal{P}|}\sum_{\boldsymbol{x}\in\mathcal{P}}\mathcal{E}\big(\boldsymbol{y}|y_{L(\boldsymbol{x})}, \boldsymbol{x}\big). \tag{4}$$

We can use the above classification error on the pool to estimate the expected error over the full distribution $P(\boldsymbol{x})$, i.e., $E_{P(\boldsymbol{x})}\mathcal{E}(\boldsymbol{y}|y_{L(\boldsymbol{x})}, \boldsymbol{x}) = \int P(\boldsymbol{x})\mathcal{E}(\boldsymbol{y}|y_{L(\boldsymbol{x})}, \boldsymbol{x})d\boldsymbol{x}$, because the pool not only provides a finite set of samples but also an estimation of $P(\boldsymbol{x})$. After selecting the pair $(\boldsymbol{x}_s, y_s)$, the expected Bayesian classification error over the pool $\mathcal{P}$ is

$$\begin{aligned}&\mathcal{E}^a(\mathcal{P})\\ &= \frac{1}{|\mathcal{P}|}\left\{\mathcal{E}\big(\boldsymbol{y}|y_s; y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big) + \sum_{\boldsymbol{x}\in\mathcal{P}\setminus\boldsymbol{x}_s}\mathcal{E}\big(\boldsymbol{y}|y_{L(\boldsymbol{x})}, \boldsymbol{x}\big)\right\}\\ &= \frac{1}{|\mathcal{P}|}\Big\{\mathcal{E}\big(\boldsymbol{y}|y_s; y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big) - \mathcal{E}\big(\boldsymbol{y}|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big)\\ &\quad + \sum_{x\in\mathcal{P}}\mathcal{E}\big(\boldsymbol{y}|y_{L(x)}, \boldsymbol{x}\big)\Big\}.\end{aligned} \tag{5}$$

Therefore, the reduction of the expected Bayesian classification after selecting $(\boldsymbol{x}_s, y_s)$ over the whole pool $\mathcal{P}$ is

$$\Delta\mathcal{E}(\mathcal{P}) = \mathcal{E}^b(\mathcal{P}) - \mathcal{E}^a(\mathcal{P}). \tag{6}$$

Our goal is to select $(\boldsymbol{x}_s^\star, y_s^\star)$ to maximize the above expected error reduction. That is,

$$\begin{aligned}\big(\boldsymbol{x}_s^\star, y_s^\star\big) &= \underset{\boldsymbol{x}_s\in\mathcal{P}, y_s\in U(\boldsymbol{x}_s)}{\arg\max}\ \Delta\mathcal{E}(\mathcal{P})\\ &= \underset{\boldsymbol{x}_s\in\mathcal{P}, y_s\in U(\boldsymbol{x}_s)}{\arg\min}\ -\Delta\mathcal{E}(\mathcal{P}).\end{aligned} \tag{7}$$

Applying Lemma 1 and Theorem 1, we have

$$\begin{aligned}-\Delta\mathcal{E}(\mathcal{P}) &= \mathcal{E}^a(\mathcal{P}) - \mathcal{E}^b(\mathcal{P}),\\ &\overset{(1)}{=} \frac{1}{|\mathcal{P}|}\{\mathcal{E}\big(\boldsymbol{y}|y_s; y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big) - \mathcal{E}\big(\boldsymbol{y}|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big)\\ &\quad + \sum_{x\in\mathcal{P}}\mathcal{E}\big(\boldsymbol{y}|y_{L(\boldsymbol{x})}, \boldsymbol{x}\big)\} - \frac{1}{|\mathcal{P}|}\sum_{\boldsymbol{x}\in\mathcal{P}}\mathcal{E}\big(\boldsymbol{y}|y_{L(\boldsymbol{x})}, \boldsymbol{x}\big)\\ &= \frac{1}{|\mathcal{P}|}\{\mathcal{E}\big(\boldsymbol{y}|y_s; y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big) - \mathcal{E}\big(\boldsymbol{y}|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big)\}\\ &\overset{(2)}{\le} \frac{1}{|\mathcal{P}|}\left\{\frac{1}{2m}\sum_{i=1}^{m}H\big(y_i|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big)\right.\\ &\quad - \frac{1}{2m}\sum_{i=1}^{m}MI\big(y_i; y_s|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big)\\ &\quad \left. - \frac{1}{m}\sum_{i=1}^{m}\mathcal{E}\big(y_i|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big)\right\}\\ &\overset{(3)}{\le} \frac{1}{|\mathcal{P}|}\left\{\frac{1}{2m}\sum_{i=1}^{m}H\big(y_i|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big)\right.\\ &\quad - \frac{1}{2m}\sum_{i=1}^{m}MI\big(y_i; y_s|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big)\\ &\quad \left. - \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}H\big(y_i|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big) - \epsilon\right)\right\}\\ &= \frac{1}{|\mathcal{P}|}\left\{\epsilon - \frac{1}{2m}\sum_{i=1}^{m}MI\big(y_i; y_s|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big)\right\}.\end{aligned} \tag{8}$$

The equality (1) comes from (4), (5). The first inequality (2) follows Theorem 1 and the second inequality (3) comes from the lower bound of Lemma 1.

Consequently, by minimizing the obtained Bayesian error bound (8), we can select the sample-label pair for annotation as

$$\begin{aligned}&\big(\boldsymbol{x}_s^*, y_s^*\big)\\ &= \underset{\boldsymbol{x}_s\in\mathcal{P}, y_s\in U(\boldsymbol{x}_s)}{\arg\min}\ \frac{1}{|\mathcal{P}|}\left\{\epsilon - \frac{1}{2m}\sum_{i=1}^{m}MI\big(y_i; y_s|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big)\right\}\\ &= \underset{\boldsymbol{x}_s\in\mathcal{P}, y_s\in U(\boldsymbol{x}_s)}{\arg\max}\ \sum_{i=1}^{m}MI\big(y_i; y_s|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big).\end{aligned} \tag{9}$$

2.4 Further Discussions

1. As discussed in Section 2.1, the proposed 2DAL approach is an active learning algorithm along two dimensions, which reduces not only *sample uncertainty* but also *label uncertainty*. The above selection

strategy (9) well reflects these two targets. The last term in (9) can be rewritten as

$$\begin{aligned}&\sum_{i=1}^{m} MI\big(y_i; y_s|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big)\\ &= MI\big(y_s; y_s|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big) + \sum_{i=1, i\neq s}^{m} MI\big(y_i; y_s|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big)\\ &= H\big(y_s|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big) + \sum_{i=1, i\neq s}^{m} MI\big(y_i; y_s|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big).\end{aligned} \tag{10}$$

As we can see, the objective selection function for 2DAL has been divided into two parts: $H(y_s|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s)$ and $\sum_{i=1,i\neq s}^{m} MI(y_i; y_s|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s)$. The former entropy measures the uncertainty of the selected pair $(\boldsymbol{x}_s^\star, y_s^\star)$ itself, which is consistent with the typical one-dimensional active learning algorithm, i.e., to select the most "informative" (uncertain) samples near the classification boundary [16], [2], [17]. On the other hand, the latter mutual information terms measure the statistical redundancy among the selected label and the rest ones. By maximizing these mutual information terms, 2DAL provides maximum information to reduce the uncertainty of the other labels. This 2DAL strategy complies with our motivation of selecting sample-label pairs reducing the uncertainties along both *sample* and *label* dimensions. Note that when there is only one label associated with each sample, the selection criterion of (10) reduces to $H(y_s|\boldsymbol{x}_s)$, which is the same as the traditional binary-based criterion, i.e., to select the most uncertain sample for annotation [17], [9]. Thus, the traditional binary-based active learning can be seen as a special case of the 2DAL strategy with a single label.

2. It is worth indicating that the posterior $P(\boldsymbol{y}|\boldsymbol{x})$ needs to capture the label correlations in the proposed 2DAL strategy. If we assume the independence among the different labels, i.e., $P(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{m} P(y_i|\boldsymbol{x})$, the corresponding mutual information terms will become $MI\big(y_i; y_s|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big) = 0, i \neq s$. In this case, the selection criterion reduces to $\big(\boldsymbol{x}_s^\star, y_s^\star\big) = \arg\max_{\boldsymbol{x}_s\in\mathcal{P}, y_s\in U(\boldsymbol{x}_s)} H\big(y_s|y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s\big)$, i.e., to select the most uncertain sample-label pair. Such a criterion neglects the label correlations and would become less efficient to reduce label uncertainty. Therefore, a statistical method that models the label correlations is required in this case. We will develop an efficient Bayesian model in the following section.
3. When computing the mutual information terms in (9), we need the distribution $P(\boldsymbol{y}|\boldsymbol{x})$. However, the true distribution is unknown, but we can estimate it using the current learner. As stated in [18], such an approximation is reasonable because the most useful labeling is usually consistent with the learner's prior belief over the majority (but not all) of the unlabeled pairs.

3 Multilabel Online Learner

Once new sample-label pairs are selected according to the 2DAL strategy, the statistical model for multilabel images should be updated accordingly. However, as stated in Section 1, the conventional offline algorithms retrain a new model on the whole historically collected training set plus the new samples. It will become intractable when hundreds of thousands of samples are accumulated into the training set over time. Therefore, an efficient online adaptation algorithm is desired to adapt the old model to the new sample without retraining it. Intuitively, such an online classification algorithm should satisfy the following requirements:

- It ought to preserve the old knowledge that has already existed in the old model. This knowledge stores the rich historical information about the previously acquired training samples.
- It can reveal the information contained in the newly arrived multilabel samples. In contrast to the traditional binary-based algorithm (e.g., one-against-rest SVM), the label correlations must be modeled in this online learner.

In this section, we will present such an online learning algorithm that satisfies the above two requirements. We begin our discussion with the definition of some notations and the online setting. Under the online setting, we are given an existing old multilabel model $P^\tau(\boldsymbol{y}|\boldsymbol{x})$, which is trained from the historically acquired images. Then a set of new images and their corresponding labels $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$ is obtained in each 2DAL iteration. Each $\boldsymbol{x}_i \in \mathbb{R}^d$ is the feature vector and $\boldsymbol{y}_i \in \{0,1\}^m$ is an m-dimensional label vector in which m is the number of image labels and each element in $\boldsymbol{y}_i$ indicates the membership for the corresponding label. Our goal is to learn a new model $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ based on the existing model $P^\tau(\boldsymbol{y}|\boldsymbol{x})$ and $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$. In contrast to the retraining-based learning algorithm, the online learner does not utilize the historical training set but only the current new coming samples. It assumes that the information about the historical training samples has been preserved in the old model $P^\tau(\boldsymbol{y}|\boldsymbol{x})$, which is learned from these old samples. Thus, with much fewer of only new coming samples, the online learner can train a new model efficiently.

As mentioned before, this new model $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ ought to satisfy the two requirements: preserving the existing knowledge in $P^\tau(\boldsymbol{y}|\boldsymbol{x})$ while revealing the information in new samples $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$. These two requirements can be satisfied by formulating the following probabilistic variational problem. In this paper, we use the Kullback-Leibler Divergence (KLD) [19] to measure the degree of the new model preserving the existing knowledge contained in the old one, under a set of multilabel constraints revealing the information contained in the new samples:

$$\widehat{P}^{\tau+1}(\boldsymbol{y}|\boldsymbol{x}) = \arg\min_{P^{\tau+1}} \big\langle D_{KL}\big(P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x}) \| P^\tau(\boldsymbol{y}|\boldsymbol{x})\big)\big\rangle_{\tilde{P}(\boldsymbol{x})}, \tag{11}$$

$$s.t.\ \langle y_i\rangle_{P^{\tau+1}(\boldsymbol{x},\boldsymbol{y})} = \langle y_i\rangle_{\tilde{P}(\boldsymbol{x},\boldsymbol{y})} + \eta_i, 1 \le i \le m, \tag{12}$$

$$\big\langle y_i y_j\big\rangle_{P^{\tau+1}(\boldsymbol{x},\boldsymbol{y})} = \big\langle y_i y_j\big\rangle_{\tilde{P}(\boldsymbol{x},\boldsymbol{y})} + \theta_{ij}, 1 \le i < j \le m, \tag{13}$$

$$\langle y_i x_l\rangle_{P^{\tau+1}(\boldsymbol{x},\boldsymbol{y})} = \langle y_i x_l\rangle_{\tilde{P}(\boldsymbol{x},\boldsymbol{y})} + \phi_{il}, 1 \le i \le m, 1 \le l \le d, \tag{14}$$

$$\sum_{\boldsymbol{y}} P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x}) = 1, \tag{15}$$

where $\langle D_{KL}(P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})\|P^{\tau}(\boldsymbol{y}|\boldsymbol{x}))\rangle_{\tilde{P}(\boldsymbol{x})}$ is the KLD between the new model $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ and the old one $P^{\tau}(\boldsymbol{y}|\boldsymbol{x})$ over the sample frequency $\widetilde{P}(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m}\delta(\boldsymbol{x}-\boldsymbol{x}_i)$ taken from $\{\boldsymbol{x}_i\}_{i=1}^{n}$, where $\delta(\cdot)$ is the indicator function. Thus, we have

$$\begin{aligned}&\left\langle D_{KL}\left(P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})\|P^{\tau}(\boldsymbol{y}|\boldsymbol{x})\right)\right\rangle_{\tilde{P}(\boldsymbol{x})}\\&=\sum_{\boldsymbol{x}}\widetilde{P}(\boldsymbol{x})D_{KL}\left(P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})\|P^{\tau}(\boldsymbol{y}|\boldsymbol{x})\right)\\&=\sum_{i=1}^{n}D_{KL}\left(P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x}_i)\|P^{\tau}(\boldsymbol{y}|\boldsymbol{x}_i)\right),\end{aligned}\tag{16}$$

where $\langle\cdot\rangle_{P^{\tau+1}(\boldsymbol{x},\boldsymbol{y})}$ and $\langle\cdot\rangle_{\tilde{P}(\boldsymbol{x},\boldsymbol{y})}$ in (12)-(14) denote the expectation w.r.t. model distribution $P^{\tau+1}(\boldsymbol{x},\boldsymbol{y})$ and empirical distribution $\widetilde{P}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{m}\sum_{i=1}^{m}\delta(\boldsymbol{x}-\boldsymbol{x}_i)\cdot\delta(\boldsymbol{y}-\boldsymbol{y}_i)$ on the training samples $\{\boldsymbol{x}_i,\boldsymbol{y}_i\}_{i=1}^{n}$, respectively. The variables y_i, y_j, and x_l in $\langle\cdot\rangle$ represent the ith and jth elements in label vectors $\boldsymbol{y}$ and lth element in feature vectors $\boldsymbol{x}$, respectively. It is worth noting that the joint model distribution $P^{\tau+1}(\boldsymbol{x},\boldsymbol{y}) = P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})\widetilde{P}(\boldsymbol{x})$, i.e., we only care about the conditional distribution $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ and thus use the sample frequency $\widetilde{P}(\boldsymbol{x})$ on the training samples to approximate the true sample distribution $P^{\tau+1}(\boldsymbol{x})$. Constraints (12)-(14) restrict the new model to comply with the statistics on the new samples. It is similar to the conventional offline model used in the previous work [20], [4]. $\eta_i \sim N(0,\sigma_\eta^2)$, $\theta_{ij} \sim N(0,\sigma_\theta^2)$, and $\phi_{il} \sim N(0,\sigma_\phi^2)$ are the estimation errors following the Gaussian distribution which serve to smooth $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ to improve the model's generalization ability. These estimation error distributions can be due to the noise in the training samples. Note that we do not assume any specific probabilistic form of $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ and $P^{\tau}(\boldsymbol{y}|\boldsymbol{x})$. Therefore, the objective function is minimized by exploring all possible input functions instead of some parameterized functions like in the conventional optimization problem. This type of optimization objective is called variational optimization (see more detail about variational optimization in [21]). With no assumption of any specific probabilistic form, the variational optimization problem can search in a much larger functional space to find a more optimal one.

The above objective function (16) has a rather intuitive explanation. From the perspective of the information theory [19], the KLD measures the distance between two different distributions. Thus, by minimizing the KLD between $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ and $P^{\tau}(\boldsymbol{y}|\boldsymbol{x})$, the new model $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ can preserve the old knowledge in the existing model $P^{\tau}(\boldsymbol{y}|\boldsymbol{x})$ as much as possible. This is consistent with the first requirement above. On the other hand, in the multilabel constraints (12)-(14), we force the new model $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ to comply with three statistics on the new samples $\{\boldsymbol{x}_i,\boldsymbol{y}_i\}_{i=1}^{n}$. It satisfies the second requirement that the new model must comply with the information contained in the new samples. It is worth noting that by modeling the label correlations in (13), the obtained model reveals the underlying correlations between different labels. Finally, the constraint (15) just serves to normalize $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$. Fig. 2 illustrates the geometry explanation of this online learner. Here, $\mathcal{D}$ denotes a space in which each point is a potential conditional distribution $P(\boldsymbol{y}|\boldsymbol{x})$ for the new model.

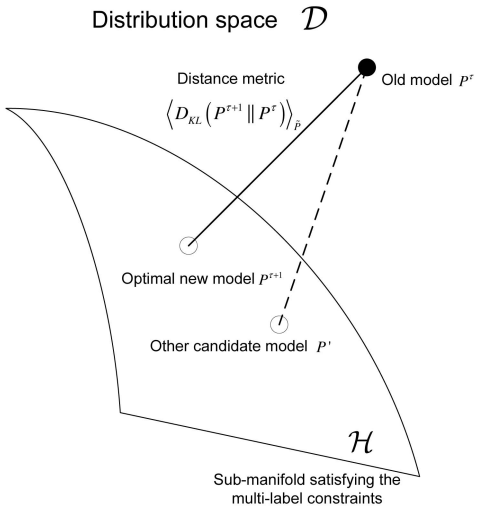

Fig. 2. A Geometry explanation of the proposed online learner. $\mathcal{D}$ is a space of all potential distributions for the new model. It is equipped with KLD as its distance metric. All the distributions satisfying the multilabel constraints constitute a submanifold $\mathcal{H}$. The optimal new model $P^{\tau+1}(\mathbf{y}|\mathbf{x})$ can be seen as the projection of the old model $P^{\tau}(\mathbf{y}|\mathbf{x})$ to the submanifold $\mathcal{H}$. In this figure, $P^{\tau+1}$ is the optimal new model and we can see its distance to the old model, which is less than that of any other candidate models P' in the submanifold $\mathcal{H}$.

This space is equipped with KLD as its distance metric. All the distributions satisfying the multilabel constraints constitute a submanifold $\mathcal{H}$ embedded in $\mathcal{D}$. Therefore, the above optimization problem is to find an optimal new model $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ in $\mathcal{H}$ with the minimum distance from the old model $P^{\tau}(\boldsymbol{y}|\boldsymbol{x})$. Equivalently, from the Geometrical prospective, the optimal solution $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ to this problem is a projection of the old model $P^{\tau}(\boldsymbol{y}|\boldsymbol{x})$ to the submanifold $\mathcal{H}$.

As stated in Section 1, when learning the new model, we should balance between the existing knowledge and the new information. The Gaussian error estimations in (12)-(14) serve to provide such a trading-off scheme. When the variances of Gaussian errors η_i, θ_{ij}, and ϕ_{il} are larger, the new model $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ will be biased to be the existing model $P^{\tau}(\boldsymbol{y}|\boldsymbol{x})$ since the multilabel constraints become more relaxed with relatively large noises η_i, θ_{ij}, and ϕ_{il} in the current training set. In contrast, the small variances will make $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ bias on the new information in $\{\boldsymbol{x}_i,\boldsymbol{y}_i\}_{i=1}^{n}$. Extremely, the removal of these error estimations will lead to a new model that completely complies with the new information. Furthermore, as to be derived later, these Gaussian errors will introduce a regularizer term in the dual form of this formulation. As suggested in [22], [4], they assume that the joint probability of estimation errors should be reasonably large, e.g., $P(\eta_i,\theta_{ij},\phi_{il}|1\le i<j\le m, 1\le l\le d)\ge\varepsilon$. Substitute $\eta_i \sim N(0,\sigma_\eta^2) = \frac{1}{\sqrt{2\pi}\sigma_\eta}\exp\{-\frac{\eta_i^2}{2\sigma_\eta^2}\}$, $\theta_{ij} \sim N(0,\sigma_\theta^2) = \frac{1}{\sqrt{2\pi}\sigma_\theta}\exp\{-\frac{\theta_{ij}^2}{2\sigma_\theta^2}\}$, and $\phi_{il} \sim N(0,\sigma_\phi^2) = \frac{1}{\sqrt{2\pi}\sigma_\phi}\exp\{-\frac{\phi_{il}^2}{2\sigma_\phi^2}\}$ into this inequality, and assume that these estimation errors are independent of each other, we have

$$\sum_i \frac{\eta_i^2}{2\sigma_\eta^2/n} + \sum_{i<j} \frac{\theta_{ij}^2}{2\sigma_\theta^2/n} + \sum_{i,l} \frac{\phi_{il}^2}{2\sigma_\phi^2/n} \le C. \tag{17}$$

Before moving further, we briefly discuss how the above online formulation tackles "concept drift" over time mentioned in Section 1. There already exist literatures working on this "concept drift" problem [16], [23]. From statistical perspective, "concept drift" can be modeled as the change of empirical distribution of the samples and the labels (i.e., the joint distribution of the samples and the labels) over time [16]. As indicated by (12)-(14), the proposed online model adapts the new empirical distribution $\widetilde{P}(\boldsymbol{x}, \boldsymbol{y})$ by complying with the first and second order statistics of $\widetilde{P}(\boldsymbol{x}, \boldsymbol{y})$. Through such a model adaptation, "concept drift" can be automatically captured by this online adaptation model.

By combining the formulations (15)-(17), according to Karush-Kuhn-Tucker (KKT) conditions, $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ can be solved by maximizing the dual Lagrange function

$$\begin{aligned}
&\mathcal{L}\big(P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x}), \eta, \theta, \phi, \boldsymbol{b}, \boldsymbol{R}, \boldsymbol{W}, \gamma, \varsigma\big) \\
&\quad = \big\langle D_{KL}\big(P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x}) \| P^{\tau}(\boldsymbol{y}|\boldsymbol{x})\big)\big\rangle_{\widetilde{P}(\boldsymbol{x})} \\
&\qquad + \sum_i b_i \big(\langle y_i \rangle_{\widetilde{P}(\boldsymbol{x},\boldsymbol{y})} + \eta_i - \langle y_i \rangle_{P^{\tau+1}(\boldsymbol{x},\boldsymbol{y})}\big) \\
&\qquad + \sum_{i<j} R_{ij} \big(\langle y_i y_j \rangle_{\widetilde{P}(\boldsymbol{x},\boldsymbol{y})} + \theta_{ij} - \langle y_i y_j \rangle_{P^{\tau+1}(\boldsymbol{x},\boldsymbol{y})}\big) \\
&\qquad + \sum_{i,l} W_{il} \big(\langle y_i x_l \rangle_{\widetilde{P}(\boldsymbol{x},\boldsymbol{y})} + \phi_{il} - \langle y_i x_l \rangle_{P^{\tau+1}(\boldsymbol{x},\boldsymbol{y})}\big) \\
&\qquad + \gamma \left(\sum_i \frac{\eta_i^2}{2\sigma_\eta^2/n} + \sum_{i<j} \frac{\theta_{ij}^2}{2\sigma_\theta^2/n} + \sum_{i,l} \frac{\phi_{ij}^2}{2\sigma_\phi^2/n} - C \right) \\
&\qquad + \sum_x \varsigma(x) \left(1 - \sum_y P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x}) \right),
\end{aligned} \tag{18}$$

where $\boldsymbol{b}, \boldsymbol{R}, \boldsymbol{W}, \gamma, \varsigma$ are Lagrangian multipliers in which $\boldsymbol{b} = [b_1, b_2, \ldots, b_m]^T$ is an $m \times 1$ column vector, $\boldsymbol{R} = [R_{ij}]_{m \times m}$ is a strict upper matrix with $R_{ij} = 0$ for $i \ge j$, and $\boldsymbol{W} = [W_{ij}]_{m \times d}$ is an $m \times d$ matrix. The above function can be maximized by taking its derivatives and setting them to zero. Specifically, the derivative of KLD w.r.t. $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ is

$$\begin{aligned}
&\frac{\partial \big\langle D_{KL}(P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x}) \| P^{\tau}(\boldsymbol{y}|\boldsymbol{x}))\big\rangle_{\widetilde{P}(\boldsymbol{x})}}{\partial P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})} \\
&\quad = \frac{\partial}{\partial P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})} \sum_\nu \widetilde{P}(\nu) \sum_h P^{\tau+1}(h|\nu) \log \frac{P^{\tau+1}(h|\nu)}{P^{\tau}(h|\nu)} \\
&\quad = \frac{\partial}{\partial P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})} \sum_\nu \widetilde{P}(\nu) \sum_h P^{\tau+1}(h|\nu) \log P^{\tau+1}(h|\nu) \\
&\qquad - \frac{\partial}{\partial P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})} \sum_\nu \widetilde{P}(\nu) \sum_h P^{\tau+1}(h|\nu) \log P^{\tau}(h|\nu) \\
&\quad = \widetilde{P}(\boldsymbol{x}) \big\{ \log P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x}) + 1 - \log P^{\tau}(\boldsymbol{y}|\boldsymbol{x}) \big\}.
\end{aligned} \tag{19}$$

Thus, the derivative of $\mathcal{L}(P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x}), \eta, \theta, \phi, \boldsymbol{b}, \boldsymbol{R}, \boldsymbol{W}, \gamma, \varsigma)$ w.r.t. $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ is

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})} = \widetilde{P}(\boldsymbol{x}) \{ &\log P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x}) + 1 - \log P^{\tau}(\boldsymbol{y}|\boldsymbol{x}) \\
&- \boldsymbol{y}^T (\boldsymbol{b} + \boldsymbol{R}\boldsymbol{y} + \boldsymbol{W}\boldsymbol{x}) \} - \varsigma(\boldsymbol{x}).
\end{aligned} \tag{20}$$

It is easy to compute the derivatives of Lagrange w.r.t. other parameters η_i, θ_{ij}, and ϕ_{il}:

$$\frac{\partial \mathcal{L}}{\partial \eta_i} = b_i + n\gamma \frac{\eta_i}{\sigma_\eta^2}; \quad \frac{\partial \mathcal{L}}{\partial \theta_{ij}} = R_{ij} + n\gamma \frac{\theta_{ij}}{\sigma_\theta^2}; \quad \frac{\partial \mathcal{L}}{\partial \phi_{il}} = W_{il} + n\gamma \frac{\phi_{il}}{\sigma_\phi^2}. \tag{21}$$

Setting the above derivatives (20), (21) of Lagrange to be zero, we can find when γ is zero, $\boldsymbol{b}$, $\boldsymbol{R}$, and $\boldsymbol{W}$ are also reduced to be the trivial solution zero. Thus, we can assume $\gamma > 0$, and we obtain

$$P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x}) \propto P^{\tau}(\boldsymbol{y}|\boldsymbol{x}) \exp\big\{ \boldsymbol{y}^T (\boldsymbol{b} + \boldsymbol{R}\boldsymbol{y} + \boldsymbol{W}\boldsymbol{x}) \big\}. \tag{22}$$

Considering the normalization condition (15), we can get

$$P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z^{\tau+1}(\boldsymbol{x})} P^{\tau}(\boldsymbol{y}|\boldsymbol{x}) \exp\big\{ \boldsymbol{y}^T (\boldsymbol{b} + \boldsymbol{R}\boldsymbol{y} + \boldsymbol{W}\boldsymbol{x}) \big\}, \tag{23}$$

$$\eta_i = -\frac{\sigma_\eta^2}{n\gamma} b_i, \quad \theta_{ij} = -\frac{\sigma_\theta^2}{n\gamma} R_{ij}, \quad \phi_{il} = -\frac{\sigma_\phi^2}{n\gamma} W_{il}, \tag{24}$$

where

$$Z^{\tau+1}(\boldsymbol{x}) = \sum_{\boldsymbol{y}} P^{\tau}(\boldsymbol{y}|\boldsymbol{x}) \exp\big\{ \boldsymbol{y}^T (\boldsymbol{b} + \boldsymbol{R}\boldsymbol{y} + \boldsymbol{W}\boldsymbol{x}) \big\} \tag{25}$$

is the partition function. Now, let us recompute the KLD between $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ and $P^{\tau}(\boldsymbol{y}|\boldsymbol{x})$ in (16) considering (23)

$$\begin{aligned}
&\big\langle D_{KL}\big(P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x}) \| P^{\tau}(\boldsymbol{y}|\boldsymbol{x})\big)\big\rangle_{\widetilde{P}(\boldsymbol{x})} \\
&\quad = \sum_{\boldsymbol{x},\boldsymbol{y}} \widetilde{P}(\boldsymbol{x}) P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x}) \log \frac{P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})}{P^{\tau}(\boldsymbol{y}|\boldsymbol{x})} \\
&\quad = \sum_{\boldsymbol{x},\boldsymbol{y}} \widetilde{P}(\boldsymbol{x}) P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x}) \log \frac{\exp\big\{ \boldsymbol{y}^T (\boldsymbol{b} + \boldsymbol{R}\boldsymbol{y} + \boldsymbol{W}\boldsymbol{x}) \big\}}{Z^{\tau+1}(\boldsymbol{x})} \\
&\quad = \sum_{\boldsymbol{x},\boldsymbol{y}} \widetilde{P}(\boldsymbol{x}) P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x}) \big\{ \boldsymbol{y}^T (\boldsymbol{b} + \boldsymbol{R}\boldsymbol{y} + \boldsymbol{W}\boldsymbol{x}) \big\} \\
&\qquad - \sum_{\boldsymbol{x}} \widetilde{P}(\boldsymbol{x}) \log Z^{\tau+1}(\boldsymbol{x}) \\
&\quad = \sum_i b_i \langle y_i \rangle_{P^{\tau+1}(\boldsymbol{x},\boldsymbol{y})} + \sum_{i<j} R_{ij} \langle y_i y_j \rangle_{P^{\tau+1}(\boldsymbol{x},\boldsymbol{y})} \\
&\qquad + \sum_{i,l} W_{il} \langle y_i x_l \rangle_{P^{\tau+1}(\boldsymbol{x},\boldsymbol{y})} - \big\langle \log Z^{\tau+1}(\boldsymbol{x}) \big\rangle_{\tilde{P}(\boldsymbol{x}).}
\end{aligned} \tag{26}$$

By substituting the above equation and (24) into (18), we can obtain the Lagrangian Function and the corresponding dual optimization problem to solve the parameters $\boldsymbol{b}, \boldsymbol{R}, \boldsymbol{W}$

$$\begin{aligned}
&\boldsymbol{b}^\star, \boldsymbol{R}^\star, \boldsymbol{W}^\star = \arg\max_{\boldsymbol{b},\boldsymbol{R},\boldsymbol{W}} \mathcal{L}(\boldsymbol{b}, \boldsymbol{R}, \boldsymbol{W}), \\
&\arg\max_{\boldsymbol{b},\boldsymbol{R},\boldsymbol{W}} \big\langle \boldsymbol{y}^T (\boldsymbol{b} + \boldsymbol{R}\boldsymbol{y} + \boldsymbol{W}\boldsymbol{x}) - \log Z^{\tau+1}(\boldsymbol{x}) \big\rangle_{\widetilde{P}(\boldsymbol{x},\boldsymbol{y})} \\
&\qquad - \frac{\alpha_{\boldsymbol{b}}}{2n} \|\boldsymbol{b}\|_2^2 - \frac{\alpha_{\boldsymbol{R}}}{2n} \|\boldsymbol{R}\|_F^2 - \frac{\alpha_{\boldsymbol{W}}}{2n} \|\boldsymbol{W}\|_F^2 \\
&\quad = \arg\max_{\boldsymbol{b},\boldsymbol{R},\boldsymbol{W}} \sum_{i=1}^n \big\{ \boldsymbol{y}_i^T (\boldsymbol{b} + \boldsymbol{R}\boldsymbol{y}_i + \boldsymbol{W}\boldsymbol{x}_i) - \log Z^{\tau+1}(\boldsymbol{x}_i) \big\} \\
&\qquad - \frac{\alpha_{\boldsymbol{b}}}{2n} \|\boldsymbol{b}\|_2^2 - \frac{\alpha_{\boldsymbol{R}}}{2n} \|\boldsymbol{R}\|_F^2 - \frac{\alpha_{\boldsymbol{W}}}{2n} \|\boldsymbol{W}\|_F^2,
\end{aligned} \tag{27}$$

with

$$\alpha_{\boldsymbol{b}} = \sigma_\eta^2/\gamma, \quad \alpha_{\boldsymbol{R}} = \sigma_\theta^2/\gamma, \quad \alpha_{\boldsymbol{W}} = \sigma_\phi^2/\gamma, \tag{28}$$

where $\|\cdot\|_2$ and $\|\cdot\|_F$ are norm-2 and Frobenius norm, respectively. Here, $-\frac{\alpha_{\boldsymbol{b}}}{2n}\|\boldsymbol{b}\|_2^2-\frac{\alpha_{\boldsymbol{R}}}{2n}\|\boldsymbol{R}\|_F^2-\frac{\alpha_{\boldsymbol{W}}}{2n}\|\boldsymbol{W}\|_F^2$ serves as the regularization term. Note that in this dual optimization problem, the old model $P^{\tau}(\boldsymbol{y}|\boldsymbol{x})$ affects the objective function through the partition function $Z^{\tau+1}(\boldsymbol{x})$ of (25). Moreover, according to the last equality in (27), the summation is only taken over the newly acquired samples, instead of all the historically accumulated training sets like the offline learner. Thus, with a rather smaller number of new samples, the above optimization problem can be solved much more efficiently for the proposed online learner.

Take the derivatives of $\mathcal{L}(\boldsymbol{b},\boldsymbol{R},\boldsymbol{W})$ w.r.t. $\boldsymbol{b},\boldsymbol{R},\boldsymbol{W}$

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial b_i}&=\langle y_i\rangle_{\widetilde{P}(\boldsymbol{x},\boldsymbol{y})}-\langle y_i\rangle_{P^{\tau+1}(\boldsymbol{x},\boldsymbol{y})}-\frac{\alpha_{\boldsymbol{b}}}{n}b_i,\\ \frac{\partial\mathcal{L}}{\partial R_{ij}}&=\langle y_iy_j\rangle_{\widetilde{P}(\boldsymbol{x},\boldsymbol{y})}-\langle y_iy\rangle_{P^{\tau+1}(\boldsymbol{x},\boldsymbol{y})}-\frac{\alpha_{\boldsymbol{R}}}{n}R_{ij},\\ \frac{\partial\mathcal{L}}{\partial W_{il}}&=\langle y_ix_l\rangle_{\widetilde{P}(\boldsymbol{x},\boldsymbol{y})}-\langle y_ix_l\rangle_{P^{\tau+1}(\boldsymbol{x},\boldsymbol{y})}-\frac{\alpha_{\boldsymbol{W}}}{n}W_{il}.\end{aligned}\tag{29}$$

Given the above derivatives, we can use the efficient gradient descent methods (such as L-BFGS [24]) to maximize (27).

Here, we would like to have a brief discussion about the optimization problem (27). By setting the larger variances $\sigma_\eta^2,\sigma_\eta^2,\sigma_\phi^2$, it says that there exists considerable noise in the new samples $\{\boldsymbol{x}_i,\boldsymbol{y}_i\}_{i=1}^n$ (see (12)-(14)). At this time, according to the regularization term in (27), the smaller parameters $\boldsymbol{b},\boldsymbol{R},\boldsymbol{W}$ are preferred. Thus, the new model $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ will approach to the old one. Conversely, $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ will more consider the new samples. So by setting different $\sigma_\eta^2,\sigma_\eta^2,\sigma_\phi^2$, we can make balance between old knowledge in $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ and new information in $\{\boldsymbol{x}_i,\boldsymbol{y}_i\}_{i=1}^n$. Such a balance is useful when the semantic meaning of image concept is changing over time, as well as the number of the new training samples is rather smaller than that of the historical training samples (see Section 1 for detail).

Note that when deriving $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ in (23), we do not assume any specific probabilistic form of the old model $P^{\tau}(\boldsymbol{y}|\boldsymbol{x})$. Thus, any statistical model can be used as $P^{\tau}(\boldsymbol{y}|\boldsymbol{x})$, such as logistic regression model and Gaussian process model. However, without loss of generality, we can assume that $P^{\tau}(\boldsymbol{y}|\boldsymbol{x})$ has the following form:

$$P^{\tau}(\boldsymbol{y}|\boldsymbol{x})=\frac{1}{Z^{\tau}(\boldsymbol{x})}\exp\{\boldsymbol{y}^T(\boldsymbol{b}^{\tau}+\boldsymbol{R}^{\tau}\boldsymbol{y}+\boldsymbol{W}^{\tau}\boldsymbol{x})\},\tag{30}$$

where $Z^{\tau}(\boldsymbol{x})=\sum_{\boldsymbol{y}}\exp\{\boldsymbol{y}^T(\boldsymbol{b}^{\tau}+\boldsymbol{R}^{\tau}\boldsymbol{y}+\boldsymbol{W}^{\tau}\boldsymbol{x})\}$ is the partition function. Thus, according to (23), the new model $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ is

$$\begin{aligned}&P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})\\&=\frac{1}{Z^{\tau+1}(\boldsymbol{x})}\exp\{\boldsymbol{y}^T((\boldsymbol{b}^{\tau}+\boldsymbol{b}^{\star})+(\boldsymbol{R}^{\tau}+\boldsymbol{R}^{\star})\boldsymbol{y}+(\boldsymbol{W}^{\tau}+\boldsymbol{W}^{\star})\boldsymbol{x}))\}.\end{aligned}\tag{31}$$

We can find that $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ has the same probabilistic form like $P^{\tau}(\boldsymbol{y}|\boldsymbol{x})$ except that their parameters have been adapted as

$$\begin{aligned}\boldsymbol{b}^{\tau+1}&\leftarrow\boldsymbol{b}^{\tau}+\boldsymbol{b}^{\star},\\ \boldsymbol{R}^{\tau+1}&\leftarrow\boldsymbol{R}^{\tau}+\boldsymbol{R}^{\star},\\ \boldsymbol{W}^{\tau+1}&\leftarrow\boldsymbol{W}^{\tau}+\boldsymbol{W}^{\star}\}\end{aligned}\tag{32}$$

Therefore, such an online adaptation can then be iterated in the same manner in each iteration. Here, the initial model P^0 can be started with the parameters $\boldsymbol{b}^0=0$, $\boldsymbol{R}^0=0,\boldsymbol{W}^0=0$.

To the best of our knowledge, we are the first to develop an online learner for multilabel classification problem, although there exist some incremental or online learners for the binary classification, which do not model the correlations between different labels [25], [26], [27].

4 Implementation Details

In this section, we discuss some implementation details about the two-dimensional active learning with the proposed online learner.

4.1 Kernelization

Note that the model in (23) is linear and can be effective on a set of samples that varies linearly. However, it will fail to capture the structure of the feature space if the variations among the samples are nonlinear. But image classification is, in this case, when one is trying to extract features from image categories that vary in their appearance, illumination conditions, and complex background clutters. Therefore, a nonlinear version is required to classify the images based on their nonlinear structure in their feature space.

Here, we extend the model in (23) to a nonlinear one so that the powerful kernel method can be adopted. A transformation ψ maps samples into a target space in which kernel function $k(\boldsymbol{x}',\boldsymbol{x})$ gives the inner product. We can rewrite (23) as

$$\begin{aligned}&P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})\\&\quad=\frac{1}{Z^{\tau+1}(\boldsymbol{x})}P^{\tau}(\boldsymbol{y}|\boldsymbol{x})\exp\big(\boldsymbol{y}^T(\boldsymbol{b}+\boldsymbol{R}y)+\boldsymbol{y}^T\psi(\boldsymbol{W})\cdot\psi(\boldsymbol{x})\big),\end{aligned}\tag{33}$$

where $\psi(\boldsymbol{W})$ is the mapped weighting matrix. According to the Representer Theorem, the optimal weighting vector of the single-label problem is a linear combination of samples. In the proposed multilabel setting, the mapped weighting matrix $\psi(\boldsymbol{W})$ can still be written as a linear combination of $\psi(\boldsymbol{x}_i)$ except that the combination coefficients are vectors instead of scalars, i.e.,

$$\begin{aligned}\psi(\boldsymbol{W})&=\sum_{i=1}^{n}\theta(\boldsymbol{x}_i)\psi^T(\boldsymbol{x}_i)\\&=[\,\theta(\boldsymbol{x}_1)\quad\theta(\boldsymbol{x}_2)\quad\cdots\quad\theta(\boldsymbol{x}_n)\,]\begin{bmatrix}\psi^T(\boldsymbol{x}_1)\\ \psi^T(\boldsymbol{x}_2)\\ \vdots\\ \psi^T(\boldsymbol{x}_n)\end{bmatrix}\\&=\Theta\cdot\begin{bmatrix}\psi^T(\boldsymbol{x}_1)\\ \psi^T(\boldsymbol{x}_2)\\ \vdots\\ \psi^T(\boldsymbol{x}_n)\end{bmatrix},\end{aligned}\tag{34}$$

where the summation is taken over the samples in the training set. $\theta(\boldsymbol{x}_i)$ is a coefficient vector and Θ is an $m \times n$ matrix in which each row is the weighting coefficients for each label. Accordingly, we have

$$\begin{aligned}&\psi(\boldsymbol{W}) \cdot \psi(\boldsymbol{x})\\ &\quad = \Theta \cdot [k(\boldsymbol{x}_1, \boldsymbol{x}) \quad \cdots \quad k(\boldsymbol{x}_n, \boldsymbol{x})]^T\\ &\quad = \Theta \cdot k(\boldsymbol{x}),\end{aligned} \tag{35}$$

and

$$\begin{aligned}&P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})\\ &\quad = \frac{1}{Z^{\tau+1}(\boldsymbol{x})} P^{\tau}(\boldsymbol{y}|\boldsymbol{x}) \exp\big(\boldsymbol{y}^T(\boldsymbol{b} + \boldsymbol{R}\boldsymbol{y} + \Theta k(\boldsymbol{x}))\big),\end{aligned} \tag{36}$$

where $k(\boldsymbol{x}) = [\, k(\boldsymbol{x}_1, \boldsymbol{x}) \ \cdots \ k(\boldsymbol{x}_n, \boldsymbol{x})\,]^T$ is an $n \times 1$ vector and it can be seen as a new representation of sample $\boldsymbol{x}$. Correspondingly, with the identity $\|\phi(\boldsymbol{W})\|_F^2 = tr(\phi(\boldsymbol{W})\phi(\boldsymbol{W})^T) = tr(\Theta K \Theta^T)$, the Lagrangian function (27) can be rewritten as

$$\begin{aligned}\mathcal{L}(\boldsymbol{b}, \boldsymbol{R}, \Theta) = &\big\langle \boldsymbol{y}^T(\boldsymbol{b} + \boldsymbol{R}\boldsymbol{y} + \Theta \cdot k(\boldsymbol{x})) - Z^{\tau+1}(\boldsymbol{x}) \big\rangle_{\widetilde{P}(\boldsymbol{x},\boldsymbol{y})}\\ &- \frac{\alpha_{\boldsymbol{b}}}{2n}\|\boldsymbol{b}\|_2^2 - \frac{\alpha_{\boldsymbol{R}}}{2n}\|\boldsymbol{R}\|_F^2 - \frac{\alpha_{\boldsymbol{W}}}{2n} tr(\Theta K \Theta^T),\end{aligned} \tag{37}$$

where $K = [k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{n\times n}$ is the kernel matrix. By maximizing (37), we can estimate the optimal parameters $\boldsymbol{b}^\star, \boldsymbol{R}^\star, \Theta^\star$ in this kernelization formulation.

Here, we do not assume any specific form of the old model $P^{\tau}(\boldsymbol{y}|\boldsymbol{x})$. Similar to the discussion in the above section, we can assume that $P^{\tau}(\boldsymbol{y}|\boldsymbol{x})$ has the following form:

$$P^{\tau}(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z^{\tau}(\boldsymbol{x})} \exp\big(\boldsymbol{y}^T(\boldsymbol{b}^{\tau} + \boldsymbol{R}^{\tau}\boldsymbol{y} + \Theta^{\tau} k^{\tau}(\boldsymbol{x}))\big). \tag{38}$$

Accordingly, the new model $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ can be

$$\begin{aligned}&P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z^{\tau+1}(\boldsymbol{x})}\\ &\cdot \exp\left\{\boldsymbol{y}^T\left((\boldsymbol{b}^{\tau} + \boldsymbol{b}^\star) + (\boldsymbol{R}^{\tau} + \boldsymbol{R}^\star)\boldsymbol{y} + [\Theta^{\tau}\ \Theta^\star]\begin{bmatrix} k^{\tau}(\boldsymbol{x}) \\ k(\boldsymbol{x}) \end{bmatrix}\right)\right\}.\end{aligned} \tag{39}$$

The above new model has a similar probabilistic form as the old one except that the parameters have been adapted as

$$\begin{aligned}\boldsymbol{b}^{\tau+1} &\leftarrow \boldsymbol{b}^{\tau} + \boldsymbol{b}^\star,\\ \boldsymbol{R}^{\tau+1} &\leftarrow \boldsymbol{R}^{\tau} + \boldsymbol{R}^\star,\\ \Theta^{\tau+1} &\leftarrow [\Theta^{\tau}\ \Theta^\star]\},\end{aligned} \tag{40}$$

and the new representation for the sample $\boldsymbol{x}$ is

$$k^{\tau+1}(\boldsymbol{x}) \leftarrow \begin{bmatrix} k^{\tau}(\boldsymbol{x}) \\ k(\boldsymbol{x}) \end{bmatrix}. \tag{41}$$

With the above equations, we can recursively update the parameters of the probabilistic model once the new samples are acquired. As stated in Section 1, such an online learning algorithm can be much more efficient than the traditional retraining algorithm, especially when the number of the accumulated samples grows rapidly.

4.2 Incomplete Labeling

Given the partially labeled training set constructed by 2DAL (see Fig. 1), we can handle the incomplete labels by integrating out the unlabeled part yielding the marginal distribution of the labeled part $P^{\tau+1}(y_{L(\boldsymbol{x})}|\boldsymbol{x}) = \sum_{y_{U(\boldsymbol{x})}} P^{\tau+1}(y_{U(\boldsymbol{x})}, y_{L(\boldsymbol{x})}|\boldsymbol{x})$. But this form will lead to intractable computations and a nonconvex objective function which results in a local optimum solution. Instead, we use the Expectation Maximization (EM) algorithm [28] to solve this incomplete labeling problem. EM algorithm can greatly reduce the computational cost. As stated in many existing literatures [29], EM iteratively optimizes a series of local lower bounds to the original objective function obtained by marginalizing over all the unlabeled part. Such local lower bounds are convex, and thus, the global optimum can be found at each M-step for these convex bounds. These related works prove that these local lower bounds can approximate the true nonconvex objective function well enough, and thus, EM can result in a good solution.

E-Step: Given the current tth step parameter estimation $\boldsymbol{b}_t, \boldsymbol{R}_t, \Theta_t$, the $\mathcal{T}$-function (i.e., the expectation of the Lagrangian (37) under the current parameters given the labeled part) can be written as

$$\begin{aligned}&\mathcal{T}(\boldsymbol{b}, \boldsymbol{R}, \Theta|\boldsymbol{b}_t, \boldsymbol{R}_t, \Theta_t)\\ &\quad = \big\langle E_{U(\boldsymbol{x})|L(\boldsymbol{x});\boldsymbol{b}_t,\boldsymbol{R}_t,\Theta_t} \boldsymbol{y}^T(\boldsymbol{b} + \boldsymbol{R}\boldsymbol{y} + \Theta k(\boldsymbol{x})) - Z^{\tau+1}(\boldsymbol{x}) \big\rangle_{\widetilde{P}(\boldsymbol{x},\boldsymbol{y})}\\ &\qquad - \frac{\alpha_{\boldsymbol{b}}}{2n}\|\boldsymbol{b}\|_2^2 - \frac{\alpha_{\boldsymbol{R}}}{2n}\|\boldsymbol{R}\|_F^2 - \frac{\alpha_{\boldsymbol{W}}}{2n} tr(\Theta K \Theta^T),\end{aligned} \tag{42}$$

where $E_{U(\boldsymbol{x})|L(\boldsymbol{x});\boldsymbol{b}_t,\boldsymbol{R}_t,\Theta_t}$ is the expectation operator given the current estimated conditional probability $P^{\tau+1}(y_{U(\boldsymbol{x})}|y_{L(\boldsymbol{x})}, \boldsymbol{x}; \boldsymbol{b}_t, \boldsymbol{R}_t, \Theta_t)$.

M-Step: Update the parameters by minimizing $\mathcal{T}$-function:

$$\boldsymbol{b}_{t+1}, \boldsymbol{R}_{t+1}, \Theta_{t+1} = \arg\max_{\boldsymbol{b},\boldsymbol{R},\Theta} \mathcal{T}(\boldsymbol{b}, R, \Theta|\boldsymbol{b}_t, \boldsymbol{R}_t, \Theta_t). \tag{43}$$

The derivatives of $\mathcal{T}$-function with respect to its parameters $\boldsymbol{b}, \boldsymbol{R}, \Theta$ is

$$\begin{aligned}\frac{\partial \mathcal{T}}{\partial b_i} &= \big\langle E_{y_i|L(\boldsymbol{x});\boldsymbol{b},\boldsymbol{R},\Theta} y_i \big\rangle_{\widetilde{P}(\boldsymbol{x},\boldsymbol{y})} - \langle y_i \rangle_{P^{\tau+1}(\boldsymbol{x},\boldsymbol{y})} - \frac{\alpha_{\boldsymbol{b}}}{n} b_i,\\ \frac{\partial \mathcal{T}}{\partial R_{ij}} &= \big\langle E_{y_i,y_j|L(\boldsymbol{x});\boldsymbol{b},\boldsymbol{R},\Theta} y_i y_j \big\rangle_{\widetilde{P}(\boldsymbol{x},\boldsymbol{y})} - \big\langle y_i y_j \big\rangle_{P^{\tau+1}(\boldsymbol{x},\boldsymbol{y})} - \frac{\alpha_{\boldsymbol{R}}}{n} R_{ij},\\ \frac{\partial \mathcal{T}}{\partial \Theta_{il}} &= \big\langle E_{y_i|L(\boldsymbol{x});\boldsymbol{b},\boldsymbol{R},\Theta} y_i k(\boldsymbol{x}_l, \boldsymbol{x}) \big\rangle_{\widetilde{P}(\boldsymbol{x},\boldsymbol{y})} - \langle y_i k(\boldsymbol{x}_l, \boldsymbol{x}) \rangle_{P^{\tau+1}(\boldsymbol{x},\boldsymbol{y})}\\ &\qquad - \frac{\alpha_{\boldsymbol{W}}}{n} \sum_{k=1}^{n} \Theta_{ik} k(\boldsymbol{x}_k, \boldsymbol{x}_l).\end{aligned} \tag{44}$$

Similarly, with the above derivatives, L-BFGS [24] can then be applied to maximize (43).

4.3 Efficient Inference

When computing the derivatives (43), we need to compute the marginal distributions of $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ over the fully connected graph on the labels, such as $P^{\tau+1}(y_i|\boldsymbol{x})$ and $P^{\tau+1}(y_i, y_j|\boldsymbol{x})$. On the other hand, these marginal distributions are also required to compute the mutual information used in 2DAL (see (9)). It is known that the computational cost will be intractable with the increment

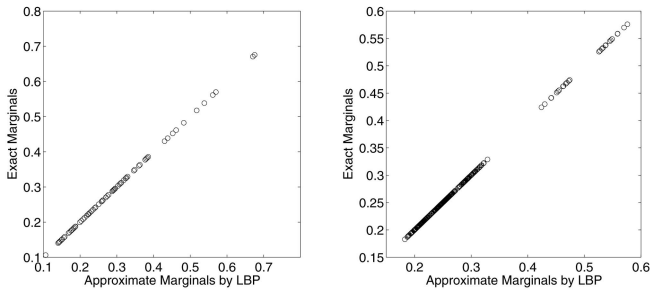

Fig. 3. Correlation plots between the exact and approximate marginals $P^{\tau+1}(y_i|\boldsymbol{x})$ and $P^{\tau+1}(y_i, y_j|\boldsymbol{x})$ for the proposed fully connected graphical models with the different number of labels: (a) 6 labels and (b) 14 labels.

of the label number m, and this limits the applicability of the algorithm. Fortunately, there exist many efficient algorithms to compute these marginals, such as Markov Chain Monte Carlo (MCMC) [30], Loopy Belief Propagation (LBP) [31], and Expectation Propagation (EP) [32]. Here, we adopt the widely used LBP to compute these marginal distributions. To apply the LBP, we need to provide the local evidences and potential functions. From the kernelized model $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ in (36), we can find its local evidences as

$$\Psi^{\tau+1}(y_i) \propto \exp\left\{ b_i^{\tau+1} y_i + \sum_{k=1}^{n} \Theta_{ik}^{\tau+1} y_i k^{\tau+1}(\boldsymbol{x}_k, \boldsymbol{x}) \right\}, \quad (45)$$

and its potentials

$$\Psi^{\tau+1}(y_i, y_j) \propto \exp\left\{ R_{ij}^{\tau+1} y_i y_j \right\}. \quad (46)$$

By propagating the above local evidences and potentials, LBP can then be used to efficiently compute all the marginal distributions simultaneously. For detailed LBP algorithm, refer to [31].

Murphy et al. [33] have conducted an empirical study and they find that LBP converges well on a variety of graphical models. Once LBP converges, the obtained marginals can give a good approximation to the exact marginals. Similarly, in order to verify this conclusion on the model represented by (42) and (43), we also conduct an empirical study to test the convergence of LBP on the two fully connected graphical models with 6 and 14 labels. A set of test samples from the data sets in following experiment (see Section 6) is used as observations $\boldsymbol{x}$. Then the posterior marginals $P^{\tau+1}(y_i|\boldsymbol{x})$ and $P^{\tau+1}(y_i, y_j|\boldsymbol{x})$ are computed by LBP. Experiments show that LBP converges in less than two iterations over all these test samples and Fig. 3 illustrates the correlation between the exact and approximate marginals. The results illustrate that LBP converges to good approximate marginals w.r.t. the exact one. The absolute errors between these approximate marginals and the exact ones are 8.50×10^{-6} and 1.31×10^{-5}, on average, in these two graphical models, respectively. If the typical LBP does not converge in some special cases, there also exist methods for preventing LBP from oscillation. For example, Murphy et al. [33] propose to use "momentum" by replacing the current messages with a weighted combination of the current message and previous ones, which can significantly reduce the chance of oscillation.

Fig. 4. The user interface that organizes image annotation by image. Users annotate all the concepts simultaneously for each image before proceeding to the next one.

5 User Interface for Annotating Images

An effective User Interface (UI) is a critical factor to improve the interaction efficiency between human beings and computers when annotating images. Traditionally, the image annotation interface is designed to annotate all the concepts simultaneously for each image before proceeding to the next one (see Fig. 4 for example of this interface). However, we argue that such an interface is not the most suitable choice to annotate images with multiple labels, because the annotators must switch their minds between different concepts to annotate these images and it can exhaust annotators' energy more quickly. An alternative choice is to annotate all the images exhaustively with a concept before proceeding to the next one. In this interface, annotators can focus on only one concept at one time so that they can quickly browse a group of images to judge if the concept exists in these images. This is because the vision system of human beings can rapidly respond to visual information of multiple images in a very short time. For example, given the images in an annotation interface illustrated in Fig. 5, annotators can find which images contain the concept "airplane" in a rather short time, instead of having to judge these images one by one. In contrast, in the traditional annotation interface of Fig. 4, annotators cannot focus on one concept to annotate all the images before proceeding to the next concept, and it reduces the annotators' efficiency. Therefore, we organize the image annotation by concept as illustrated in Fig. 5 rather than by image as in Fig. 4. In addition to the consideration of annotation efficiency, organizing annotation by concept can also lead to more accurate and complete annotation than annotating all concepts for each image simultaneously. Past experience has shown that the latter can cause many concepts to be missed (i.e., causing false negative labeling) [34].

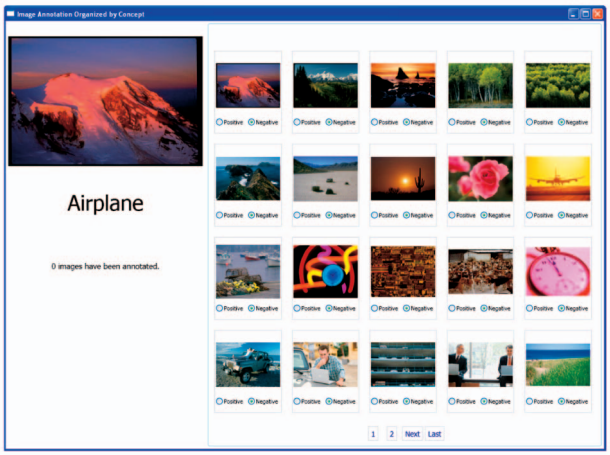

Fig. 5. The user interface that organizes image annotation by concept. Users annotate all the images exhaustively with a concept before proceeding to the next one.

In the user interface that organizes image annotation by concept as Fig. 5, the annotation workload is proportional to the number of sample-label pairs rather than the number of distinct images associated with these pairs, because the time used to annotate a certain concept for an image (i.e., a sample-label pair) can be assumed to be a constant on average. Therefore, in this interface, it is reasonable to compare the performances of 2DAL against the other active learning approaches under the same number of selected sample-label pairs as the basic unit to measure the annotation costs.

To verify the assumption that image annotation workload is proportional to the number of label-sample pairs rather than the number of distant images in the annotation interface organized by concept, we conduct a user study on this interface. Each of the three users annotated 200 sample-label pairs that are associated with *varying* number of images in each step. This step was repeated 10 times, yielding 2,000 sample-label pairs per user. First, we compare the annotation efficiency on the interface organizing annotation by image as Fig. 4 and by concept as Fig. 5. In Table 1, we report the time used on these two interfaces by these three users. Obviously, organizing annotation by concept is much more efficient than that by image, and thus, we will adopt the former one to annotate images. Second, we verify that the image annotation workload is related to the number of label-sample pairs on the interface as Fig. 5. Fig. 6 illustrates the linear relationship between the annotation time and the number of annotated sample-label pairs on this interface. We can find that the annotation time is proportional to the number of annotated sample-label pairs. It verifies that the annotation workload can be measured by the number of sample-label pairs. Therefore, in the later experiments, we will compare the performances of different active learning approaches under the same number of sample-label pairs.

6 Experiments on Two Benchmark Data Sets

In this section, we conduct experiments on two publicly available data sets to evaluate the proposed algorithm.

TABLE 1
Comparison between the Interface Organizing Annotation by Image and by Concept

User ID	Organizing annotation by image	Organizing annotation by concept
1	3660	1386
2	3138	1230
3	3138	1162

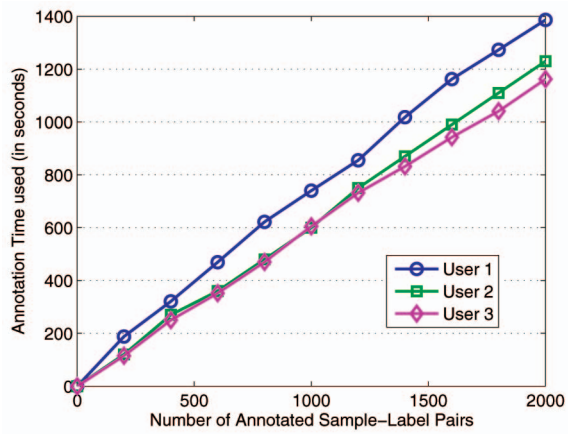

Fig. 6. User study for the user interface for image annotation organized by concept.

6.1 Natural Scene Classification

This natural scene data set is first used in a previous research on the multilabel image scene classification problem [13].[1] It contains 2,407 natural images belonging to one or more of six natural scene categories including beach, sunset, fall foliage, field, mountain, and urban. Since the data sets are multilabeled, there are 14,442 sample-label pairs in this set.

An image is first converted into CIE Luv color space and then the first and second color moments (mean and variance) are extracted over a 7×7 grid on the image. The end result is a $49 \times 2 \times 3 = 294$-dimension feature vector [13].

In this experiment, we compare the following three active learning strategies:

1. The proposed 2DAL strategy (2DAL+online): using the proposed sample-label pair selection criterion in Section 2.2 associated with the online adaptation statistical model (23) as the underlying classifier.
2. 1D active learning strategy (1DAL+SVM): using the mean-max loss active learning strategy that has been proposed in [12] on multilabel active learning. As stated in Section 1, this strategy selects only along the sample dimension. It does not take advantage of the label correlations to reduce human labeling cost. To the best of our knowledge, there exist very limited literatures on the multilabel active learning [12], [14] and 1DAL, here, is among these methods. All these existing methods are 1D-style active learning, which only selects samples rather than

1. This data set is publicly available at http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multilabel.html#scene-classification.

TABLE 2
Scene Data Set

Class	Total	Class	Total
Beach	369	Beach+Mountain	38
Sunset	364	Foliage+Mountain	13
Foliage	360	Field+Mountain	75
Field	327	Field+Foliage+Mountain	1
Beach+Field	1	Urban	405
Foliage+Field	23	Beach+Urban	19
Mountain	405		

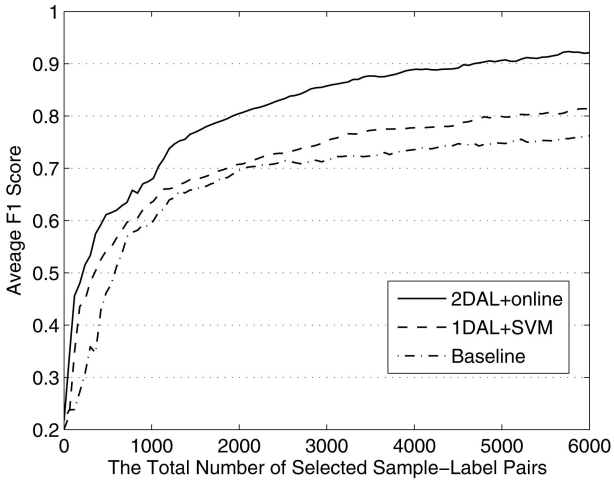

Fig. 7. The performance of three active learning strategies over the *Scene* data set.

sample-label pairs. Thus, we compare with the existing 1DAL method to verify the effectiveness of the proposed 2DAL.

3. Baseline: the random strategy—selecting the sample-label pairs at random. For the sake of fair comparison with the proposed 2DAL, we also use the online adaptation learner (23) in Section 3 as the classifier.

We use the average $F1$ score over all different labels for performance evaluation, i.e., $F1 = \frac{2rp}{r+p}$, where p and r are precision and recall, respectively. It is the harmonic mean of precision and recall. In statistics, the $F1$ score measures a test's accuracy and has been widely used in information retrieval. For this Scene data set (shown in Table 2), we use 241 (10 percent) images as the initial training set. In each iteration, 60 sample-label pairs are selected by the 2DAL. Note that for 1DAL, it requests annotation on the basis of samples rather than sample-label pairs, so in each iteration, it selects 10 images for annotating all the six labels or equivalently 60 image-label pairs. The average $F1$ score is then computed over all the remaining unlabeled data. In Fig. 7, we show the performance of these three strategies versus the number of selected sample-label pairs. The proposed 2DAL has the best performance in all the iterations. With the number of selected pairs increasing, the improvement becomes more and more significant. Table 3 compares the $F1$ scores after 100 iterations over all the six scene categories, which illustrates that 2DAL outperforms the other strategies on all the labels. In particular, the improvement is obvious on "Urban." Such an improvement is obtained by considering its significant correlations with other labels, such as "Mountain" and "Fall foliage." It confirms that 2DAL can obviously improve the classification performance.

TABLE 3
$F1$ Scores after 100 Iterations on Six Scene Categories

Class	2DAL+online	1DAL+SVM	Baseline
Beach	0.9523	0.8652	0.6744
Sunset	0.9916	0.9421	0.9002
Fall Foliage	0.9887	0.9338	0.8927
Field	0.9588	0.8813	0.8071
Mountain	0.7806	0.6457	0.6122
Urban	0.8534	0.6162	0.6856

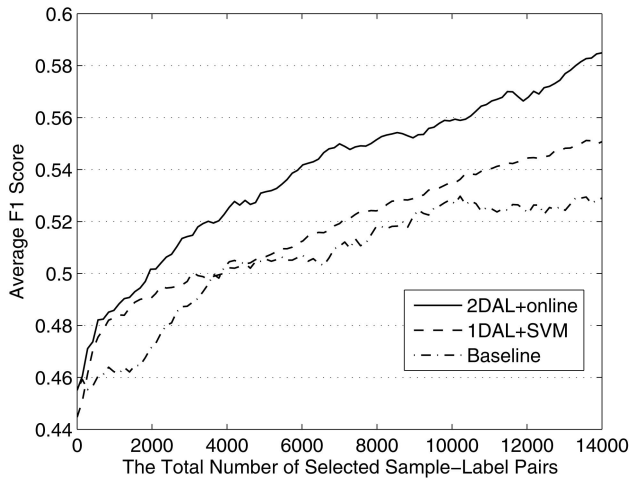

Fig. 8. The performance of three active learning strategies over the *Yeast* data set.

Note that in the above comparison, the performances of different active learning approaches are compared under the same number of sample-label pairs. As indicated in Section 5, the labor cost for image annotation is linear to the number of sample-label pairs in the user interface as Fig. 5. Therefore, it is reasonable to prove the superiority of an active learning algorithm if it can improve the classification accuracy with the same number of selected sample-label pairs.

6.2 Gene Classification

The second benchmark data set is the Yeast data set [11], which consists of microarray expression data and phylogenetic profiles with 2,417 genes. Each gene in the set belongs to one or more of 14 different functional classes,[2] yielding 33,838 sample-label pairs. Fig. 9 illustrates the distribution of the label numbers for the gene samples on this Yeast data set. The detailed description about this biological data set can be found in [35].

In the experiment, 242 (10 percent) genes with their labels are used as the initial training set. In each iteration, 140 sample-label pairs are selected. Similar to the above section, 1DAL selects 10 samples for annotating all their labels, which is 140 sample-label pairs. Fig. 8 compares the performance of the three strategies.

From the above two experiments, we have observed the following:

2. This data set is publicly available at http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multilabel.html#yeast.

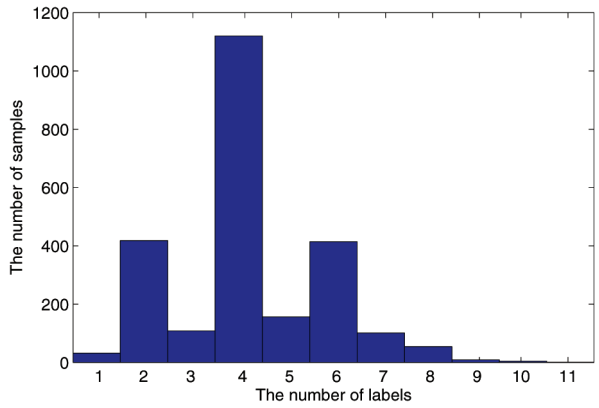

Fig. 9. The distribution of the label numbers for the gene samples on the *Yeast* data set.

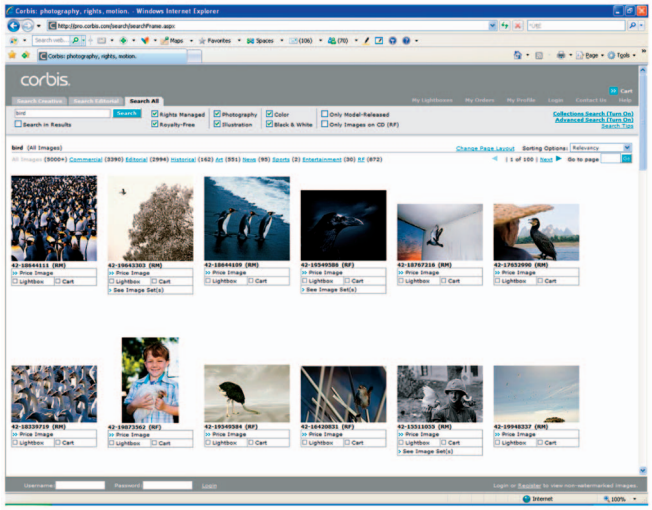

Fig. 10. Snapshot of the image sharing Web site—Corbis: http://www.corbis.com. This Web site provides a service for searching the images by their keywords. Such a service proposes the requirement to automatically annotate the labels of the images. In this example, this service returns the images that have been labeled by the keyword "bird."

1. Given a fixed number of annotations, 2DAL outperforms 1DAL over all the active learning iterations. This is because the former considers both sample and label uncertainty for selecting sample-label pair, while 1DAL only considers the sample uncertainty. Therefore, the informative label correlations associated with each sample can help reduce the expensive human labor needed to construct the labeled pool.
2. The proposed 2DAL gives good performance on diverse data sets, ranging from natural scenes to gene images. This is an important character of a good algorithm to be used in real-world applications.

7 Real-World Image Classification

In this section, we evaluate the proposed online active learning algorithm on a real-world image data set. This data set is obtained from an image sharing Web site—Corbis (http://www.corbis.com). Fig. 10 illustrates the snapshot of this Web site. We construct a realistic image data set from this Web site. This data set contains 28,868 images uploaded by the users. Each image is annotated from a set of 23 labels. Table 4 shows these 23 labels and the image numbers associated with them. Fig. 11 illustrates some example images and their associated labels. We can find that these real-world images are usually annotated by multiple labels.

TABLE 4
The Number of Images Associated with Each Label in the Corbis Data Set

Label Name	Total	Label Name	Total
airplane	350	face	998
bicycle	5146	flower	512
bird	13899	grass	5178
boat	1617	road	956
body	4532	sheep	3455
book	1200	sign	1262
building	8736	sky	1091
car	2808	tree	235
cat	1041	water	175
chair	3910	mountain	1225
cow	2454	horse	126
dog	1462		

Different from the global features extracted from the images in the *Scene* data set [13], we extract the dense local features as suggested in [36]. In more detail, each image is first normalized into 256×256 pixels and then represented by a two-dimensional array of local patches. Each local patch has 16×16 pixels over a grid with spacing of eight pixels. We extract three kinds of features for each patch as follows:

- Color moments (nine-dimensional)—the first, second, and third color moments in each component of CIE Luv color space.
- Co-occurrence texture (16-dimensional)—it computes the occurrence distribution of the 16 different patterns in a local patch.
- SIFT descriptor—the 128-dimensional SIFT descriptor is processed by principal component analysis (PCA) to reduce its dimensionality to 50.

With the above grid-based two-dimensional representation of images, we can compute a kernel between the images into the proposed algorithm, as depicted in Section 4. In this paper, we use the joint appearance-spatial kernel proposed in our previous work [37]. This kernel is based on the distance between two Dependent Tree-HMMs (DT-HMMs) [38]—a variant Two-Dimensional Hidden Markov Model (2DHMM). Fig. 12 illustrates an example of DT-HMM. With this structure of 2D Markov fields on the images, we can tractably compute a tighter upper bound of the Kullback-Leibler Divergence (KLD) between these DT-HMMs. Accordingly, a kernel can be defined on the basis of the KLD by exponentiating them. The reason that we choose this kernel in this paper is twofold. First, different from many other kernels which only measure the appearance similarity between images, this kernel can also measure the spatial similarity simultaneously. In fact, the spatial structure of local patches is rather important cue for an image classification problem. Thus, a joint appearance-spatial kernel like this one can bring significant advantage during the classification. Second, we extract three different kinds of feature modalities as above. Effective fusion of these modalities in classification attracts much research attentions [39]. This kernel algorithm can compute the similarity measure between images over all the modalities

Fig. 11. Some examples of the images acquired from the Corbis Web site and their associated labels. From these examples, we can find that most of the real-world images can be annotated by multiple labels simultaneously.

simultaneously without a late fusion [39] of the classification results on each single modality. The details about computing these kernels can be found in [37].

7.1 Performance Comparison with Previous Work

In this experiment, we also compare the proposed 2DAL+online approach with 1DAL+SVM and Baseline like the experiments on the benchmark data sets.

At the beginning, we use 10,000 images as the initial training set. In each iteration, 1DAL selects 100 images for annotating all 23 labels, containing $100 \times 23 = 2{,}300$ sample-label pairs. Equivalently, 2DAL and the random baseline request annotation of 2,300 sample-label pairs. A separate set of 5,000 samples are used as the validation set and the average $F1$ score is computed on it to compare the performances of different approaches. Such active learning iterations are repeated 100 times. Fig. 13 illustrates the performance of the three active learning strategies. We can find that

1. the 2DAL+online has the best performance of all the three strategies in all the 100 iterations.
2. after 100 iterations, $F1$ score of the 2DAL+online obtains 0.772 in contrast to 0.629 for 1DAL+SVM and 0.557 for the baseline. In other words, 2DAL+online gains 22.7 and 38.6 percent improvements compared to 1DAL+SVM and the baseline in terms of $F1$ score.

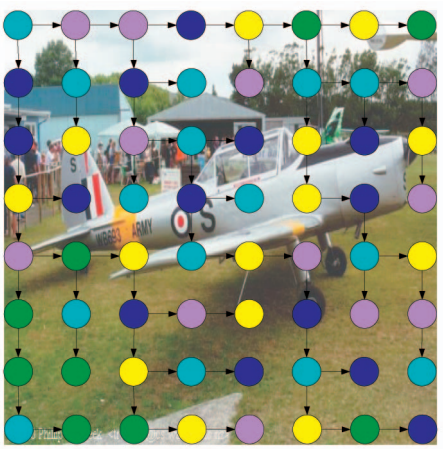

Fig. 12. An example of Dependent Tree Hidden Markov Model (DT-HMM) [38]. A joint appearance-spatial kernel can be defined by computing the Kullback-Leibler Divergence between these DT-HMMs [37]. This kernel can capture the image similarity of both the appearance and spatial structures in a unifying formulation.

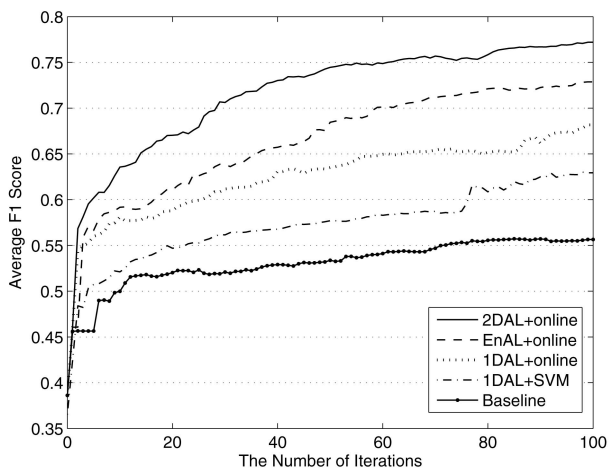

Fig. 13. Comparison of various active learning approaches. 1) 2DAL+online: using 2DAL strategy with the online learner; 2) EnAL+online: using the most uncertainty criterion with the online learner; 3) 1DAL+online: using 1DAL strategy with online learner; 4) 1DAL+SVM: using 1DAL strategy with SVM learner; and 5) Baseline: using the random strategy.

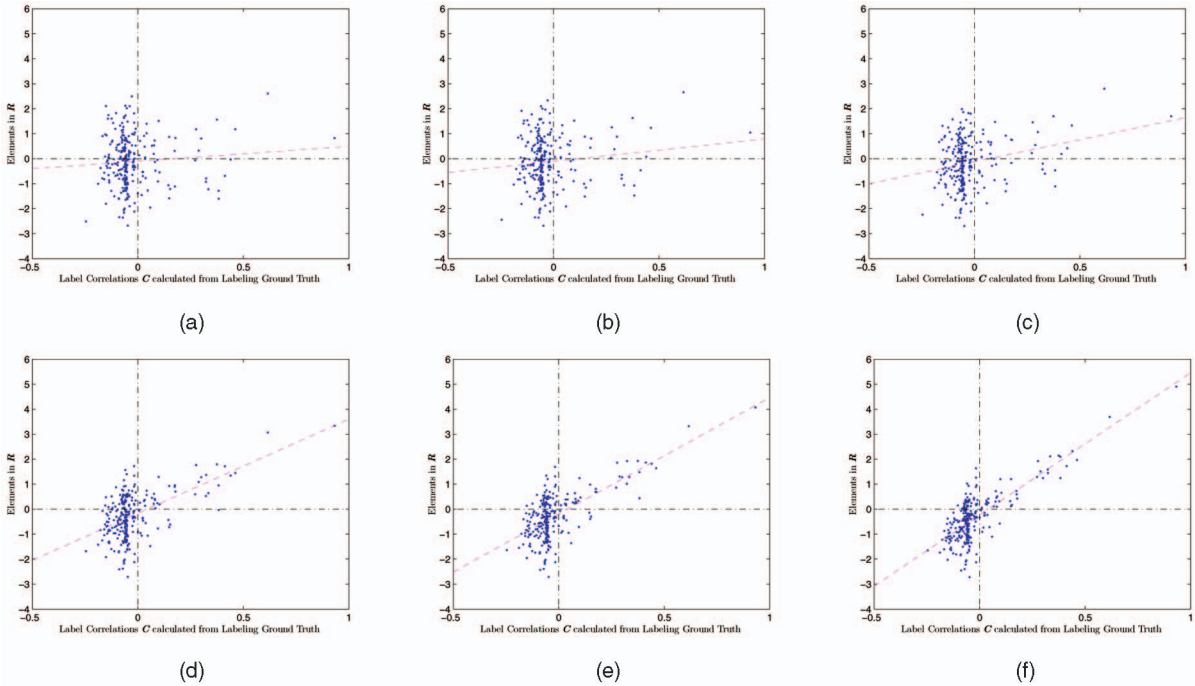

Fig. 14. The horizontal axis in each figure denotes the linear correlation coefficients $\mathbf{C}$ between the labels calculated from the labeling ground truth. Here, the linear correlation coefficients can be seen as the label correlation ground truth. The vertical axis denotes the elements in $\mathbf{R}$. The dashed line is the linear regression from the points in figures. These figures illustrate the relation between the label correlation ground truth (horizontal axis) and the corresponding elements in $\mathbf{R}$ (vertical axis). With the increment of the iteration number, the compliance between the label correlation ground truth $\mathbf{C}$ and the elements in $\mathbf{R}$ becomes more and more significant. That is to say, a large C_{ij} usually has a large R_{ij} and vice versa. Such a compliance indicates that the learned $\mathbf{R}$ can reflect the label correlations contained in the labeling ground truth of the Corbis data set. (a) Iteration number $= 20$; (b) Iteration number $= 40$; (c) Iteration number $= 50$; (d) Iteration number $= 60$; (e) Iteration number $= 80$; and (f) Iteration number $= 100$.

7.2 Does Label Correlation Really Help in 2DAL?

To show the superiority of the proposed 2DAL+online, we conduct an experiment which also selects the sample-label pairs like 2DAL+online but ignores the label correlations. As presented in (10), the 2DAL strategy can be divided into two terms: one is the self-entropy for the selected sample-label pair and the other is its correlations with the other labels. If we ignore the second term, this strategy reduces to an algorithm similar to the traditional active learning approach that only selects the most "informative" pair but ignores the label correlations. Here, we test the strategy without the second term to see if the label correlation really contributes to the performance improvement. We call this strategy EnAL (Entropy-based Active Learning) since it only uses the first entropy term. For the sake of fair comparison with 2DAL, we also use the online learner as the underlying classifier in EnAL. The experimental results are illustrated in Fig. 13 denoted by EnAL+online. From this illustration, we can observe that

- 2DAL+online outperforms EnAL+online in all the 100 iterations. This result proves that utilizing the label correlations in active learning is a better strategy.

7.3 Where the Superiority of 2DAL+Online Comes from, 2DAL Selection Strategy or Online Learner?

It is worth noting that 2DAL+online is based on the proposed online learner for the multilabel classification in Section 3 while 1DAL+SVM proposed in [12] depends on SVM learner. We conduct extra experiments here to testify whether the performance gain in 2DAL+online comes from the 2DAL selection strategy proposed in Section 2 or from the online learner developed in Section 3. Three compared experiments are conducted as follows:

1. *Experiment I:* It uses the online learner as the underlying statistical model and selects the sample-label pairs according to the proposed 2DAL strategy, which is 2DAL+online.
2. *Experiment II:* It also uses the online learner to train the label prediction model. However, it trains the classification model based on the image samples selected by 1DAL algorithm, that is, 1DAL+online.
3. *Experiment III:* It uses the typical SVM as the underlying learner and the 1DAL strategy to select image samples, that is, 1DAL+SVM.

Experiments I and III are the same as the above 2DAL+online and 1DAL+SVM algorithms. In contrast, Experiment II combines the sample selection strategy in Experiment III and the online learner in Experiment I, i.e., in each iteration, the underlying learner in Experiment II is trained from the same selected data set in Experiment III. Thus, we denote experiment II by 1DAL+online. Fig. 13 compares the results of these three experiments. From these results, we can observe that

1. Under the same underlying learner (i.e., online learner), 2DAL+online has the better performance than the 1DAL+online. It indicates that 2DAL selection strategy is superior to the 1DAL selection strategy.

TABLE 5
The Average Computing Time Used for Sample Selection in Each Iteration

Sample Selection	Computing Time
2DAL	149.5 seconds
1DAL	92.9 seconds

TABLE 6
The Average Computing Time for Updating/Retraining Models in Each Iteration

Learning Model	Computing Time
Online Learner	7.7 minutes
SVM	67.3 minutes

2. On the other hand, the underlying learners of 1DAL+online and 1DAL+SVM are trained on the same selected data set. It demonstrates that the online learner for multilabel classification performs better than the typical SVM learner.

From the above observations, we can conclude that both the 2DAL selection strategy and the online learner contribute to the performance improvement of the proposed 2DAL algorithm (2DAL+online).

The superiority of the online learner is owed to the rich multilabel correlations in data set while the typical SVM learner ignores them. To demonstrate it, we illustrate the label correlations and their corresponding parameters in the matrix $\boldsymbol{R}$ in Fig. 14. As discussed in Section 3, each element R_{ij} in $\boldsymbol{R}$ is related to the label correlations between labels i and j. In Fig. 14, we illustrate the label correlation ground truth C_{ij} in the horizontal axis versus the learned R_{ij} in the vertical axis $(1 \leq i < j \leq m)$. Here, C_{ij} denotes the linear correlation coefficient between labels i and j calculated from the labeling ground truth over the whole data set. From this figure, we can find that when the iteration number of active learning increases from 20, 40, 50, 60, 80 to 100, the compliance between the label correlation ground truth $\boldsymbol{C}$ and the learned model parameter $\boldsymbol{R}$ becomes more and more significant. That is to say, a larger C_{ij} usually has a larger R_{ij} and vice versa. It confirms that the learned model parameters $\boldsymbol{R}$ can reflect the real labeling correlations of the data set. In other words, the online learner captures the label correlation evolution contained in the selected sample-label pairs during the active learning procedure.

7.4 Computing Time

With the use of the proposed online learner, 2DAL is significantly more efficient than 1DAL which is based on the offline SVM learner [12] . We summarize and compare the average computing time of the proposed 2DAL and 1DAL in each active learning iteration in Tables 5 and 6. They are both performed on a 3.0 GHz PC with 1 GB RAM memory. The computing time consists of two parts. One part is for the sample selection in active learning and the other is for updating online model for 2DAL or retraining the SVM model for 1DAL. As for the computing time spent on sample selection, 2DAL strategy spends a little more time than 1DAL. They use average 149.5 and 92.9 seconds, respectively, in each iteration. On the other hand, the online learner is much more efficient than the SVM in each iteration of updating/retraining their respective models. From Table 6, the online learner is nearly one-order magnitude faster than the typical SVM, because it can be updated by only using the newly acquired training samples while SVM needs to be retrained on all the training samples.

8 Conclusion

In this paper, we develop a two-dimensional multilabel active learning algorithm asking for human annotation along both sample and label dimensions. In contrast to the typical one-dimensional active learning algorithm that asks to annotate all the labels of the selected image samples, this 2DAL strategy only requires to annotate a part of label set given the selected samples and the remaining part can be inferred according to the learned label correlations. This strategy can effectively reduce the unnecessary annotation labor to construct the training set for a classifier. Specifically, we derive a multilabel Bayesian error bound and the new sample-label pairs are selected to minimize it.

Furthermore, we also develop an online adaptation model which updates the existing model with only the newly acquired samples. Most of the existing learning algorithm has to be retrained with all the historically acquired training samples. This learning scheme becomes impractical when more and more training samples are accumulated into the training set over time during the active learning iterations. Instead, the proposed online learner can efficiently update the existing model without retraining it. In detail, the new model is obtained by minimizing its Kullback-Leibler distance from the existing one under a set of multilabel constraints.

Finally, we conduct experiments on two benchmark data sets and a real-world image data set obtained from an image sharing Web site—Corbis. The experiments prove the superiority of the proposed 2DAL strategy to the other existing multilabel active learning algorithm. Furthermore, it also demonstrates the efficiency of the online learner compared to the other learning algorithm with retraining mode.

Appendix A

Proof of Lemma 1

Here, we give the proof of Lemma 1.

Proof. Since the selected y_s can take on two values $\{0, 1\}$, there are two possible posterior distributions for the unlabeled y_i, i.e., $P(y_i|y_s = 0; y_{L(\boldsymbol{x})}, \boldsymbol{x})$ and $P(y_i|y_s = 1; y_{L(\boldsymbol{x})}, \boldsymbol{x})$. If $y_s = 1$ holds, then the Bayesian classification error is [15]:

$$\begin{aligned}\mathcal{E}(y_i|y_s = 1; y_{L(x)}, x) = \min\{&P(y_i = 1|y_s = 1; y_{L(x)}, x),\\ &P(y_i = 0|y_s = 1; y_{L(x)}, x)\}.\end{aligned} \tag{47}$$

Given the inequality $\frac{1}{2}H(p) - \epsilon \leq \min\{p, 1-p\} \leq \frac{1}{2}H(p), \epsilon = \frac{1}{2}\log\frac{5}{4}$ (see Fig. 15), we have

$$\begin{aligned}\frac{1}{2}H(y_i|y_s = 1; y_{L(\boldsymbol{x})}, \boldsymbol{x}) - \epsilon \leq \mathcal{E}(y_i|y_s = 1; y_{L(\boldsymbol{x})}, \boldsymbol{x})\\ \leq \frac{1}{2}H(y_i|y_s = 1; y_{L(\boldsymbol{x})}, \boldsymbol{x}).\end{aligned} \tag{48}$$

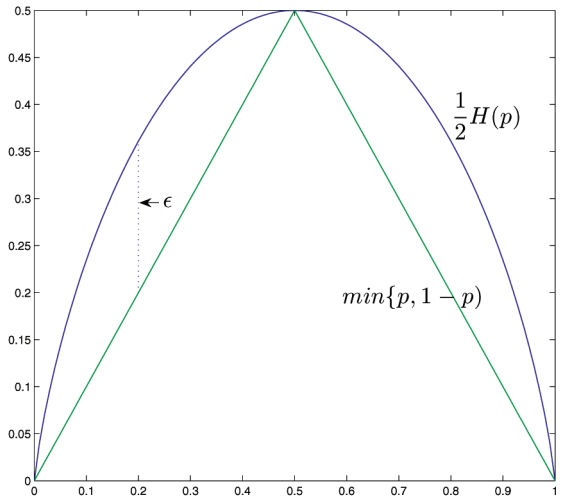

Fig. 15. Illustration of the inequality $\frac{1}{2}H(p) - \epsilon \leq \min\{p, 1-p\} \leq \frac{1}{2}H(p)$, $\epsilon = \frac{1}{2}\log\frac{5}{4}$.

Similarly, if $y_s = 0$ holds,

$$\begin{aligned}\frac{1}{2}H(y_i|y_s=0;y_{L(\boldsymbol{x})},\boldsymbol{x}) - \epsilon &\leq \mathcal{E}(y_i|y_s=0;y_{L(\boldsymbol{x})},\boldsymbol{x}) \\ &\leq \frac{1}{2}H(y_i|y_s=0;y_{L(\boldsymbol{x})},\boldsymbol{x}).\end{aligned} \tag{49}$$

Therefore, the Bayesian classification error bound given the selected y_s can be computed as

$$\begin{aligned}&\mathcal{E}(y_i|y_s;y_{L(\boldsymbol{x})},\boldsymbol{x}) \\ &= P(y_s=1|y_{L(\boldsymbol{x})},\boldsymbol{x})\mathcal{E}(y_i|y_s=1;y_{L(\boldsymbol{x})},\boldsymbol{x}) \\ &\quad + P(y_s=0|y_{L(\boldsymbol{x})},\boldsymbol{x})\mathcal{E}(y_i|y_s=0;y_{L(\boldsymbol{x})},\boldsymbol{x}) \\ &\leq \frac{1}{2}P(y_s=1|y_{L(\boldsymbol{x})},\boldsymbol{x})H(y_i|y_s=1;y_{L(\boldsymbol{x})},\boldsymbol{x}) \\ &\quad + \frac{1}{2}P(y_s=0|y_{L(\boldsymbol{x})},\boldsymbol{x})H(y_i|y_s=0;y_{L(\boldsymbol{x})},\boldsymbol{x}) \\ &= \frac{1}{2}H(y_i|y_s;y_{L(\boldsymbol{x})},\boldsymbol{x}).\end{aligned} \tag{50}$$

The last equality follows the definition of conditional entropy. And similarly

$$\begin{aligned}&\mathcal{E}(y_i|y_s;y_{L(\boldsymbol{x})},\boldsymbol{x}) \\ &= P(y_s=1|y_{L(\boldsymbol{x})},\boldsymbol{x})\mathcal{E}(y_i|y_s=1;y_{L(\boldsymbol{x})},\boldsymbol{x}) \\ &\quad + P(y_s=0|y_{L(\boldsymbol{x})},\boldsymbol{x})\mathcal{E}(y_i|y_s=0;y_{L(\boldsymbol{x})},\boldsymbol{x}) \\ &\geq \frac{1}{2}P(y_s=1|y_{L(\boldsymbol{x})},\boldsymbol{x})\{H(y_i|y_s=1;y_{L(\boldsymbol{x})},\boldsymbol{x}) - 2\epsilon\} \\ &\quad + \frac{1}{2}P(y_s=0|y_{L(\boldsymbol{x})},\boldsymbol{x})\{H(y_i|y_s=0;y_{L(\boldsymbol{x})},\boldsymbol{x}) - 2\epsilon\} \\ &= \frac{1}{2}H(y_i|y_s;y_{L(\boldsymbol{x})},\boldsymbol{x}) - \epsilon.\end{aligned}$$

□

REFERENCES

[1] L. Fei-Fei, R. Fergus, and P. Perona, "One-Shot Learning of Object Categories," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594-611, Apr. 2006.

[2] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrel, "Active Learning with Gaussian Processes for Object Categorization," *Proc. IEEE Int'l Conf. Computer Vision*, 2007.

[3] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative Multi-Label Video Annotation," *Proc. ACM Conf. Multimedia*, 2007.

[4] S. Zhu, X. Ji, W. Xu, and Y. Gong, "Multi-Labelled Classification Using Maximum Entropy Method," *Proc. Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, 2005.

[5] G.-J. Qi, Y. Song, X.-S. Hua, L.-R. Dai, and H.-J. Zhang, "Video Annotation by Active Learning and Cluster Tuning," *Proc. Int'l Workshop Semantic Learning Applications in Multimedia (in association with CVPR)*, 2006.

[6] S.C.H. Hoi and M.R. Lyu, "A Semi-Supervised Active Learning Framework for Image Retrieval," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, 2005.

[7] A. Dong and B. Bhanu, "Active Concept Learning for Image Retrieval in Dynamic Databases," *Proc. IEEE Int'l Conf. Computer Vision*, 2003.

[8] R. Yan, J. Yang, and A. Hauptmann, "Automatically Labeling Data Using Multi-Class Active Learning," *Proc. IEEE Int'l Conf. Computer Vision*, 2003.

[9] S. Tong and E.Y. Chang, "Support Vector Machine Active Learning for Image Retrieval," *Proc. ACM Conf. Multimedia*, 2001.

[10] E.Y. Chang, S. Tong, K. Goh, and C. Chang, "Support Vector Machine Concept-Dependent Active Learning for Image Retrieval," *IEEE Trans. Multimedia*, 2005.

[11] A. Krause, A. Singh, and C. Guestrin, "Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies," *J. Machine Learning Research*, vol. 9, pp. 235-284, 2008.

[12] X. Li, L. Wang, and E. Sung, "Multi-Label SVM Active Learning for Image Classification," *Proc. Int'l Conf. Image Processing*, 2004.

[13] M.R. Boutell, J. Luo, X. Shen, and C.M. Brown, "Learning Multi-Label Scene Classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757-1771, 2004.

[14] K. Brinker, "On Active Learning in Multi-Label Classification," *From Data and Information Analysis to Knowledge Eng.*, Springer, 2006.

[15] M.E. Hellman and J. Raviv, "Probability of Error, Equivocation, and the Chernoff Bound," *IEEE Trans. Information Theory*, vol. 16, no. 4, pp. 368-372, July 1970.

[16] A. Kapoor and E. Horvitz, "On Discarding, Caching, and Recalling Samples in Active Learning," *Proc. Conf. Uncertainty and Artificial Intelligence*, 2007.

[17] F. Jing, M. Li, and H.-J. Zhang, "Entropy-Based Active Learning with Support Vector Machine for Content-Based Image Retrieval," *Proc. IEEE Int'l Conf. Multimedia and Expo*, 2004.

[18] N. Roy and A. McCallum, "Toward Optimal Active Learning through Sampling Estimation of Error Reduction," *Proc. Int'l Conf. Machine Learning*, 2001.

[19] T. Cover and J. Thomas, *Elements of Information Theory*, second ed. John Wiley and Sons, 2006.

[20] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang, "Two Dimensional Active Learning for Image Classification," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, 2008.

[21] C. Bishop, "Approximate Inference," *Pattern Recognition and Machine Learning*, pp. 461-473, Springer, 2006.

[22] S.F. Chen and R. Rosenfeld, "A Gaussian Prior for Smoothing Maximum Entropy Models," Technical Report CMU-CS-99-108, School of Computer Science, Carnegie Mellon Univ., 1999.

[23] J. Wu, X.-S. Hua, and B. Zhang, "Tracking Concept Drifting with Gaussian Mixture Model," *Proc. Int'l Conf. Visual Comm. and Image Processing*, 2005.

[24] D.C. Liu and J. Nocedal, "On the Limited Memory BFGS Method for Large Scale Optimization," *Math. Programming B*, vol. 45, nos. 1-3, pp. 503-528, 1989.

[25] N. Syed, H. Liu, and K. Sung, "Incremental Learning with Support Vector Machines," *Proc. Workshop Support Vector Machines, at the Int'l Joint Conf. Artificial Intelligence*, 1999.

[26] G. Cauwenberghs and T. Poggio, "Incremental and Decremental Support Vector Machine," *Proc. Conf. Neural Information Processing Systems*, 2000.

[27] J. Yang, R. Yan, and A. Hauptmann, "Cross-Domain Video Concept Detection Using Adaptive svms," *Proc. ACM Conf. Multimedia*, 2007.

[28] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum-Likelihood from Incomplete Data via EM Algorithm," *J. Royal Statistical Soc.*, vol. 39, no. 1, pp. 1-38, 1977.
[29] R. Neal and G. Hinton, *A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants*, M. Jordan, ed. Kluwer Academic Press, 1998.
[30] R.M. Neal, "Probabilistic Inference Using Markov Chain Monte Carlo Methods," Technical Report CRG-TR-93-1, Univ. of Toronto, 1993.
[31] B.J. Frey and D.J.C. MacKay, "A Revolution: Belief Propagation in Graphs with Cycles," *Advances in Neural Information Processing Systems*, vol. 10, MIT Press, 1998.
[32] T. Minka, "Expectation Propagation for Approximate Bayesian Inference," *Proc. 17th Conf. Uncertainty in Artificial Intelligence*, 2001.
[33] K.P. Murphy, Y. Weiss, and M.I. Jordan, "Loopy Belief Propagation for Approximate Inference: An Empirical Study," *Proc. Conf. Uncertainty in Artificial Intelligence*, 1999.
[34] T. Volkmer, J.R. Smith, and A. Natsev, "A Web-Based System for Collaborative Annotation of Large Image and Video Collections," *Proc. ACM Int'l Conf. Multimedia*, 2005.
[35] A. Elisseeff and J. Weston, "A Kernel Method for Multi-Labelled Classification," *Proc. Conf. Neural Information Processing Systems*, 2002.
[36] L. Fei-Fei and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, 2005.
[37] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, Z.-J. Zha, and H.-J. Zhang, "A Joint Appearance-Spatial Distance for Kernel-Based Image Categorization," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, 2008.
[38] B. Merialdo, J. Jiten, E. Galmar, and B. Huet, "A New Approach to Probabilistic Image Modeling with Multidimensional Hidden Markov Models," *Proc. Fourth Int'l Workshop Adaptive Multimedia Retrieval*, 2006.
[39] C.G.M. Snoek, M. Worring, J.C. Gemert, J.-M. Geusebroek, and A.W.M. Smeulders, "The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia," *Proc. ACM Int'l Conf. Multimedia*, pp. 421-430, Oct. 2006.

Guo-Jun Qi received the BE degree in automation from the University of Science and Technology of China, Hefei, Anhui, China, in 2005. He is currently with the Image Formation and Processing Group at the University of Illinois at Urbana-Champaign. From July 2006 to November 2008, he worked as a research intern in the Internet Media Group at Microsoft Research Asia. His research interests include computer vision, multimedia, and machine learning, especially content-based image/video retrieval, analysis, management, and sharing. He was the winner of the Best Paper Award at the 15th ACM International Conference on Multimedia, Augsburg, Germany, 2007. He is a student member of the ACM.

Xian-Sheng Hua received the BS and PhD degrees in applied mathematics from Peking University, Beijing, China, in 1996 and 2001, respectively. Since 2001, he has been with Microsoft Research Asia, Beijing, where he is currently a lead researcher with the Internet Media Group. His current research interests include video content analysis, multimedia search, management, authoring, sharing, and advertising. He has authored more than 130 publications in these areas and has more than 30 filed patents or pending applications. He is an adjunct professor at the University of Science and Technology of China, and serves as an associate editor of the *IEEE Transactions on Multimedia* and an editorial board member of *Multimedia Tools and Applications*. He won the Best Paper Award and the Best Demonstration Award at ACM Multimedia 2007 and also won the TR35 2008 Young Innovator Award from the *MIT Technology Review*. He is a member of the ACM and the IEEE.

Yong Rui received the BS degree from Southeast University, the MS degree from Tsinghua University, and the PhD degree from the University of Illinois at Urbana-Champaign (UIUC). He serves as the director of Strategy of the Microsoft China R&D (CRD) Group. Prior to this, he spent seven years managing the Multimedia Collaboration Team at Microsoft Research Redmond, Washington. He is an associate editor of the *ACM Transactions on Multimedia Computing, Communication, and Applications* (TOMCCAP), the *IEEE Transactions on Multimedia*, and *the IEEE Transactions on Circuits and Systems for Video Technologies*. He was an editor of the *ACM/Springer Multimedia Systems Journal* (2004-2006), the *International Journal of Multimedia Tools and Applications* (2004-2006), and the *IEEE Transactions on Multimedia* (2004-2008). He serves on the Advisory Board of the *IEEE Transactions on Automation Science and Engineering* and also holds an Executive Training Certificate from the Wharton Business School at the University of Pennsylvania. He contributes significantly to the research communities in computer vision, signal processing, machine learning, and their applications in communication, collaboration, and multimedia systems. His contribution to relevance feedback in image search created a new research area in multimedia. He has published 12 books and book chapters and more than 70 refereed journal and conference papers. He holds 30 issued and pending US patents. He was on the organizing committees and program committees of ACM Multimedia, IEEE CVPR, IEEE ECCV, IEEE ACCV, IEEE ICIP, IEEE ICASSP, IEEE ICME, SPIE ITCom, ICPR, CIVR, among others. He was a general chair of the International Conference on Image and Video Retrieval (CIVR) 2006, a program chair of ACM Multimedia 2006, and a program chair of the Pacific-Rim Conference on Multimedia (PCM) 2006. He is a senior member of the ACM and the IEEE.

Jinhui Tang received the BE and PhD degrees from the University of Science and Technology of China in July 2003 and July 2008, respectively. He is currently a postdoctoral research fellow in the School of Computing at the National University of Singapore. From June 2006 to February 2007, he worked as a research intern in the Internet Media Group at Microsoft Research Asia, and, from February 2008 to May 2008, he worked as a research intern in the School of Computing at the National University of Singapore. He is a recipient of the 2008 President Scholarship of the Chinese Academy of Science and a corecipient of the Best Paper Award at ACM Multimedia 2007. His current research interests include content-based image retrieval, video content analysis, and pattern recognition. He is a member of the ACM and a student member of the IEEE.

Hong-Jiang Zhang received the BS and PhD degrees in electrical engineering from Zhengzhou University, Henan, China, in 1982, and the Technical University of Denmark, Lyngby, in 1991, respectively. From 1992 to 1995, he was with the Institute of Systems Science, National University of Singapore, where he led several projects in video and image content analysis and retrieval and computer vision. From 1995 to 1999, he was a research manager at Hewlett-Packard Labs, Palo Alto, California, where he was responsible for research and development in the areas of multimedia management and intelligent image processing. In 1999, he joined Microsoft Research, where he is currently the managing director of the Advanced Technology Center in Beijing. He has coauthored/coedited four books, more than 350 papers and book chapters, numerous special issues of international journals on image and video processing, content-based media retrieval, and computer vision, as well as more than 60 granted patents. He currently serves as the editor-in-chief of the *IEEE Transactions on Multimedia* and is on the editorial board of the *Proceedings of IEEE*. He is a fellow of the IEEE and the ACM.