Online Multi-Label Active Learning for Large-Scale Multimedia Annotation

Xian-Sheng Hua
Microsoft Research Asia, Beijing, China
xshua@microsoft.com

Guo-Jun Qi[†]
University of Science and Technology of China, Hefei, China
qgj@mail.ustc.edu.cn

ABSTRACT

Existing video search engines have not taken the advantages of video content analysis and semantic understanding. Video search in academia uses semantic annotation to approach content-based indexing. We argue this is a promising direction to enable real content-based video search. However, due to the complexity of both video data and semantic concepts, existing techniques on automatic video annotation are still not able to handle large-scale video set and large-scale concept set, in terms of both annotation accuracy and computation cost. To address this problem, in this paper, we propose a scalable framework for annotation-based video search, as well as a novel approach to enable large-scale semantic concept annotation, that is, online multi-label active learning. This framework is scalable to both the video sample dimension and concept label dimension. Large-scale unlabeled video samples are assumed to arrive consecutively in batches with an initial pre-labeled training set, based on which a preliminary multi-label classifier is built. For each arrived batch, a multi-label active learning engine is applied, which automatically selects and manually annotates a set of unlabeled sample-label pairs. And then an online learner updates the original classifier by taking the newly labeled sample-label pairs into consideration. This process repeats until all data are arrived. During the process, new labels, even without any pre-labeled training samples, can be incorporated into the process anytime. Experiments on TRECVID dataset demonstrate the effectiveness and efficiency of the proposed framework.

Keywords

Video Annotation, Online Learning, Multi-Label Active Learning.

1. INTRODUCTION

We will propose a novel semantic annotation scheme to enable content-based video search, which is scalable to large-scale video samples as well as large-scale semantic concepts.

Existing video search engines are all based on indexing "direct text" with few support from video content understanding. That is, surrounding text (the text in a webpage that may be related to the video), video descriptions (the textual description of the target video, including title, author, content description, tags, comments, etc.), recognized speech, and video metadata (such as format, bitrates, frame size, etc) are used to index videos. The advantage of this design is that, we can directly use text-based indexing and ranking technologies to do video search. The disadvantage is that this "direct text" frequently is far from the real content of the video.

Video search in academia, mainly under TRECVID dataset [26-32, 35], moves one-step ahead, which takes content analysis, especially semantic concept annotation, into consideration. Machine learning techniques are applied to convert visual information into textual description to realize content-based video search. We argue this is a promising direction to enable real content-based video search. However, due to the complexity of both video data and semantic concepts, existing techniques on automatic video annotation are suffering the difficulties in dealing with large-scale video set and large-scale concept set, in terms of both annotation accuracy and computation cost. Generally, to overcome these difficulties, huge training datasets and high computation powers are required, but they are difficult, if not impossible, to be obtained in many real-world large-scale problems.

To address this difficulty, in this paper, we propose a scalable framework (in terms of both video samples and concept labels) for annotation-based video search and a novel online multi-label active learning approach to enable large-scale semantic concept annotation. In this framework, large-scale unlabeled video samples are assumed to arrive consecutively in batches with an initial pre-labeled training set as the first batch, based on which an initial multi-label classifier is built. For each arrived batch, an online multi-label active learning engine is applied to efficiently update the classifier which maximizes its performance on all currently-available data. This process repeats until all data are arrived and resumes when a new data batch is available, during which new concept labels are allowed to be introduced into the online multi-label active learning framework at any batch, even though these labels have no any pre-labeled training samples.

The core approach, online multi-label active learning (Online MLAL), has three major modules, *multi-label active learning*, *online multi-label learning* and *new label learning from zero-knowledge*.

Active learning is well-known for its efficiency in saving labeling cost by exploiting the redundancy in samples. The multi-label active learning strategy in this paper exploits the redundancy both in samples and semantic labels to further reduce the labor cost of labeling. The primary difference between multi-label active learning and typical active learning is that, we will iteratively ask people to confirm the labels of a selected set of *sample-label pairs* instead of *samples* (with all labels by default) to minimize an estimated classification error.

Online learning is designed to reduce the computational cost in multi-label active learning. Generally, model retraining is required when new sample-label pairs are obtained, which is computation-intensive. The computation cost increases remarkably with the increase of labeled sample-label pairs. On the contrary, online multi-label learning is able to incrementally update the multi-label classifier by adapting the original classifier to the newly labeled data, which significantly saves computation cost. Different from existing online learning approaches, the one proposed in this paper exploits the correlations among multiple labels to improve the performance of the classifier.

New label learning is to make the proposed framework scalable to new semantic labels. Existing semantic annotation schemes are only applicable for a closed concept set, which is not practical for real-world video search engines. The aforementioned online learner can be naturally extended to handling new labels, even though these new labels have no any pre-labeled training data. The annotation performance of the new labels will be gradually improved through the iterative active learning process.

1.1 Related Work

We will review related work along three threads, including multi-label learning, active learning and online learning.

1.1.1 Multi-Label Learning for Video Annotation

A large semantic concept corpus is desired to enable effective semantic annotation based multimedia search. Therefore, multimedia annotation is actually a multi-label classification problem. Research on multi-label video annotation evolved through three paradigms [3]: individual concept annotation, context-based conceptual fusion (CBCF) [20] annotation, and integrated multi-label annotation. A detailed review of these three paradigms can be found in our previous work [3].

The proposed multi-label active learning approach in this paper is based on a similar idea to the third paradigm of multi-label annotation scheme, which also exploits the correlations or redundancy among different labels. Instead of using a SVM-like algorithm in [3], which is difficult to handle large-scale data, we use an online learner in the proposed approach to handle the problems of high computation cost as well as “label incompletion”. It is to be detailed in Section 3 and 4.

1.1.2 Active Learning for Video Annotation

Active learning is one of the widely-used approaches in image and video classification, as it can significantly reduce human cost in labeling training samples [5-7]. Specifically, active learning approaches iteratively annotate a set of elaborately selected samples so that the expected generalization error is minimized in each step. As a result, the total number of training samples that need to be labeled in active learning is smaller than that in non-active learning approaches. It is clear that one of the core problems of active learning approaches is the sample selection strategy. In the past decade, a number of active learning approaches were developed by using different sample selection strategies [8, 9, 2, 5]. Most of these approaches focus on the binary classification scenario. However, in many real-world applications [10, 11, 3], a sample is usually associated with multiple labels rather than a single one. Under such a multi-label setting, each sample will be annotated as “positive” or “negative” for each and every label. As a result, active learning with multi-labeled samples is much more challenging than that with binary-labeled ones, especially when the number of labels is large.

A direct way to tackle active learning under multi-label 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 [10, 12, 33, 34] three research groups have proposed or evaluated such binary-based active learning algorithms for multi-label classification problem. However, this type of approaches does not take into account the inherent relationships among multiple labels.

The proposed multi-label active learning (MLAL) instead iteratively selects sample-label pairs to minimize the expected classification error. Specifically, in each iteration, the annotators are only required to annotate/confirm selected labels of selected samples while the remaining unlabeled part will be inferred according to the label correlations. An intuitive explanation of this strategy is that there exist both sample and label redundancies for multi-labeled samples.

1.1.3 Online Learning

The main purpose of online learning is to save computation cost by incrementally updating classifiers in which new training samples are taken into account, instead training the classifier again from the beginning on the combined training set. For example, [16] proposes an incremental online algorithm to update the SVM model by training it with the previous support vectors and the newly arrived samples. Gauwenberghs et al. [17] present not only an incremental SVM but also a decremental one. Recently, Yang et al. [18] proposes a cross-domain online algorithm based on SVM to adapt it to a set of cross-domain dataset with a different distribution. All these algorithms are designed for single labeled samples. To the best of our knowledge, we are the first to develop a multi-label online algorithm by considering the correlations between different labels. Moreover, this new online algorithm also allows introducing new labels without any pre-labeled training data.

2. A NOVEL DESIGN OF CONTENT-BASED VIDEO SEARCH

As aforementioned, existing video search engines have not taken the advantages of video content analysis, and the video search approaches in academia uses video annotation but existing annotation scheme is not scalable to the size of video set and concept set. The proposed new design of content-based video search system is also based on semantic annotation, but the difficulty on scaling up is addressed by actively leveraging users’ interactions during the learning process, which is realized through the proposed online multi-label active learning algorithm.

Figure 1 shows the work flow of how the entire video dataset is annotated with online active learning. We assume the size of this dataset is keeping increasing, simulating the increase of the video database in a real video search engine due to the continuous video crawling. And we also assume the dataset is increasing batch by batch, denoted by B_0, $B_{1,}$ …, and B_N. For simplicity, we also use the same batch symbol to denote the set of sample-label pairs in the corresponding batch (a batch with n samples and m semantic concepts will have $m \times n$ sample-label pairs), and we assume B_0 is the initial pre-labeled training set. We denote the initial classifier built on B_0 as P^0, Then the learning procedure of the online multi-label active learning approach can be summarized as follows.

(1) *Active Learning on* (B_0+B_1). Based on the knowledge in classifier $P^{\,0}$, an iterative multi-label active learning process is applied on B_1. In each round, a certain number of sample-label pairs are selected to be annotated manually, and a new classifier will be built through an online learner based on current classifier and the newly labeled data. We denote all the selected sample-label pairs in this multiple-round active learning process as L_1 (a subset of B_1), and the final new classifier is $P^{\,1}$, which is gradually built by the online learner based on $P^{\,0}$ and L_1.

(2) *From the iteration* t = 2 *to N*, *Active Learning on* (B_0+B_1+…+ B_t). Based on the knowledge in classifier $P^{\,t\text{-}1}$, the active learning process is applied on the set of all available unlabeled sample-pairs (B_0+B_1+…+ B_t)\(L_1+…+ $L_{t\text{-}1}$). We denote all the selected sample-label pairs in this multiple-round active learning process as L_t, and the final new classifier is $P^{\,t}$, which is built by the online learner based on $P^{\,t\text{-}1}$ and L_t step-by-step.

(3) *Learning New Label*s. During any step in (2), the multi-label classifier can be extended to handle new labels (which can be obtained from query log analysis, for example, from the list of query terms that are most-frequently used), and with the arrival of next data batch, the new sample-label pair set will cover the new labels, and they will be selected by the sample-label pair selection engine in the active learning prodedure. The correlations between the new labels and existing labels will be gradually exploited with the increase of labeled sample-label pairs.

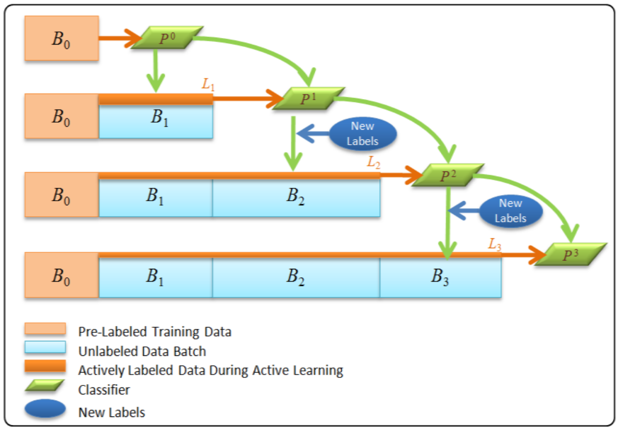


Figure 1. The Process of Online MLAL.

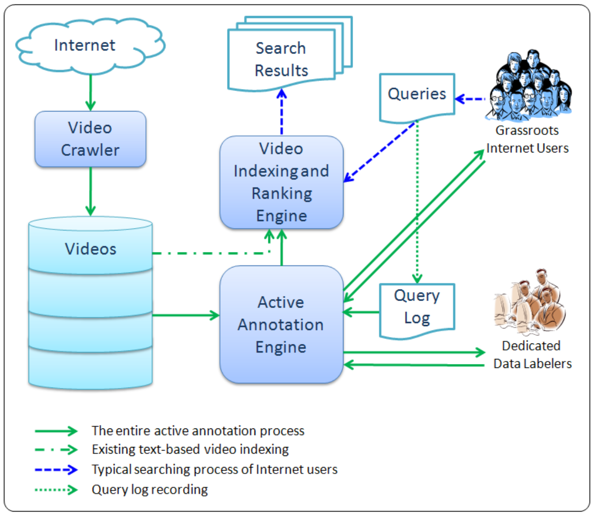

Internet
Search Results
Queries
Grassroots Internet Users
Video Crawler
Video Indexing and Ranking Engine
Videos
Active Annotation Engine
Query Log
Dedicated Data Labelers
The entire active annotation process
Existing text-based video indexing
Typical searching process of Internet users
Query log recording

Figure 2. New Design of Annotation-Based Video Search.

With the support of the above online MLAL process, Figure 2 shows a new design of annotation-based video search system. The main difference between this design and existing systems is that we have two groups of people involved in the annotation process, that is, dedicated data labelers and grassroots users who may also contribute to data labeling. All the labeling requests in the above active annotation process will be delivered to these two types of labelers. We argue it is rational for a video search engine to have a certain number of dedicated data labelers, though grassroots labelers only is also workable for the video search system.

To attract average Internet users to label given data, the following scenarios can be applied.

(1) *Game*. Designing attractive games and during playing game the players will be asked to confirm labels of video clips with a friendly interface. A good example is the ESP game [1], but it has no support from learning algorithms.

(2) *Pay*. Pay by the number of labeled sample-label pairs. The pay can be real currency or virtual ones which can be used to buy online products/content. A good example is to use MTurk [38].

(3) *reCAPTCHA*. CAPTCHA is a type of challenge-response test used to determine that the response is not generated by a computer. A typical CAPTCHA can be generating an image with distorted text which is supposed that it can only be recognized by human beings [24]. Recently some researchers have implemented methods by which some of the effort and time spent by people who are responding to CAPTCHA challenges can be regarded as a distributed work system. This system, called reCAPTCHA, includes "solved" and "unrecognized" elements (such as images of text which were not successfully recognized via OCR) in each challenge. The respondent thus answers both elements and roughly half of his or her effort validates the challenge while the other half is collected as useful information [25]. This idea can also be applied to do image and video labeling.

The design of the above incentive programs to attractive grassroots users to label data is out of the scope of this paper. And, if sufficient grassroots join the process, one sample-label pair can be confirmed by multiple users, which is able to reduce noises (grassroots labeling generally has lower quality compared with dedicated labelers). Moreover, how to improve and evaluate the labeling quality for both dedicated labelers and grassroots labelers, as well as anti-spam, are also important research topics, which are also out of the scope of this paper. Some attempts can be found in [40]. Another point that is worth to mention is that, although this paper takes video search as the underlying application, this active learning based strategy is applicable for improving image search and even text search relevance.

3. MULTI-LABEL ACTIVE LEARNING

The research work on multi-label active learning has been published in our previous work [14], but for completion of this proposed system, we will briefly introduce the basic idea and the algorithm workflow.

Different from the typical binary active learning formulation that selects the most "informative" samples for manual labeling, we jointly select both the samples and labels simultaneously. The underlying assumption is that different labels of a certain sample have different contributions to minimizing the expected classification error of the to-be-trained classifier. And annotating a well-selected portion of labels may provide sufficient information for learning the classifier.

In essence, the multi-label classifiers do have uncertainty along different labels as well as different samples. Traditional active learning algorithms can be seen as a one-dimension active selection approach, which only reduces the sample uncertainty. In contrast, multi-label active learning is a two-dimensional active learning strategy, which selects the most "informative" *sample-label pairs* to reduce the uncertainty along the dimensionalities of both samples and labels. More specifically, along label dimension all of the labels correlatively interact. Therefore, once partial labels are annotated, the left unlabeled concepts can then be inferred based on label correlations.

This approach significantly saves the labor cost for data labeling compared with fully annotating all labels. Thus, it is far more efficient when the number of labels is huge. For instance, an image may be associated with thousands and hundreds of concepts. That means a full annotation strategy will pay large labor costs for only one image. On the contrary, multi-label active learning only manually annotates the most informative labels.

It is worth noting that during multi-label active learning process, some samples may be lack of some labels since only a partial of labels are annotated. This is different from traditional active learning algorithm. The missing labels for a certain sample can be seen as hidden variables and the corresponding classifier with such incomplete labeling can be trained by the standard Expectation-Maximum (EM) [19] algorithm accordingly.

Before we move further, we first define some notations. For each sample $\boldsymbol{x}$, it has m labels $y_i(1 \le i \le m)$ and each of them indicates whether its corresponding semantic concept occurs. As stated before, in each active learning iteration, some of these labels have already been annotated while others not. Let $U(\boldsymbol{x}) = \{i \mid (\boldsymbol{x}, y_i)$ is unlabeled sample-label pair.$\}$ denote the set of indices of unlabeled part and $L(\boldsymbol{x}) = \{i \mid (\boldsymbol{x}, y_i)$ is labeled sample-label pair.$\}$ denote the labeled part. Note that $L(\boldsymbol{x})$ can be an empty set Ø, 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 $P(\boldsymbol{x})$ be the marginal sample distribution.

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

$$\xi^b(\mathbf{P}) = \frac{1}{|\mathbf{P}|}\sum_{x \in P} \xi(\boldsymbol{y} \mid y_{L(x)}, \boldsymbol{x}) \tag{4}$$

We can use the above classification error on the pool to estimate the expected error over the full distribution $P(\boldsymbol{x})$, i.e., $E_{P(x)}\xi(\boldsymbol{y} \mid y_{L(x)}, \boldsymbol{x}) = \int P(\boldsymbol{x})\xi(\boldsymbol{y} \mid y_{L(x)}, \boldsymbol{x})d\boldsymbol{x}$, 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 **P** is

$$\begin{aligned}\xi^a(\mathbf{P}) &= \frac{1}{|\mathbf{P}|}\{\xi(\boldsymbol{y} \mid y_s; y_{L(x_s)}, \boldsymbol{x}_s) + \textstyle\sum_{x\in\mathbf{P}\setminus x_s} \xi(\boldsymbol{y} \mid y_{L(x)}, \boldsymbol{x})\} \\ &= \frac{1}{|\mathbf{P}|}\{\xi(\boldsymbol{y} \mid y_s; y_{L(x_s)}, \boldsymbol{x}_s) - \xi(\boldsymbol{y} \mid y_{L(x_s)}, \boldsymbol{x}_s)\} + \textstyle\sum_{x\in\mathbf{P}} \xi(\boldsymbol{y} \mid y_{L(x)}, \boldsymbol{x})\end{aligned} \tag{5}$$

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

$$\Delta\xi(\mathbf{P}) = \xi^b(\mathbf{P}) - \xi^a(\mathbf{P}) \tag{6}$$

Thus our goal is to select a best $(\boldsymbol{x}^*_s, y^*_s)$ to maximize the above expected error reduction. That is,

$$\begin{aligned}(\boldsymbol{x}^*_s, y^*_s) &= \arg\max_{\boldsymbol{x}_s\in\mathbf{P}, y_s\in U(\boldsymbol{x}_s)} \Delta\xi(\mathbf{P}) \\ &= \arg\min_{\boldsymbol{x}_s\in\mathbf{P}, y_s\in U(\boldsymbol{x}_s)} -\Delta\xi(\mathbf{P})\end{aligned} \tag{7}$$

Applying Lemma 1 and Theorem 1 in [14], we have

$$\begin{aligned}-\Delta\xi(\mathbf{P}) &= \xi^a(\mathbf{P}) - \xi^b(\mathbf{P}) \\ &\le \frac{1}{|\mathbf{P}|}\{\varepsilon - \frac{1}{2m}\textstyle\sum_{i=1}^{m} MI(y_i; y_s \mid y_{L(\boldsymbol{x}_s)}, \boldsymbol{x}_s)\}\end{aligned} \tag{8}$$

where $MI(y_i; y_s \mid y_{L(x_s)}, x_s)$ is the mutual information between the random variables y_i and y_s given the known label x_s. Consequently, by minimizing the obtained error bound in Eqn. (8), we can select the sample-label pair for annotation as

$$\begin{aligned}&(\boldsymbol{x}^*_s, y^*_s) \\ &= \arg\min_{\boldsymbol{x}_s\in\mathbf{P}, y_s\in U(\boldsymbol{x}_s)} \frac{1}{|\mathbf{P}|}\{\varepsilon - \frac{1}{2m}\textstyle\sum_{i=1}^{m} MI(y_i; y_s \mid y_{L(x_s)}, \boldsymbol{x}_s)\} \\ &= \underset{\boldsymbol{x}_s\in\mathbf{P}, y_s\in U(\boldsymbol{x}_s)}{\arg\max} \textstyle\sum_{i=1}^{m} MI(y_i; y_s \mid y_{L(x_s)}, \boldsymbol{x}_s) \\ &= \underset{\boldsymbol{x}_s\in\mathbf{P}, y_s\in U(\boldsymbol{x}_s)}{\arg\max} \left\{H(y_s \mid y_{L(x_s)}, x_s) + \sum_{i=1, i\neq s}^{m} MI(y_i; y_s \mid y_{L(x_s)}, x_s)\right\}\end{aligned} \tag{9}$$

As this multi-label active learning strategy exploits the redundancy along sample dimension and label dimension simultaneously, we also name it Two-Dimensional Active Learning (MLAL). On the contrary, conventional single label active learning approaches are called One-Dimensional Active Learning (SLAL). More details and analysis of this sample-label pair selection strategy can be found in reference [14].

4. ONLINE MULTI-LABEL LEARNING

Once new sample-label pairs are selected according to the multi-label active learning strategy, the statistical model for multi-labeled images needs to be updated accordingly. However, as stated in Section I, the conventional offline algorithms retrain a new model on the 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 algorithm is desired to adapt the old model to the new samples without retraining. In this section we introduce such an online learner which has been detailed in our previous work [36]. Intuitively, such an online classification algorithm should satisfy the following requirements:

(1) It preserves the old knowledge that has already existed in the old model. This knowledge captures the rich historical information about the previously-acquired training samples.

(2) It reveals the information contained in the newly-arrived multi-labeled samples. In contract to the traditional binary based algorithm (e.g., one-against-rest SVM), the label correlations are modeled by this online learner.

In this section, we will propose 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 multi-label classification model $P^{\tau}(\boldsymbol{y}|\boldsymbol{x})$, which is trained from the historically-acquired images. Then a set of newly-labeled images and their corresponding ground truth $\{x_i, y_i\}_{i=1}^{n}$ is obtained in each active learning iteration. 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 new samples $\{x_i, y_i\}_{i=1}^{n}$ .In contrast to the retraining-based learning algorithm, the online learner does not utilize the historical training set but only the old model and new coming samples. Actually, the online learner assumes the information about the historical training samples has been preserved in the old model $P^{\tau}(\boldsymbol{y}|\boldsymbol{x})$, because this model is learned from historical training samples. Thus with a small number of new training samples, the online learner trains a new model much more efficiently.

As aforementioned, this new model $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ should satisfy the two requirements: preserving the existing knowledge in $P^{\tau}(\boldsymbol{y}|\boldsymbol{x})$ while revealing the information in new samples $\{x_i, y_i\}_{i=1}^{n}$. These two requirements can be satisfied by formulating the following probabilistic variational problem, where Kullback-Leibler Divergence (KLD) [13] is applied to measure the degree of the new model preserving the existing knowledge contained in the old one, under a set of multi-label constraints revealing the information contained in the new samples:

$$\hat{P}^{\tau+1}(\boldsymbol{y} \mid \boldsymbol{x}) = \arg\min_{P^{\tau+1}} \left\langle D_{KL}(P^{\tau+1}(\boldsymbol{y} \mid \boldsymbol{x}) \,\|\, p^{\tau}(\boldsymbol{y} \mid \boldsymbol{x})) \right\rangle_{\tilde{P}} \tag{11}$$

$$s.t. \quad \left\langle y_i \right\rangle_{P^{\tau+1}} = \left\langle y_i \right\rangle_{\tilde{P}} + \eta_i, 1 \le i \le m \tag{12}$$

$$\left\langle y_i y_j \right\rangle_{P^{\tau+1}} = \left\langle y_i y_j \right\rangle_{\tilde{P}} + \theta_{ij}, 1 \le i < j \le m \tag{13}$$

$$\left\langle y_i x_l \right\rangle_{P^{\tau+1}} = \left\langle y_i x_l \right\rangle_{\tilde{P}} + \varphi_{il}, 1 \le i \le m, 1 \le l \le d \tag{14}$$

$$\sum\nolimits_{y} P^{\tau+1}(\boldsymbol{y} \mid \boldsymbol{x}) = 1 \tag{15}$$

$$\sum_{i} \frac{\eta_i^2}{2\sigma_\eta^2 / n} + \sum_{i<j} \frac{\theta_{ij}^2}{2\sigma_\theta^2 / n} + \sum_{i,l} \frac{\phi_{il}^2}{2\sigma_\phi^2 / n} \le C \tag{16}$$

where $\left\langle D_{KL}(P^{\tau+1}(\boldsymbol{y} \mid \boldsymbol{x}) \,\|\, P^{\tau}(\boldsymbol{y} \mid \boldsymbol{x})) \right\rangle_{\tilde{P}}$ 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 $\tilde{P}(\boldsymbol{x}) = \frac{1}{m}\sum\nolimits_{i=1}^{m} \delta(\boldsymbol{x} - \boldsymbol{x}_i)$ taken from $\{\boldsymbol{x}_i\}_{i=1}^{n}$ ($\delta(\cdot)$ is the indicator function). It is worth noting that joint model distribution $P^{\tau+1}(\boldsymbol{x}, \boldsymbol{y}) = P^{\tau+1}(\boldsymbol{y} \mid \boldsymbol{x})\tilde{P}(\boldsymbol{x})$, i.e., we only care about the conditional distribution $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ and thus use the sample frequency $\tilde{P}(x)$ on the training samples to approximate the true sample distribution $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$. d is the dimension of the feature vector $\boldsymbol{x}$ and x_i represents its i-th element.

Constraints (12) - (14) constrain 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 [14] [4]. $\eta_i \sim N(0, \sigma_\eta^2)$, $\theta_{il} \sim N(0, \sigma_\theta^2)$ 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 caused by the noises in the training samples. As suggested in [15] [4], they assume the joint probability of estimation errors should be reasonably large, and thus we can add Eqn. (16) as a constraint of Eqn. (11).

As aforementioned, when learning the new model we need to balance between the existing knowledge and the new information. Actually, the Gaussian error estimations in (12) (13) (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 multi-label constraints become more relaxed with relatively large noises η_i, θ_{ij} and ϕ_{il} in the new training sample set. On the contrary, the small variances will make $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ bias on the new information in $\{x_i, y_i\}_{i=1}^{n}$. Extremely, the removal of these error estimations will lead to a new model that completely complies with the new information.

From the above formulations, by formulating their dual problem according to Karush-Kuhn-Tucker (KKT) conditions, the solution of $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ can be found as

$$P^{\tau+1}(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{1}{Z^{\tau+1}(\boldsymbol{x})} P^{\tau}(\boldsymbol{y} \mid \boldsymbol{x}) \exp\{\boldsymbol{y}^T(\boldsymbol{b} + \boldsymbol{R}\boldsymbol{y} + \boldsymbol{W}\boldsymbol{x})\} \tag{18}$$

where $\boldsymbol{b},\boldsymbol{R},\boldsymbol{W}$ are Lagrangian multipliers, in which $\boldsymbol{b} = [b_1, b_2, \cdot\cdot\cdot, b_m]^T$ is a $m \times 1$ colum vector, $\boldsymbol{R} = [R_{ij}]_{m\times m}$ is a strict upper matrix with $R_{ij} = 0$ for $i \ge j$, $\boldsymbol{W} = [W_{ij}]_{m\times d}$ is a $m\times d$ matrix, and

$$Z^{\tau+1}(\boldsymbol{x}) = \sum\nolimits_{y} P^{\tau}(\boldsymbol{y} \mid \boldsymbol{x}) \exp\{\boldsymbol{y}^T(\boldsymbol{b} + \boldsymbol{R}\boldsymbol{y} + \boldsymbol{W}\boldsymbol{x})\} \tag{19}$$

is the partition function. The derivation of the above results is missed here due to limited space and the detailed derivation can be found in [36].

The parameters $\boldsymbol{b}$, $\boldsymbol{R}$, $\boldsymbol{W}$ can be solved by the following dual problem:

$$
\begin{aligned}
&\boldsymbol{b}^*, \boldsymbol{R}^*, \boldsymbol{W}^* = \arg\max_{b,R,W} L(\boldsymbol{b}, \boldsymbol{R}, \boldsymbol{W}) \\
&\arg\max_{b,R,W} \left\langle \boldsymbol{y}^T(\boldsymbol{b} + \boldsymbol{R}\boldsymbol{y} + \boldsymbol{W}\boldsymbol{x}) - \log Z^{\tau+1}(\boldsymbol{x}) \right\rangle_{\tilde{P}} \\
&-\frac{\alpha_b}{2n}\|\boldsymbol{b}\|_2^2 - \frac{\alpha_R}{2n}\|\boldsymbol{R}\|_F^2 - \frac{\alpha_W}{2n}\|\boldsymbol{W}\|_F^2 \\
&= \arg\max_{b,R,W} \sum_{i=1}^{n} \{\boldsymbol{y}_i^T(\boldsymbol{b} + \boldsymbol{R}\boldsymbol{y}_i + \boldsymbol{W}\boldsymbol{x}_i) - \log Z^{\tau+1}(\boldsymbol{x}_i)\} \\
&-\frac{\alpha_b}{2n}\|\boldsymbol{b}\|_2^2 - \frac{\alpha_R}{2n}\|\boldsymbol{R}\|_F^2 - \frac{\alpha_W}{2n}\|\boldsymbol{W}\|_F^2
\end{aligned}
\tag{20}
$$

where $\| \|_2$ and $\| \|_F$ are norm-2 and Frobenius norm respectively. Here, $-\frac{\alpha_b}{2n}\|b\|_2^2 - \frac{\alpha_R}{2n}\|R\|_F^2 - \frac{\alpha_W}{2n}\|W\|_F^2$ serves as regularization term, and $\alpha_b, \alpha_R, \alpha_W$ are the trading-off parameters which balance between the old knowledge in the model $P^\tau(\boldsymbol{y}|\boldsymbol{x})$ and the information in the new samples. The larger they are, the more old knowledge is preserved in the new model.

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

$$
\begin{aligned}
\frac{\partial L}{\partial b_i} &= \langle y_i \rangle_{\tilde{P}} - \langle y_i \rangle_{P^{\tau+1}} - \frac{\alpha_{\boldsymbol{b}}}{n} b_i \\
\frac{\partial L}{\partial R_{ij}} &= \langle y_i y_j \rangle_{\tilde{P}} - \langle y_i y \rangle_{P^{\tau+1}} - \frac{\alpha_{\boldsymbol{R}}}{n} R_{ij} \\
\frac{\partial L}{\partial W_{il}} &= \langle y_i x_l \rangle_{\tilde{P}} - \langle y_i x_l \rangle_{P^{\tau+1}} - \frac{\alpha_{\boldsymbol{W}}}{n} W_{il}
\end{aligned}
\tag{21}
$$

Given the above derivatives, we can use the efficient gradient descent methods (such as L-BFGS [19]) to maximize (21).

Note that we do not assume any specific probabilistic form of the old model $P^\tau(\boldsymbol{y}|\boldsymbol{x})$, so any statistical model can be used as $P^\tau(\boldsymbol{y}|\boldsymbol{x})$. However, for simplicity, we can assume $P^\tau(\boldsymbol{y}|\boldsymbol{x})$ has the following form

$$
P^\tau(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{1}{Z^\tau(\boldsymbol{x})} \exp\{\boldsymbol{y}^T(\boldsymbol{b}^\tau + \boldsymbol{R}^\tau \boldsymbol{y} + \boldsymbol{W}^\tau \boldsymbol{x})\}
\tag{22}
$$

where $Z^\tau(\boldsymbol{x}) = \sum_y \exp\{\boldsymbol{y}^T(\boldsymbol{b}^\tau + \boldsymbol{R}^\tau \boldsymbol{y} + \boldsymbol{W}^\tau \boldsymbol{x})\}$ is the partition function. Consequently, according to Eqn. (18), the new model $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ is

$$
\begin{aligned}
&P^{\tau+1}(y \mid x) \\
&= \frac{1}{Z^{\tau+1}(x)} \exp\{\boldsymbol{y}^T((\boldsymbol{b}^\tau + \boldsymbol{b}^*) + (\boldsymbol{R}^\tau + \boldsymbol{R}^*)\boldsymbol{y} + (\boldsymbol{W}^\tau + \boldsymbol{W}^*)\boldsymbol{x})\}
\end{aligned}
\tag{23}
$$

We can find $P^{\tau+1}(\boldsymbol{y}|\boldsymbol{x})$ has the same probabilistic form like $P^\tau(\boldsymbol{y}|\boldsymbol{x})$. Therefore, such an online adaption can then be iterated in the same manner in each active learning iteration.

As mentioned in Section 3, to handle the problem of label incompletion caused by the sample-label pair selection strategy in the multi-label active learning algorithm, an EM-algorithm is applied where the missing labels for a certain sample are taken as hidden variables. This algorithm is very similar to the EM procedure adopted in [14].

5. NEW LABEL LEARNING

The online learner presented in Section 4 can be naturally extended to handling new labels. Suppose we have *m* original labels, then parameters $\boldsymbol{b}^\tau$ of the current classifier is an *m*-dimensional vector, $\boldsymbol{R}^\tau$ is an $m \times m$ matrix, and $\boldsymbol{W}^\tau$ is an $m \times d$ matrix. With the introduction of a new label, the parameter $\boldsymbol{b}^\tau$, $\boldsymbol{R}^\tau$ and $\boldsymbol{W}^\tau$ will be changed to $(m+1)$-dimension vector, $(m+1)\times(m+1)$ matrix and $(m+1)\times d$ matrix, respectively, in which all the newly introduced entries of the vector/matrixes are set to zero. Then we use Eqn. (23) to move to the next data batch.

After introducing a new label as above, in the next active learning iteration with a new data batch, according to the last row of Eqn. (9), there is actually no any information about the label at initial stage. That is to say, the marginal distribution $P(y_{m+1}|\boldsymbol{x})$ for the new $(m+1)$-*th* label is uniformly distributed. Therefore, the sample selection strategy tends to select the new label with nearly random samples at initial steps. If prior knowledge about the correlations between the new label and existing labels is introduced, the selected sample-label pairs will be more efficient. This is one of our future work items.

A disadvantage of this solution is that the performance of the new label at the beginning most likely will not be good enough. The reason is the corresponding new labels have very limited number of training data. A solution to this problem is to add a separate single label active learning or initial training set construction technique (such as the approach in [21]) to get a certain number of samples labeled in terms of this new label. And then, we continue the regular online active multi-label learning process.

6. EXPERIMENTS

As TRECVID video data is the only publicly-available medium-scale benchmark dataset, and we also use it as the primary dataset to test the proposed algorithm. Though the scale of this dataset is still limited compared with the scale of the data on the Internet, we will show the potential of the proposed scheme in saving labeling cost and computation cost. Another reason that we choose TRECVID dataset, as well as another even smaller benchmark dataset for some experiments, is that the offline learning algorithms (which we will compare the online learn with) requires very high computation power – which is intractable with limited computation resource within limited time.

We will first conduct experiments on a small benchmark dataset to compare the performances and computation costs between online and offline learning algorithms. We will find the online learner can retain a comparable performance compared with the offline learner while its efficiency is much more competitive than offline learner.

6.1 Online Learner and Offline Learner

We will first compare the labeling accuracy and the computation cost between the proposed online learner in Section 4 and the traditional offline learner previously proposed in [4]. These two algorithms are both applied to multi-label classification problem by exploiting the correlations between the different labels. The experiments are conducted on a multi-labeled *Scene* dataset previously used in [11]. This dataset contains 2,407 natural images belonging to one or more of six natural scene categories including *beach*, *sunset*, *fall foliage*, *field*, *mountain*, and *urban*. Each sample in this set can be assigned one or more than one label at the same time. For the features used in this experiment, 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 49×2×3 = 294 dimension feature vector [11].

In the experiments, both algorithms are initially given 300 samples as training set and 100 training samples will be added sequentially at each round. This round will be iterated 10 times and we compare their computation costs and the labeling accuracies on the remaining samples, as shown in Table 1. Here we use F1 score to compare the classification accuracies. In statistics, the F1 score measures a test's accuracy and has been widely used in information retrieval. It can be computed as F1=2*rp*/(*r*+*p*) where *p* and *r* are precision and recall respectively. We can see that the online learner has a comparable performance with the offline learner. At the same time, we can find in Figure 3 that the computation cost of online learner is much lower than the offline learner. With the same number of new samples in each round, the computation cost of online learner stays stably while that of the offline learner grows rapidly when the increase of the total number of the training samples.

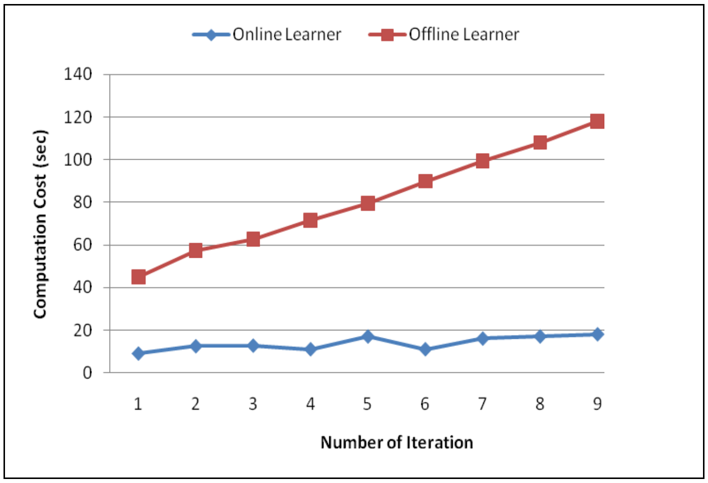


Figure 3: Comparision of computation costs between online and offline learners in each round.

6.2 Online MLAL on TRECVID Dataset

In this section, we evaluate the proposed online multi-label active learning algorithm on the widely-used TRECVID dataset. This version of TRECVID set we used is the development set used in TRECVID 2006 competition [26]. The reason we use it as the evaluation dataset is it contains the complete annotation on all the 39 concepts while the other versions (such as TRECVID 2006 test set) only have been partially annotated. With the fully-annotated labeling, we can compare the performances between different algorithms. This set contains 61,901 shots segmented from collections of international broadcasting news in Arabic, English and Chinese. These video shots are processed to extract block-wise color moment features from order 1 to order 3 in Lab color space. Many previous research efforts support that these color moment features have the best performance over the other features on TRECVID dataset [23].

Table 1: Comparison of F1 Score between online and offline learners on Scene dataset.

# of labeled samples in total	Online Learner	Offline Learner
300	**0.4059**	0.4049
400	**0.4098**	0.4095
500	0.4642	**0.4649**
600	0.4534	**0.4539**
700	0.4850	**0.4854**
800	0.4973	**0.4975**
900	0.5246	**0.5251**
1000	**0.5556**	0.5546
1100	0.5786	**0.5787**
1200	0.5612	**0.5615**

6.2.1 Experiment Setup

To test the proposed active learning schemes in Section 2 for content-based video search, we conduct two experiments. The first experiment is designed to evaluate the multi-label active learning algorithm when new sample-label pairs are selected by the two-dimensional active learning strategy introduced in Section 3. We compare such the multi-label active learning strategy with the previously proposed single-label strategy in [10] which requests to annotate all the labels of the selected samples. We denote the multi-label strategy by MLAL and the single-label strategy by SLAL, respectively. We also conduct experiment on using randomly select sample-label pairs, denoted by RND, with the proposed online learner as underlying classifier. On the other hand, we also conduct the second experiment when new labels are involved into the online learning system. We will see how the performance is improved on these new labels in the online active learning system.

In both experiments, an initial training set is constructed by the first 10,000 samples with all their 39 complete labels. Based on this initial training set, a classifier P^0 is trained. Then at each round, MLAL selects and annotates 39,000 sample-label pairs and then incorporated into the training set to train a new classifier. In contrast, SLAL also selects 39,000 sample-label pairs but these pairs come from the 1000 samples together with all their 39 labels. That is to say, SLAL selects and annotates 1000 samples with all labels. Similarly, these samples and their annotation are also incorporated into the training set to train the new classifier. Such a round will be repeated step-by-step.

For the active learning system with new labels, we do the same experiments but only use 37 labels at the first step, where two concepts, “Building” and “Explosion_Fire”, are excluded. That is, the initial training set will consist of 10,000 samples with only 37 concepts. These two excluded concepts will be involved into the active learning system at the second round. Note that “Building” is strongly correlated with many concepts in the initial training set, such as “Urban”, “Road”, “Outdoor” and “Court”. Similarly, “Explosion_Fire” is also related to “Military”, “Natural-Disaster” and “Car”. We expect the two concepts can be improved by their correlation with these related concepts. These correlations are learned in a data-driven manner compared to prior knowledge from WordNet or Google Distance [37].

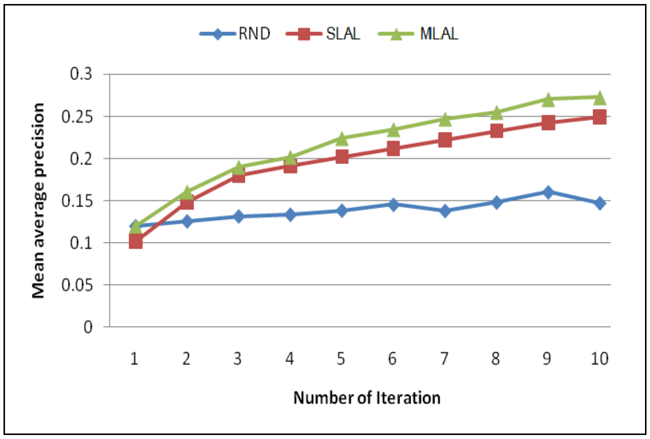
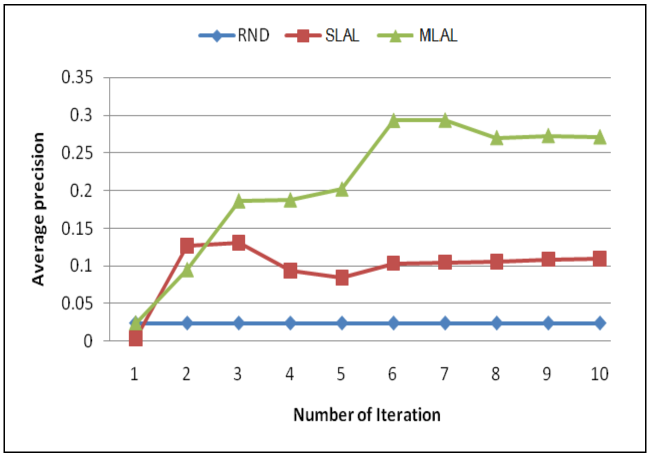
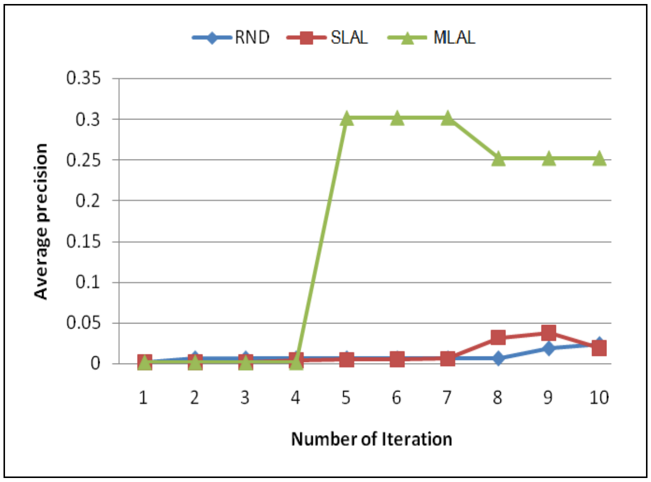


Figure 4: Performance comparison of MLAL, SLAL and RND in terms of mean average precision.



(a) People-Marching



(b) Maps

Figure 5: Performance comparison of MLAL, SLAL and RND on two concepts: (a) People-Marching; (b) Maps.

6.2.2 Experiment I – Active Learning with Online Learner

In this subsection, we report the results of the first experiment which compares different sample-label pair selection strategies, that is, MLAL, SLAL and RND. Figure 4 illustrates the overall performance on all the 39 concepts in terms of mean average precision (MAP). Among all the 10 iterations, MLAL has the best performance compared with SLAL and RND. We also compare the performances of MLAL with online and offline learner as the underlying classifiers in Table 2. We can find the online learner also obtains a competitive results compared to the offline learner. Moreover, on some concepts, MLAL has insignificantly improvement compared to SLAL. For example, Figure 5 illustrates the comparison between MLAL and SLAL on two concepts “People-Marching” and “Maps”. From this illustration, we can find MLAL outperforms SLAL significantly on these two concepts. This is probably due to that the label correlations of these labels with other labels are informative. For example, “People-Marching” can be well induced by its correlation with “Walking_Running”, “Person”, “Crowd”, “Face” and so on. The same reason is applied to “Maps” which is correlated to “Charts”, “Computer_TV-Screen”, etc.

Another advantage of the proposed MLAL algorithm is its efficiency. SLAL uses the SVM as its underlying classifier. So in each round, the SVM must be retrained on the whole training set. For example, in the first iteration, it takes about half and an hour to train the SVM on 10,000 samples. When more samples are involved into the training set, we must spend more and more time on retraining the SVM on the quickly dilated training set. In our experiment, we spend nearly 6 hours to retrain the SVM in the last round. In contrast, we only need invest once to train a starting online model for MLAL on the initial training set. In this experiment, it takes about 3.5 hours to obtain the initial online model. But after that, in each round, this model can be efficiently updated in a few minutes at each round, and this time cost does not increase with the increase of the size of the overall training set. The saving of computation cost is huge especially when more and more samples are labeled.

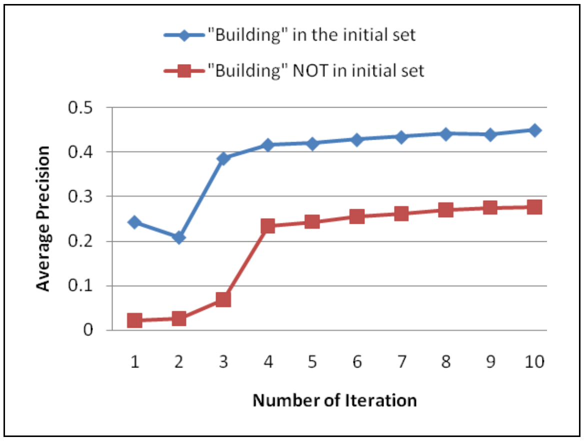
Table 2: Comparison of MAP between online and offline learners on TRECVID dataset.

Iteration Number	Online Learner	Offline Learner
1	**0.12038**	0.11327
2	0.16071	**0.16159**
3	**0.19009**	0.18918
4	0.20179	**0.20894**
5	**0.22422**	0.22364
6	**0.23469**	0.22852
7	0.24708	**0.24795**
8	**0.25502**	0.25370
9	**0.27041**	0.26346
10	**0.27259**	0.26314

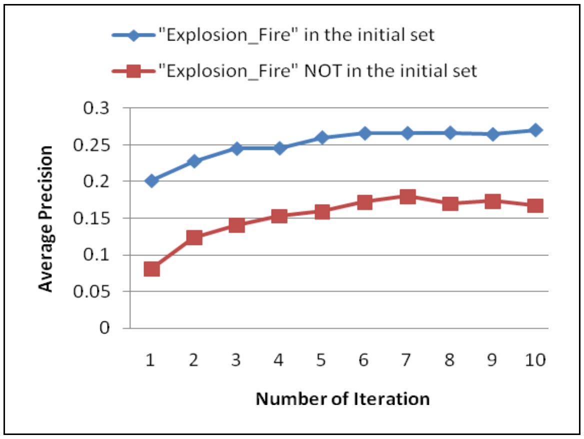
6.2.3 Experiment II – Actively Learning New Labels

We report the experimental results of the second experiment when new labels “Building” and “Explosion_Fire” are involved into the active learning system. Figure 6 shows the promising results of learning new labels.

Although these two concepts start with relatively low accuracy, their performance can be gradually improved once they are incorporated into the system. Note that for the new labels, when they are involved into the system, there is nearly no knowledge of them in the system. So their corresponding marginal distributions of these labels are uniformly distributed. According to the last row of Eqn. (9), the self entropy for these labels are relatively high and it makes the active learning system tends to select the sample-label pairs associated with these labels for annotation. This will be helpful to reduce the uncertainty of these labels and improve their accuracies. However, without prior knowledge, such a process is gradual in a data-driven manner. A data-driven manner means the label correlation is exploited from training samples rather than directly from some prior knowledge. As it does not involve expensive expert’s effort, the accuracy of these labels is relatively lower compared with the performance when these labels exist in the initial training set. In our next plan, we will study how to use some existing knowledge to accelerate the learning rate when new labels are involved, such as through WordNet and Google Distance, or our recent effort called Flickr Distance [39].



(a) Building



(b) Explosion_Fire

Figure 6: The performance of learning two labels (compared with the case when these labels are labeled in the initial training set).

7. CONCLUSION AND FUTURE WORK

We have proposed an online multi-label active learning algorithm for semantic video annotation, which is scalable to the size of both video dataset and semantic concept set. With the support of this algorithm, a new system design of active annotation-based video search system is proposed, which is able to deal with large-scale video data and large-scale concept set. In this design, large-scale unlabeled video samples are assumed to arrive consecutively in batches with an initial pre-labeled training set, based on which an initial multi-label classifier is built. For each arrived batch, an online multi-label active learning engine is applied to efficiently update the classifier which maximizes its performance on all current-available data. This process repeats until all data are arrived, during which new concept labels are allowed to be introduced into the online multi-label active learning framework at any batch, even though these labels have no any pre-labeled training samples. This new framework significantly saved both labeling cost and computation cost but keeps comparable or even better performance compared with the state-of-the-art approaches. Moreover, new label learning without pre-labeled sample demonstrated very promising results

There are still a number of problems need to be addressed under this framework. As aforementioned, the design of the incentive programs for attracting grassroots users to participate active labeling is an interesting problem. And the quality evaluation and insurance of manual labeling, as well as anti-spam, are also essential for the entire framework. Enabling region-level annotation in this active learning process will increase the relevance of video search results, and multi-label multi-instance learning [22] is a possible approach that can be integrated in. Investigating the effects of the Gaussian noises in the online learner, speeding up new label learning, and testing the framework on larger and real world video datasets are also desired. In addition, applying it on image search, text search, and other multi-label classification problems is also our future work.

8. REFERENCES

[1] L. Ahn, and L. Dabbish, “Labeling Images with a Computer Game,” in Proc. of ACM CHI, 2004.

[2] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrel, “Active Learning with Gaussian Processes for Object Recognition,” in *Proc. of IEEE International Conference on 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,” in *Proc. of International ACM Conference on Multimedia*, 2007.

[4] S. Zhu, X. Ji, W. Xu, and Y. Gong, “Multi-Labeled Classification using Maximum Entropy Method,” in *Proc. of ACM SIGIR*, 2005.

[5] S. C. H. Hoi and M. R. Lyu, “*A Semi-Supervised Active Learning Framework for Image Retrieval*,” in Proc. of IEEE *Conference on Computer Vision and Patter Recognition*, 2005.

[6] A. Dong and B. Bhanu, “Active Learning for Image Retrieval in Dynamic Databases,” in *Proc. of IEEE International Conference on Computer Vision,*, 2003.

[7] R. Yan, J. Yang, and A. Hauptmann, “Automatic Labeling Data using Multi-Class Active Learning, *Proc. of IEEE International Conference on Computer Vision*, 2003.

[8] S. Tong, and Edward Chang, “Support Vector Machine Active Learning for Image Retrieval,” in Proc. of *International ACM Conference on Multimedia*, 2001.

[9] E. Chang, S. Tong, K. Goh, and C. Chang, “Support Vector Machine Concept-Dependent Active Learning for Image Retrieval,” *IEEE Transactions on Multimedia*, 2005.

[10] X. Li, L. Wang, and E. Sung, "Multi-Label SVM Active Learning for Image Classification," in *Proc. of IEEE International Conference on Image Processing*, 2004.

[11] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning Multi-Label Scene Classification," *Pattern Recognition*, 2004.

[12] K. Brinker, "On active learning in multi-label classification," in *Book "From Data and Information Analysis to Knowledge Engineering" of Book Series "Studies in Classification, Data Analysis, and Knowledge Organization"*, Springer, 2006.

[13] T. Cover and J. Thomas, *Elements of Information Theory*, Second Edition, New York: Wiley Series in Telecommunications, John Wiley and Sons, 2006.

[14] G.-J. Qi, X.-S. Hua, et al., "Two-Dimensional Active Learning for Image Classification," in *Proc. of IEEE Conference on Computer Vision and Patter Recognition*, 2008.

[15] S. F. Chen and R. Rosenfeld, "A Gaussian Prior for Smoothing Maximum Entropy Models," School of Computer Science, Carnegie Mellon University, Tech. Rep. CMU-CS-99-108, 1999.

[16] N. Syed, H. Liu, and K. Sung, "Incremental Learning with Support Vector Machines," in *Workshop on Support Vector Machines, at the IJCAI, 1999*.

[17] G. Cauwenberghs and T. Poggio, "Incremental and Decremental Support Vector Machine," in *Proc. of NIPS,* 2000.

[18] J. Yang, R. Yan, and A. Hauptmann, "Cross-Domain Video Concept Detection using Adaptive SVMs," in *Proc. of International ACM Conference on Multimedia,* 2007.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-Likelihood from Incomplete Data via EM Algorithm," *Journal of the Royal Statistical Society (Series B)*, 1977.

[20] W. Jiang, S.-F. Chang, and A. Loui, "Active Concept-Based Concept Fusion with Partial User Labels," in *Proc. of IEEE International Conference on Image Processing*, 2006.

[21] J. Tang, Y. Song, X.-S. Hua, T. Mei, X. Wu, "To Construct Optimal Training Set For Video Annotation," in *Proc. of International ACM Conference on Multimedia*, 2006.

[22] Z.-J. Zha, X.-S. Hua, et al., "Joint Multi-Label Multi-Instance Learning for Image Classification," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[23] X.-S. Hua, T. Mei, W. Lai, M. Wang, J. Tang, G.-J. Qi, L. Li, Z. Gu, "Microsoft Research Asia TRECVID 2006: High-Level Feature Extraction and Rushes Exploitation," *In TREC Video Retrieval Evaluation Online Proceeding*, 2006.

[24] Luis von Ahn, Manuel Blum and John Langford. Telling Humans and Computers Apart Automatically. In Communications of the ACM.

[25] reCAPTCHA. http://recaptcha.net/.

[26] Smeaton, A. F., Over, P., and Kraaij, W. Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval* (MIR). 2006.

[27] C. Ngo, Y. Jiang, X. Wei, F. Wang, W. Zhao, H. Tan and X. Wu. Experimenting VIREO-374: Bag-of-Visual-Words and Visual-Based Ontology for Semantic Video Indexing and search. *In TREC Video Retrieval Evaluation Online Proceeding*, 2007.

[28] S. Chang, W. Jiang, A. Yanagawa, and E. Zavesky. Columbia University TRECVID 2007 High-Level Feature Extraction. *In TREC Video Retrieval Evaluation Online Proceeding*, 2007.

[29] M. Campbell, et al. IBM Research TRECVID-2007 Video Retrieval System. *In TREC Video Retrieval Evaluation Online Proceeding*, 2007.

[30] C.G.M. Snoek, et al. The MediaMill TRECVID 2007 Semantic Video Search Engine. *In TREC Video Retrieval Evaluation Online Proceeding*, 2007.

[31] J. Yuan, et al. THU and ICRC at TRECVID 2007. *In TREC Video Retrieval Evaluation Online Proceeding*, 2007.

[32] M. Naphade, J. R. Smith, et al. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86 – 91, 2006.

[33] S. Ayache and G. Qu´enot. Evaluation of active learning strategies for video indexing. Signal Processing: *Image Communication*, 2007.

[34] Stéphane Ayache.Georges Quénot. TRECVID 2007: Collaborative Annotation using Active Learning. *In TREC Video Retrieval Evaluation Online Proceeding*, 2007.

[35] Q. Zhang, et al. The COST292 experimental framework for TRECVID 2007. *In TREC Video Retrieval Evaluation Online Proceeding, 2007*.

[36] G.-J. Qi, X.-S. Hua, Y. Rui and H.-J. Zhang. Two-Dimensional Multi-Label Active Learning with An Efficient Online Adaption Model for Image Classification, *Pre-prints of submission of IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[37] R. L. Cilibrasi, P. M. B. Vitanyi. The Google Similarity Distance, *in IEEE Trans. Knowledge and Data Engineering, 19:3(2007), 370-383.*

[38] Alexander Sorokin, David Forsyth. Utility data annotation with Amazon Mechanical Turk. First International Workshop on Internet Vision (in conjunction with CVPR), 2008.

[39] Lei Wu, Xian-Sheng Hua, et al. Flickr Distance. ACM Multimedia 2008.

[40] Yang Yang, Bin B. Zhu, et al. A Comprehensive Human Computation Framework – With Application to Image Labeling. ACM Multimedia 2008.