Semantic-Gap-Oriented Active Learning for Multilabel Image Annotation

Jinhui Tang, Zheng-Jun Zha, Dacheng Tao, and Tat-Seng Chua

***Abstract*—User interaction is an effective way to handle the semantic gap problem in image annotation. To minimize user effort in the interactions, many active learning methods were proposed. These methods treat the semantic concepts individually or correlatively. However, they still neglect the key motivation of user feedback: to tackle the semantic gap. The size of the semantic gap of each concept is an important factor that affects the performance of user feedback. User should pay more efforts to the concepts with large semantic gaps, and vice versa. In this paper, we propose a semantic-gap-oriented active learning method, which incorporates the semantic gap measure into the information-minimization-based sample selection strategy. The basic learning model used in the active learning framework is an extended multilabel version of the sparse-graph-based semisupervised learning method that incorporates the semantic correlation. Extensive experiments conducted on two benchmark image data sets demonstrated the importance of bringing the semantic gap measure into the active learning process.**

***Index Terms*—Active learning, image annotation, multilabel, semantic gap, sparse graph.**

I. Introduction

In recent years, a lot of research work has been devoted to automatic image annotation [1]–[3]. However, the purely automatic image annotation techniques are still far from satisfactory due to the well-known problem of semantic gap. User interactions and feedback provide a possible solution to handle this issue. To fully utilize human effort, active learning aims to actively select the most effective samples to present to the users for feedback [4].

A typical active learning framework consists of two parts, i.e., a learning engine and a sample selection engine. It is an iterative process. In each round, the learning engine trains a model to predict the labels of unlabeled samples based on the training set, whereas the sample selection engine selects the most effective unlabeled samples based on a certain strategy for manual labeling. These samples are then added to the training set for the next round of learning. The basic objective of the sample selection engine is to select the samples that are more useful than the randomly selected samples for training.

It is obvious that the sample selection strategy plays a crucial role in an active learning framework. Lots of approaches were proposed to reduce the number of manually labeled images required for effective learning [5]. However, the existing methods do not directly tackle the semantic gap, which is a key motivation of user feedback. The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation [6]. Traditionally, an image is relevant to multiple semantic concepts while different concepts have different sizes of semantic gaps [7]. The size of each concept's semantic gap is an important factor that affects the user interactions in a multilabel image annotation framework. User should spend more efforts to the concepts with large semantic gaps, and vice versa.

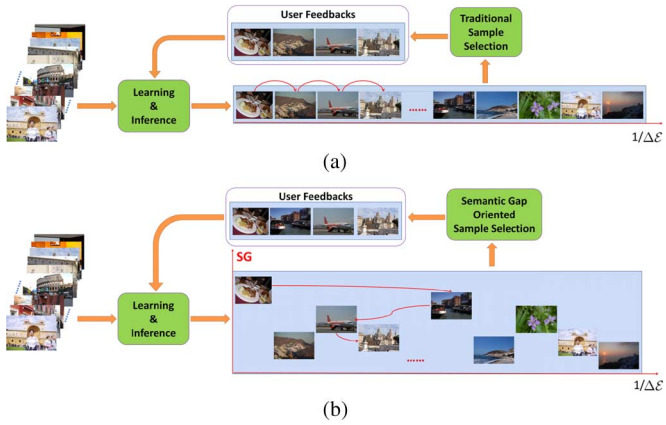

Fig. 1. Comparison of the frameworks of semantic-gap-oriented active learning and traditional active learning. (a) Framework of traditional active learning. (b) Framework of semantic-gap-oriented active learning.

In this paper, we propose a semantic-gap-oriented active learning method for multilabel image annotation. It combines the semantic gap measures of concepts into the information minimization criterion to take into the effects of semantic gap in the sample selection strategy. Fig. 1 visualizes the comparison between the frameworks of semantic-gap-oriented active learning and traditional active learning. The basic learning model used in this framework is the multilabel sparse-graph-based semisupervised learning, which is an extended version of the method proposed by Tang *et al.* [8] to a multilabel scenario by incorporating the semantic correlation. Extensive experiments are conducted on the benchmark image data sets NUS-WIDE [9] and Corel [10] to show the effectiveness of integrating the semantic gap measure into an active learning framework.

The rest of this paper is organized as follows. Section II introduces the related work on sample selection. In Section III, we derive the semantic-gap-oriented active learning strategy based on the expected classification risk reduction. The quantitative measure of semantic gap is presented in Section IV, whereas the correlative sparse-graph-based semisupervised learning framework is introduced in Section V. Section VI details the experimental evaluation on a real-world image set. Finally, we conclude the work in Section VII.

II. Related Work

The optimal sample selection strategy is based on the expected classification risk reduction [11]. However, estimating the reduced expected risk is very computationally intensive [12]. For image annotation, we usually need to deal with the large-scale data. Thus, several kinds of heuristic sample selection strategies were proposed. As discussed by Wang and Hua [5], these heuristic criteria can be categorized into four groups.

- The most commonly used criterion is *uncertainty*, which means that the most uncertain samples should be selected. A typical mea-

Manuscript received May 04, 2011; revised September 12, 2011 and December 09, 2011; accepted December 09, 2011. Date of publication December 21, 2011; date of current version March 21, 2012. This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant 61103059, in part by the National Basic Research Program of China (973 Program) under Grant 2012CB316304, in part by the NSFC under Grant 61173104, in part by the Natural Science Foundation of Jiangsu Province under Grant BK2011700, and in part by by the Australian Research Council discovery project (ARC DP-120103730). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Min Wu.

J. Tang is with the School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: jinhuitang@mail.njust.edu.cn).

Z.-J. Zha and T.-S. Chua are with the School of Computing, National University of Singapore, Singapore 117417 (e-mail: zhazj@comp.nus.edu.sg; chuats@comp.nus.edu.sg).

D. Tao is with the Center for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology, Sydney, N.S.W. 2007, Australia (e-mail: dacheng.tao@uts.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2011.2180916

sure that estimates uncertainty is the information entropy. The uncertainty criterion can be also viewed as a greedy strategy to reduce risk (without model updating, the method to reduce maximal expected risk is to select the most uncertain samples). This criterion has been widely explored for its simplicity [13]–[16].

- The second criterion is *diversity*, which was first investigated in batch mode active learning [17]. In many applications, we need to select a batch of samples instead of just one in an active learning iteration. Recent work shows that the selected samples in a batch or even all the labeled samples should be diverse [15], [16], [18].
- The third criterion is *density*, which favors the selection of samples within the regions of high density [19], [20].
- The last criterion is *relevance*, which is the degree of the retrieved image relevant to the query. It is usually applied in multilabel image annotation and retrieval. Of course, the aforementioned criteria such as *uncertainty* can be also applied together with this criterion. However, in many cases, it is found that the use of relevance criterion alone, i.e., directly selecting the samples that have the highest probabilities to be relevant, is more effective [21], [22].

The four criteria reflect four different aspects of samples' effectiveness. In many cases, these criteria are combined explicitly or implicitly [23]. The *semantic-gap*-based strategy can be regarded as a supplement to the traditional four criteria while it can be also combined with other strategies.

Actually, the above four criteria are general sample selection strategies for active learning. For the complex case of multilabel image annotation, several special strategies were proposed. Qi *et al.* [24] proposed a 2-D active learning method that selects sample-concept pairs for manual annotation in a correlative multilabel learning approach. The sample-concept selection is derived based on the reduction in multilabel Bayesian error bound. Vijayanarasimhan and Grauman [25] treated the manual labeling costs to be different for different images and label types. They proposed a method that trades off between the labeling effort and the information gain.

The objective of our work is to demonstrate the usefulness of the semantic gap measure in the sample selection strategy. Thus, we derive the semantic-gap-oriented active learning approach basically from the information minimization criterion, while discarding the other criteria. Certainly, the other criteria such as diversity and relevance can be also integrated into the whole framework.

III. Semantic-Gap-Oriented Active Learning

We use $\mathbf{f}$ to denote the predicted multilabel vector of image $\mathbf{x}$, where f_i gives the predicted label of the ith semantic concept of $\mathbf{x}$. By using $\mathcal{U}$ to denote the pool of unlabeled samples to be selected for learner, the Bayesian classification error over all samples in $\mathcal{U}$ before labeling a selected sample $\mathbf{x}_s$ is

$$\mathcal{E}^b(\mathcal{U}) = \frac{1}{|\mathcal{U}|}\sum_{\mathbf{x}\in\mathcal{U}} \mathcal{E}(\mathbf{f}|\mathcal{L}^b,\mathbf{x}) \tag{1}$$

where $\mathcal{L}^b = (\mathbf{x}_1,\mathbf{y}_1),(\mathbf{x}_2,\mathbf{y}_2),\ldots,(\mathbf{x}_{|\mathcal{L}^b|},\mathbf{y}_{|\mathcal{L}^b|})$ represents the set of labeled image-label pairs. After labeling, the expected classification error is

$$\begin{aligned}\mathcal{E}^a(\mathcal{U}) &= \frac{1}{|\mathcal{U}|}\left\{\mathcal{E}(\mathbf{f}|\mathbf{f}_s;\mathcal{L}^b,\mathbf{x}_s) + \sum_{\mathbf{x}\in\mathcal{U}\setminus\mathbf{x}_s}\mathcal{E}(\mathbf{f}|\mathcal{L}^a,\mathbf{x})\right\}\\ &= \frac{1}{|\mathcal{U}|}\sum_{\mathbf{x}\in\mathcal{U}\setminus\mathbf{x}_s}\mathcal{E}(\mathbf{f}|\mathcal{L}^a,\mathbf{x})\end{aligned} \tag{2}$$

where $\mathcal{L}^a = \{\mathcal{L}^b,(\mathbf{x}_s,\mathbf{y}_s)\}$.

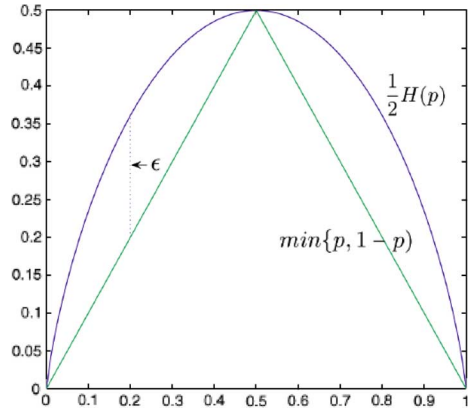

Fig. 2. Illustration of the inequality $(1/2)H(p) - \epsilon \leqslant \min\{p, 1-p\} \leqslant (1/2)H(p), \epsilon = (1/2)\log(5/4)$ [24].

Then, the error reduction can be represented as

$$\begin{aligned}\Delta\mathcal{E}(\mathcal{U}) =& \mathcal{E}^b(\mathcal{U}) - \mathcal{E}^a(\mathcal{U})\\ =& \frac{1}{|\mathcal{U}|}\left\{\mathcal{E}(\mathbf{f}|\mathcal{L}^b,\mathbf{x}_s) + \sum_{\mathbf{x}\in\mathcal{U}\setminus\mathbf{x}_s}\mathcal{E}(\mathbf{f}|\mathcal{L}^b,\mathbf{x}) - \sum_{\mathbf{x}\in\mathcal{U}\setminus\mathbf{x}_s}\mathcal{E}(\mathbf{f}|\mathcal{L}^a,\mathbf{x})\right\}\\ =& \Delta\mathcal{E}_1 + \Delta\mathcal{E}_2\end{aligned} \tag{3}$$

where $\Delta\mathcal{E}_1 = (1/|\mathcal{U}|)\mathcal{E}(\mathbf{f}|\mathcal{L}^b,\mathbf{x}_s)$ and $\Delta\mathcal{E}_2 = (1/|\mathcal{U}|)\{\sum_{\mathbf{x}\in\mathcal{U}\setminus\mathbf{x}_s}\mathcal{E}(\mathbf{f}|\mathcal{L}^b,\mathbf{x}) - \sum_{\mathbf{x}\in\mathcal{U}\setminus\mathbf{x}_s}\mathcal{E}(\mathbf{f}|\mathcal{L}^a,\mathbf{x})\}$. Thus, our objective is to select the $\mathbf{x}_s$ that maximize the above expected error reduction $\Delta\mathcal{E}_1 + \Delta\mathcal{E}_2$.

Let us first consider $\Delta\mathcal{E}_1$

$$\begin{aligned}\Delta\mathcal{E}_1 = \frac{1}{|\mathcal{U}|}\sum_{i=1}^{m}\mathcal{E}(f_i|\mathcal{L}^b,\mathbf{x}_s) &\leqslant \frac{1}{2|\mathcal{U}|}\sum_{i=1}^{m}H(f_i|\mathcal{L}^b,\mathbf{x}_s)\\ &= \frac{1}{2|\mathcal{U}|}H(\mathbf{f}|\mathcal{L}^b,\mathbf{x}_s)\end{aligned} \tag{4}$$

where

$$H(\mathbf{f}|\mathcal{L}^b,\mathbf{x}_s) = \sum_{i=1}^{m}H(f_i|\mathcal{L}^b,\mathbf{x}_s) \tag{5}$$

and $H(f_i|\mathcal{L}^b,\mathbf{x}_s)$ is defined as

$$\begin{aligned}&H(f_i|\mathcal{L}^b,\mathbf{x}_s)\\ &\quad= -P(f_i=1|\mathcal{L}^b,\mathbf{x}_s)\times\log P(f_i=1|\mathcal{L}^b,\mathbf{x}_s)\\ &\qquad- P(f_i=0|\mathcal{L}^b,\mathbf{x}_s)\times\log P(f_i=0|\mathcal{L}^b,\mathbf{x}_s).\end{aligned} \tag{6}$$

The inequality in (4) can be proven as follows: the Bayesian classification error is [26]: $\mathcal{E}(f_i|\mathcal{L}^b,\mathbf{x}_s) = \min\{P(f_i = 1|\mathcal{L}^b,\mathbf{x}_s), P(f_i = 0|\mathcal{L}^b,\mathbf{x}_s)\}$. From Fig. 2, we can see that the inequality $(1/2)H(p) - \epsilon \leqslant \min\{p,1-p\} \leqslant (1/2)H(p)$ holds and $\epsilon = (1/2)\log(5/4)$. Thus, we have$(1/2)H(f_i|\mathcal{L}^b,\mathbf{x}_s) - \epsilon \leqslant \mathcal{E}(f_i|\mathcal{L}^b,\mathbf{x}_s) \leqslant (1/2)H(f_i|\mathcal{L}^b,\mathbf{x}_s)$.

Regarding $\Delta\mathcal{E}_2$, the expected errors $\mathcal{E}(\mathbf{f}|\mathcal{L}^a,\mathbf{x})$ and $\mathcal{E}(\mathbf{f}|\mathcal{L}^b,\mathbf{x})$ are not computable due to the unknown marginal distribution $P(\mathbf{x})$ and conditional distribution $P(\mathbf{y}|\mathbf{x})$. Fortunately, our learning framework

is based on the sparse graph representation that means every sample is only related to very few other samples. Then, we can rewrite $\Delta\mathcal{E}_2$ as

$$\begin{aligned}\Delta\mathcal{E}_2 &= \frac{1}{|\mathcal{U}|}\left\{\sum_{\mathbf{x}\in\mathcal{U}\setminus\mathbf{x}_s}\mathcal{E}(\mathbf{f}|\mathcal{L}^b,\mathbf{x}) - \sum_{\mathbf{x}\in\mathcal{U}\setminus\mathbf{x}_s}\mathcal{E}(\mathbf{f}|\mathcal{L}^a,\mathbf{x})\right\}\\ &= \frac{1}{|\mathcal{U}|}\left\{\sum_{\mathbf{x}\in\mathcal{R}_s}\mathcal{E}(\mathbf{f}|\mathcal{L}^b,\mathbf{x}) + \sum_{\mathbf{x}\in\overline{\mathcal{R}}_s}\mathcal{E}(\mathbf{f}|\mathcal{L}^b,\mathbf{x})\right.\\ &\qquad\left. - \sum_{\mathbf{x}\in\mathcal{R}_s}\mathcal{E}(\mathbf{f}|\mathcal{L}^a,\mathbf{x}) - \sum_{\mathbf{x}\in\overline{\mathcal{R}}_s}\mathcal{E}(\mathbf{f}|\mathcal{L}^a,\mathbf{x})\right\}\\ &= \frac{1}{|\mathcal{U}|}\sum_{\mathbf{x}\in\mathcal{R}_s}\left\{\mathcal{E}(\mathbf{f}|\mathcal{L}^b,\mathbf{x}) - \mathcal{E}(\mathbf{f}|\mathcal{L}^a,\mathbf{x})\right\} \qquad (7)\end{aligned}$$

where $\mathcal{R}_s$ is the sample set in which each sample is reconstructed by using $\mathbf{x}_s$, $\overline{\mathcal{R}}_s$ is the sample set in which neither sample is reconstructed by using $\mathbf{x}_s$, $\mathcal{R}_s\bigcup\overline{\mathcal{R}}_s = \mathcal{U}\setminus\mathbf{x}_s$, and $\mathcal{R}_s\bigcap\overline{\mathcal{R}}_s = \emptyset$. The third equality stands because the samples in $\overline{\mathcal{R}}_s$ are not related to $\mathbf{x}_s$; thus, $\sum_{\mathbf{x}\in\overline{\mathcal{R}}_s}\mathcal{E}(\mathbf{f}|\mathcal{L}^b,\mathbf{x}) = \sum_{\mathbf{x}\in\overline{\mathcal{R}}_s}\mathcal{E}(\mathbf{f}|\mathcal{L}^a,\mathbf{x})$.

Until now, $\Delta\mathcal{E}_2$ is still not computable. However, we can find that a key factor affecting the quantity is the size of $\mathcal{R}_s$. Assume that the obtained error reduction $\mathcal{E}(\mathbf{f}|\mathcal{L}^b,\mathbf{x}) - \mathcal{E}(\mathbf{f}|\mathcal{L}^a,\mathbf{x})$ for every $\mathbf{x}\in\mathcal{R}_s$ is similar when we add a new training pair $(\mathbf{x}_s,\mathbf{y}_s)$; we can achieve more error reduction linearly when the size of $\mathcal{R}_s$ increases. Thus, we can rewrite (7) as

$$\Delta\mathcal{E}_2 \doteq \frac{1}{2|\mathcal{U}|}\left(\rho|\mathcal{R}_s|\right) \qquad (8)$$

where $|\mathcal{R}_s|$ is the number of samples in $\mathcal{R}_s$, and ρ is an incomputable variable, while we regarded as a parameter.

To this end, we can obtain that

$$\Delta\mathcal{E}(\mathcal{U}) \leqslant \frac{1}{2|\mathcal{U}|}\left\{H(\mathbf{f}|\mathcal{L}^b,\mathbf{x}_s) + \rho|\mathcal{R}_s|\right\}. \qquad (9)$$

A reasonable approach is to select the $\mathbf{x}_s$ that maximizes the above expected error reduction $\Delta\mathcal{E}(\mathcal{U})$, which is difficult to estimate. Fortunately, it is easy to calculate the upper bound of the error reduction. Thus, we can alternatively select the $\mathbf{x}_s$ that maximizes the upper bound of the error reduction. That is

$$\mathbf{x}_s^* = \text{argmax}_{\mathbf{x}_s\in\mathcal{U}}\left\{H(\mathbf{f}|\mathcal{L}^b,\mathbf{x}_s) + \rho|\mathcal{R}_s|\right\}. \qquad (10)$$

However, we know that different concepts have different sizes of semantic gaps [7]. The automatic annotation of some concepts works quite well, such as *sunset*. For this kind of concepts, sometimes it is not necessary to manually label them. For the concepts with large semantic gaps, such as *dog*, the automatic annotation performance is not sufficient to fulfill the needs of real applications. Thus, we need to pay more attention to manually labeling those images that are informative to concepts with large semantic gaps.

We use the significance vector $\Gamma = [\gamma_1,\ldots,\gamma_m]$ to denote the semantic gap measures of the m different concepts. By using $\mathcal{E}_\Gamma(\mathbf{f}|\mathcal{L},\mathbf{x})$ to represent the semantically weighted classification error for $\mathbf{x}$, and note that $|\mathcal{R}_s|$ is unrelated to the semantic gap of concepts, formulas (3) and (4) can be rewritten as

$$\begin{aligned}\Delta\mathcal{E}_\Gamma(\mathcal{U}) &= \mathcal{E}_\Gamma^b(\mathcal{U}) - \mathcal{E}_\Gamma^a(\mathcal{U}) = \Delta\mathcal{E}_{\Gamma 1} + \Delta\mathcal{E}_{\Gamma 2}\\ &= \frac{1}{|\mathcal{U}|}\sum_{i=1}^m\gamma_i\mathcal{E}(f_i|\mathcal{L}^b,\mathbf{x}_s) + \frac{1}{2|\mathcal{U}|}\rho|\mathcal{R}_s|\\ &\leqslant \frac{1}{2|\mathcal{U}|}\left\{\sum_{i=1}^m\gamma_i H(f_i|\mathcal{L}^b,\mathbf{x}_s) + \rho|\mathcal{R}_s|\right\}\\ &= \frac{1}{2|\mathcal{U}|}\left\{H_\Gamma(\mathbf{f}|\mathcal{L}^b,\mathbf{x}_s) + \rho|\mathcal{R}_s|\right\}. \qquad (11)\end{aligned}$$

To this end, our objective becomes to select the $\mathbf{x}_s$ that maximizes the above expected semantically weighted classification error reduction $\Delta\mathcal{E}_\Gamma(\mathcal{U})$. Thus

$$\mathbf{x}_s^* = \text{argmax}_{\mathbf{x}_s\in\mathcal{U}}\left\{H_\Gamma(\mathbf{f}|\mathcal{L}^b,\mathbf{x}_s) + \rho|\mathcal{R}_s|\right\} \qquad (12)$$

where

$$H_\Gamma(\mathbf{f}|\mathcal{L}^b,\mathbf{x}_s) = \sum_{i=1}^m\gamma_i H(f_i|\mathcal{L}^b,\mathbf{x}_s) \qquad (13)$$

is the semantically weighted information of sample $\mathbf{x}_s$'s labels.

Considering that these labels have semantic correlations, we should incorporate the mutual information of the $\mathbf{x}_s$'s labels into formula (13). Here, we only consider the first-order semantic correlation information, $\Delta\mathcal{E}_{\Gamma 1}$ becomes

$$\begin{aligned}\Delta\mathcal{E}_{\Gamma 1} &= \frac{1}{|\mathcal{U}|}\sum_{i=1}^m E\left[\mathcal{E}_\Gamma(f_i|\mathcal{L}^b,\mathbf{x}_s,f_j)\right]\\ &= \frac{1}{(m-1)|\mathcal{U}|}\sum_{i=1}^m\sum_{j\neq i}\mathcal{E}_\Gamma(f_i|\mathcal{L}^b,\mathbf{x}_s,f_j)\\ &= \frac{1}{(m-1)|\mathcal{U}|}\sum_{i=1}^m\sum_{j\neq i}\gamma_i\mathcal{E}(f_i|\mathcal{L}^b,\mathbf{x}_s,f_j)\\ &\leqslant \frac{1}{2(m-1)|\mathcal{U}|}\sum_{i=1}^m\sum_{j\neq i}\gamma_i H(f_i|\mathcal{L}^b,\mathbf{x}_s,f_j)\\ &= \frac{1}{2(m-1)|\mathcal{U}|}\sum_{i=1}^m\sum_{j\neq i}\gamma_i\\ &\qquad\times\left\{H(f_i|\mathcal{L}^b,\mathbf{x}_s) - MI(f_i,f_j|\mathcal{L}^b,\mathbf{x}_s)\right\}\\ &= \frac{1}{2|\mathcal{U}|}\left\{\sum_{i=1}^m\gamma_i H(f_i|\mathcal{L}^b,\mathbf{x}_s)\right.\\ &\qquad\left. - \frac{1}{m-1}\sum_{i=1}^m\sum_{j\neq i}\gamma_i MI(f_i,f_j|\mathcal{L}^b,\mathbf{x}_s)\right\}\\ &= \frac{1}{2|\mathcal{U}|}\left\{\sum_{i=1}^m\gamma_i H(f_i|\mathcal{L}^b,\mathbf{x}_s)\right.\\ &\qquad\left. - \frac{1}{m-1}\sum_{i,j:i\neq j}^m\frac{\gamma_i+\gamma_j}{2}MI(f_i,f_j|\mathcal{L}^b,\mathbf{x}_s)\right\} \qquad (14)\end{aligned}$$

where $E[\cdot]$ is the mathematical expectation, and the last equality comes from the fact that $MI(f_i,f_j|\mathcal{L}^b,\mathbf{x}_s) = MI(f_j,f_i|\mathcal{L}^b,\mathbf{x}_s)$. Finally, we can obtain the correlation-based sample selection strategy

$$\begin{aligned}\mathbf{x}_s^* = \text{argmax}_{\mathbf{x}_s\in\mathcal{U}}&\left\{\sum_{i=1}^m\gamma_i H(f_i|\mathcal{L}^b,\mathbf{x}_s)\right.\\ &\left. - \frac{1}{m-1}\sum_{i,j:i\neq j}^m\frac{\gamma_i+\gamma_j}{2}MI(f_i,f_j|\mathcal{L}^b,\mathbf{x}_s) + \rho|\mathcal{R}_s|\right\}. \qquad (15)\end{aligned}$$

Here, $H(f_i|\mathcal{L}^b,\mathbf{x}_s)$ is calculated according to the formula (6), and $MI(f_i,f_j|\mathcal{L}^b,\mathbf{x}_s)$ is calculated as

$$\begin{aligned}&MI(f_i,f_j|\mathcal{L}^b,\mathbf{x}_s)\\ &\quad= H(f_i|\mathcal{L}^b,\mathbf{x}_s) + H(f_j|\mathcal{L}^b,\mathbf{x}_s) - H(f_i,f_j|\mathcal{L}^b,\mathbf{x}_s) \qquad (16)\end{aligned}$$

where

$$H(f_i, f_j|\mathcal{L}^b, \mathbf{x}_s) = -\sum_{f_i, f_j \in [1,0]} P(f_i, f_j|\mathcal{L}^b, \mathbf{x}_s) \log P(f_i, f_j|\mathcal{L}^b, \mathbf{x}_s). \quad (17)$$

Since we have utilized the label correlation information in the learning model, the joint posterior probability $P(f_i, f_j|\mathcal{L}^b, \mathbf{x}_s)$ can be approximated with a logistic function on the output of the classifier: $P(f_i = 1, f_j = 1|\mathcal{L}^b, \mathbf{x}_s) = 1/(1 + e^{-\beta(\hat{y}_i + \hat{y}_j)})$, $P(f_i = 1, f_j = 0|\mathcal{L}^b, \mathbf{x}_s) = 1/(1 + e^{-\beta(\hat{y}_i - \hat{y}_j)})$, $P(f_i = 0, f_j = 1|\mathcal{L}^b, \mathbf{x}_s) = 1/(1 + e^{-\beta(-\hat{y}_i + \hat{y}_j)})$, and $P(f_i = 0, f_j = 0|\mathcal{L}^b, \mathbf{x}_s) = 1/(1 + e^{-\beta(-\hat{y}_i - \hat{y}_j)})$, followed by a linear normalization to ensure $\sum_{f_i, f_j \in [1,0]} P(f_i, f_j|\mathcal{L}^b, \mathbf{x}_s) = 1$. Here, $\hat{y}_i = y_i - \tau_i$, $\tau_i = (|\mathcal{T}_i|/|\mathcal{L}|)$, where $|\mathcal{T}_i|$ represents the number of the relevant samples for the ith concept in the training set, and $|\mathcal{L}|$ is the number of all samples in the training set, and β is set to 3 empirically. Then, we can marginalize the joint probability to get: $P(f_i = 1|\mathcal{L}^b, \mathbf{x}_s) = P(f_i = 1, f_j = 1|\mathcal{L}^b, \mathbf{x}_s) + P(f_i = 1, f_j = 0|\mathcal{L}^b, \mathbf{x}_s)$, and $P(f_i = 0|\mathcal{L}^b, \mathbf{x}_s) = P(f_i = 0, f_j = 1|\mathcal{L}^b, \mathbf{x}_s) + P(f_i = 0, f_j = 0|\mathcal{L}^b, \mathbf{x}_s)$.

IV. Quantitative Measure of Semantic Gap

Semantic gap can be regarded as the inconsistency between the distributions of the low-level visual features and the high-level semantic concepts. Currently few research efforts have been done on how to quantitatively measure the semantic gaps of different concepts, except for the recent work done be Lu *et al.* [7]. We adopt a similar method as that of [7] in our framework to measure the semantic gap. Here, we briefly introduce the procedure.

First, for each image $\mathbf{x}_i$, we search for its k nearest neighbors based on their visual similarities. Meanwhile, the semantic distance $\mathrm{dis\_sem}(\mathbf{x}_i, \mathbf{x}_j)$ between $\mathbf{x}_i$ and each of its neighbors $\mathbf{x}_j$ is measured by the cosine distance between the vectors of their tags: $1 - (\mathbf{y}_i \cdot \mathbf{y}_j / |\mathbf{y}_i||\mathbf{y}_j|)$. The semantic gap of the current image $\mathbf{x}_i$ is then measured as

$$\mathrm{Im\_SG}(\mathbf{x}_i) = \frac{1}{k} \sum_{\mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)} \mathrm{dis\_sim}(\mathbf{x}_i, \mathbf{x}_j) \quad (18)$$

where $\mathcal{N}(\mathbf{x}_i)$ represents the set of the k nearest neighbors of $\mathbf{x}_i$ in the visual space. We can see that $\mathrm{Im\_SG}(\mathbf{x}_i)$ interprets the consistency of visual features and semantic concepts of image $\mathbf{x}_i$.

Second, the images with the t-smallest semantic gaps are selected, and the fast K-Means algorithm [27] is used to cluster these images into K clusters C_i $(i = 1, \ldots, K)$. Here, instead of employing the affinity propagation as in [7], we adopt the fast K-Means algorithm since the affinity graph construction is very slow.

Finally, the semantic gap measure of a certain concept c_j is defined based on the relevance scores of the tags to the clusters. Thus

$$\mathrm{SG}(c_j) = 1/\sum_{\mathcal{C}_i \in \mathcal{C}} \mathrm{RS}(c_j, \mathcal{C}_i) \quad (19)$$

where $\mathrm{RS}(c_j, \mathcal{C}_i)$ represents the relevance score of the concept c_j to cluster $\mathcal{C}_i$, and C is the cluster pool. Many strategies can be applied to measure RS. Lu *et al.* [7] showed that the strategy of image frequency inverse tag frequency is effective. Accordingly, RS is defined as

$$\mathrm{RS}(c_j, C_i) = \begin{cases} \frac{\mathrm{OC}(c_j, C_i)}{\ln(\#(\mathcal{T}_i)+1)}, & \text{otherwise} \\ 0, & W_i = \Phi \text{ or } c_j \notin \mathcal{T}_i \end{cases} \quad (20)$$

where $\mathrm{OC}(c_j, C_i)$ represents the number of concept c_j in the associated tags of images that belong to cluster C_i, and $\mathcal{T}_i$ is the set of tags associated to the images of cluster C_i. $\#(\mathcal{T}_i)$ denotes the number of unique tags in $\mathcal{T}_i$.

TABLE I
Order of Semantic Gap Measures of Several Concepts

book (1.00)	>	dancing (0.823)	>	cars (0.545)	>
cow (0.283)	>	birds (0.0935)	>	cityscape (0.0710)	>
bridge (0.0587)	>	house (0.0354)	>	mountain (0.0185)	>
grass (0.0070)	>	sunset (0.0069)	>	water (0.0022)	—

Table I shows the order of semantic gap measures of several concepts and their normalized quantitative values. They are calculated from the training set in our experiments on NUS-WIDE-Lite [1]. The parameters are empirically set to: $k = 500$, $K = 50$, and $t = 5,000$.

V. Correlative Sparse-Graph-Based Semisupervised Learning

Until now, we have detailed the sample selection engine. For the learning engine, we extend the kNN-sparse-graph-based semisupervised learning [8] to a multilabel scenario by incorporating the semantic correlation as the basic learning model.

A. kNN-Sparse Graph Construction

Most traditional graph-based semisupervised learning algorithms construct the graphs according to visual distance and are thus very sensitive to the noise in visual features. Moreover, constructing the graph based only on visual distance will bring in semantically unrelated links between samples due to the *semantic gap*.

It has been found in neural science that the human vision system seeks a sparse representation for the incoming image using a few words in a feature vocabulary [28]. Wright *et al.* [29] demonstrated that the ℓ_1-norm-based linear reconstruction error minimization can naturally lead to a sparse representation for the images. The sparse reconstruction is robust to the noise in features and has been shown to have the ability to ensure that the images selected to reconstruct the test image are semantically related to the test image [1]. However, the one-versus-all sparse reconstruction is computationally very complex. Thus, we adopt the method proposed by Tang *et al.* [8] to construct the one-versus-kNN sparse graph by reconstructing each sample from its k nearest neighbors instead of all the other samples.

Suppose we have an underdetermined system of linear equations: $\mathbf{x} = \mathbf{D}\mathbf{w}$, where $\mathbf{x} \in \mathbb{R}^d$ is the feature vector of the image to be reconstructed, $\mathbf{w} \in \mathbb{R}^n$ is the vector of the unknown reconstruction coefficients, and $\mathbf{D} \in \mathbb{R}^{d \times n}$ $(d < n)$ is a matrix formed by the feature vectors of the other images in the data set. The sparse solution for $\mathbf{w}$ can be obtained by solving the following convex optimization problem [30]:

$$\min_{\mathbf{w}} \|\mathbf{w}\|_1, \quad s.t. \quad \mathbf{x} = \mathbf{D}\mathbf{w}. \quad (21)$$

In practice, there may exist noise on certain elements of $\mathbf{x}$, and a natural way to recover these elements and provide a robust estimation of $\mathbf{w}$ is to formulate $\mathbf{x} = \mathbf{D}\mathbf{w} + \xi$, where $\xi \in \mathbb{R}^d$ is the noise term. We can then solve the following ℓ_1-norm minimization problem with respect to both reconstruction coefficients and feature noise

$$\min_{\hat{\mathbf{w}}} \|\hat{\mathbf{w}}\|_1, \quad s.t. \quad \mathbf{x} = \mathbf{B}\hat{\mathbf{w}} \quad (22)$$

where $\mathbf{B} = [\mathbf{D}; \mathbf{I}] \in \mathbb{R}^{d \times (n+d)}$ and $\hat{\mathbf{w}} = [\mathbf{w}^{\mathrm{T}}; \xi^{\mathrm{T}}]^{\mathrm{T}}$. This optimization problem is convex and can be transformed into a general linear programming problem. There exists a globally optimal solution, and the optimization can be efficiently solved using many available ℓ_1-norm optimization toolboxes such as [31].

Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_l, \mathbf{x}_{l+1}, \ldots, \mathbf{x}_N\}$ be the set of feature vectors for the N images in the data set, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the ith

sample in the data set, and $\mathcal{G} = \{\mathcal{X}, \mathbf{W} = \{w_{ij}\}\}$ be the sparse graph with the samples in set $\mathcal{X}$ as graph vertices and $\mathbf{W}$ as the edge weight matrix. The construction of the kNN-sparse graph can be summarized as follows [8].

1) For each sample $\mathbf{x}_i$, search its k nearest neighbors $\mathcal{N}(\mathbf{x}_i)$. Here, an approximate method [32] can be applied to accelerate the process.
2) Form the matrix $\mathbf{B}_i$ with all samples $\mathbf{x}_{i_p} \in \mathcal{N}(\mathbf{x}_i)$: $\mathbf{B}_i = [\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \ldots, \mathbf{x}_{i_k}, \mathbf{I}] \in \mathbb{R}^{d\times(d+k)}$, where $p \in \{1, 2, \ldots, k\}$ and $i_p \in \{1, 2, \ldots, N\}$. Then, the vector of the reconstruction coefficients for $\mathbf{x}_i$ can be obtained by solving the following ℓ_1-norm minimization problem:
$$\min_{\mathbf{w}_i} \|\mathbf{w}_i\|_1, \quad s.t. \quad \mathbf{x}_i = \mathbf{B}_i \mathbf{w}_i \tag{23}$$
where $\mathbf{w}_i \in \mathbb{R}^{d+k}$. We call this one-versus-kNN sparse reconstruction. Note that if we set $\mathcal{N}(\mathbf{x}_i) = \{\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_N\}$, then it becomes the one-versus-all sparse reconstruction and $k = N - 1$.
3) Set the edge weight w_{ij} from the sample $\mathbf{x}_j$ to the sample $\mathbf{x}_i$ as
$$w_{ij} = \begin{cases} \mathbf{w}_i(p), & \text{if } \mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i) \text{ and } j = i_p \\ 0, & \text{if } \mathbf{x}_j \notin \mathcal{N}(\mathbf{x}_i) \end{cases} \tag{24}$$
where $i, j \in \{1, 2, \ldots, N\}$, and $\mathbf{w}_i(p)$ denotes the pth element of vector $\mathbf{w}_i$.

B. Correlative Sparse-Graph-Based Inference

We reorder the samples in image set $\mathcal{X}$ and have $\mathcal{X} = \mathcal{L} \cup \mathcal{U}$, where $\mathcal{L} = \{\mathbf{x}_1, \ldots, \mathbf{x}_l\}$ contains the first l samples labeled as $\mathbf{y}_i$ with $y_{i,j} \in \{1, 0\}$ for jth concept [1], and $\mathcal{U} = \{\mathbf{x}_{l+1}, \ldots, \mathbf{x}_N\}$ contains the unlabeled ones. We denote the matrix of the predicted labels of all samples for all concepts as $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N]^{\mathrm{T}}$.

In the multilabel scenario, the labels assigned to an image are usually consistent with the inherent label correlations [33]. For example, "beach" and "sea" are usually coassigned to a certain image since they often simultaneously appear. Motivated by this, we extended the sparse-graph-based semisupervised learning method in [8] to a multilabel framework by bringing in the semantic correlation.

Similar to the assumption in linear neighborhood propagation algorithm [34], we assume that the labels of each sample can be reconstructed from the other samples' labels, while the reconstruction coefficients are the same as those for the sparse reconstruction of sample vectors. We call this *label reconstruction assumption*. Thus, the linear reconstruction coefficients in the constructed sparse matrix can be used to predict the labels of the unlabeled samples. This prediction is based on the intuition that the weight w_{ij} reflects the likelihood for sample $\mathbf{x}_i$ to have the same label as sample $\mathbf{x}_j$. Considering the correlation between the semantic concepts, we also assume that the label prediction function over the graph should be consistent with the label correlations [35], which we call *label correlation consistency assumption*.

To capture the label correlations, we introduce a $K \times K$ symmetric matrix $\mathbf{C}$ with c_{ij} representing the correlation between label i and label j. Then, given a label matrix $\mathbf{F}_{\mathcal{L}} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_l]^{\mathrm{T}}$ on the training data points with certain labeling, $\mathbf{C}'$ is calculated as $c'_{ij} = \exp(-\|\mathbf{f}'_i - \mathbf{f}'_j\|^2/2\sigma_c^2)$, where $\mathbf{f}'_i$ is the ith column of $\mathbf{F}_{\mathcal{L}}$, and $\sigma_c^2 = E(\|\mathbf{f}'_i - \mathbf{f}'_j\|)$ is the average distance. Then, $\mathbf{C}$ is defined as $\mathbf{C} = \mathbf{C}' - \mathbf{D}_c$, where $\mathbf{D}_c$ is a diagonal matrix with the (i, i)-element equals to the sum of the ith row of $\mathbf{C}'$. Let us define e_i as $\mathbf{f}_i^{\mathrm{T}}\mathbf{C}\mathbf{f}_i$. Intuitively, e_i reflects the coherence between the inherent correlation and the label vector $\mathbf{f}_i$ assigned to $\mathbf{x}_i$. That is to say, the larger the e_i, $\mathbf{f}_i$ is more coherent with the label correlations. Consequently, we add a regularization $\mathrm{tr}(\mathbf{F}\mathbf{C}\mathbf{F}^{\mathrm{T}})$ to make the predicted multiple labels for each sample satisfy the correlations, where $\mathrm{tr}(\mathbf{M})$ is the trace of matrix $\mathbf{M}$.

Based on the *label reconstruction assumption* and the *label correlation consistency assumption*, we can infer the labels of the unlabeled samples by minimizing the label reconstruction error. Thus
$$\begin{aligned} \min_{\mathbf{f}} \quad & \sum_{i=1}^{N} \left(\mathbf{f}_i - \sum_{j\neq i} w_{ij}\mathbf{f}_j\right)^{\mathrm{T}} \left(\mathbf{f}_i - \sum_{j\neq i} w_{ij}\mathbf{f}_j\right) \\ & + \sum_{i,j} c'_{i,j}\left(\mathbf{f}'_i - \mathbf{f}'_j\right)^2 \\ s.t. \quad & \mathbf{f}_i = \mathbf{y}_i, \text{ if } \mathbf{x}_i \in \mathcal{L}. \end{aligned} \tag{25}$$
This formulation can be represented in matrix form as
$$\begin{aligned} \min_{\mathbf{F}} \quad & \mathrm{tr}\left[(\mathbf{I} - \mathbf{W})\mathbf{F}\right]^{\mathrm{T}}\left[(\mathbf{I} - \mathbf{W})\mathbf{F}\right] + \mathrm{tr}(\mathbf{F}\mathbf{C}\mathbf{F}^{\mathrm{T}}), \\ s.t. \quad & \mathbf{F}_{\mathcal{L}} = \mathbf{Y} \end{aligned} \tag{26}$$
where $\mathbf{Y}$ is the multilabel matrix for the first l samples. Let $\mathbf{B} = (\mathbf{I} - \mathbf{W})^{\mathrm{T}}(\mathbf{I} - \mathbf{W})$ and differentiate the right side of (26) with respect to $\mathbf{F}$, we obtain
$$\mathbf{M}\mathbf{F} + \mathbf{F}\mathbf{C} = \mathbf{0} \tag{27}$$
where $\mathbf{M} = (\mathbf{B} + \mathbf{B}^{\mathrm{T}})/2$ is a symmetric matrix. By splitting the matrix $\mathbf{M}$ after the lth row and lth column, we have $\mathbf{M} = \begin{bmatrix} \mathbf{M}_{\mathcal{L}\mathcal{L}} & \mathbf{M}_{\mathcal{L}\mathcal{U}} \\ \mathbf{M}_{\mathcal{U}\mathcal{L}} & \mathbf{M}_{\mathcal{U}\mathcal{U}} \end{bmatrix}$. Then, rewriting (27), we can obtain
$$\mathbf{M}_{\mathcal{U}\mathcal{U}}\mathbf{F}_{\mathcal{U}} + \mathbf{F}_{\mathcal{U}}\mathbf{C} = -\mathbf{M}_{\mathcal{U}\mathcal{L}}\mathbf{F}_{\mathcal{L}} \tag{28}$$
where $\mathbf{F}_{\mathcal{U}} = [\mathbf{f}_{l+1}, \mathbf{f}_{l+2}, \ldots, \mathbf{f}_N]^{\mathrm{T}}$.

Equation (28) is essentially a Sylvester equation [36], which is widely used in control theory. It is well known that (28) has a unique solution if and only if the eigenvalues $\eta_1, \eta_2, \ldots, \eta_{N-L}$ of $\mathbf{M}_{\mathcal{U}\mathcal{U}}$ and $\gamma_1, \gamma_2, \ldots, \gamma_K$ of $\mathbf{C}$ satisfy $\eta_i + \gamma_j \neq 0$ $(i = 1, 2, \ldots, N - L; j = 1, 2, \ldots, K)$. This condition can be easily satisfied in the real-world multilabel learning scenario. Solving the equation array (28) using the LYAP function in Matlab, we can obtain the final multilabel annotation $\mathbf{F}_{\mathcal{U}}$ of the unlabeled images.

VI. Experimental Evaluation

To demonstrate the effectiveness of the semantic gap measure in the active learning framework, we compare the proposed semantic-gap-oriented sample selection strategy with the traditional uncertainty-based sample selection strategies, which do not consider the semantic gap. The experiments are conducted on a large-scale real-world data set NUS-WIDE-Lite [1] on 81 labels. The data set is divided into two parts: development part, which contains 27 807 images, and testing part, which contains 27 808 images. The basic model is trained from the development part, and the new training samples are actively selected from the testing part. We also compare the results, learned with or without semantic correlation, to show the effectiveness of incorporating the correlation into the sparse-graph-based learning model. The low-level features we used to represent the images include 5×5 block-based color moments (225-D), edge direction histogram (73-D), and wavelet texture (128-D).

We conduct the label inference for ten rounds with 100 sampled images for each round. As shown in Table II, we found that the times of the sample used to reconstruct other samples do not affect the performance too much. Thus, we discard the $\Delta\mathcal{E}_2$ part in the following experiments. We compare four approaches by discarding $\rho|\mathcal{R}_s|$: AL_noCorr (traditional active learning without correlation), AL_Corr (traditional active

[1]It is worth to note that, in this section, the definition of $\mathcal{L}$ is a bit different from Section III. Here, $\mathcal{L}$ only includes the samples but without the labels.

TABLE II
COMPARISONS OF THE STRATEGIES DERIVED FROM $\Delta\mathcal{E}_1$, AND $\Delta\mathcal{E}_1 + (1/2|\mathcal{U}|)(\rho|\mathcal{R}_s|)$, WHERE $\rho = 0.01$ IS A TUNED TO NEAR OPTIMAL MANUALLY

AL with $\Delta\mathcal{E}_1$	0.24663	0.25547	0.26371	0.27191	0.27960	0.28536	0.29230	0.29867	0.30390	0.30963
AL with $\Delta\mathcal{E}_1 + \frac{1}{2\|\mathcal{U}\|}(\rho\|\mathcal{R}_s\|)$	0.24656	0.25596	0.26351	0.27342	0.27989	0.28571	0.29390	0.29920	0.30393	0.30965

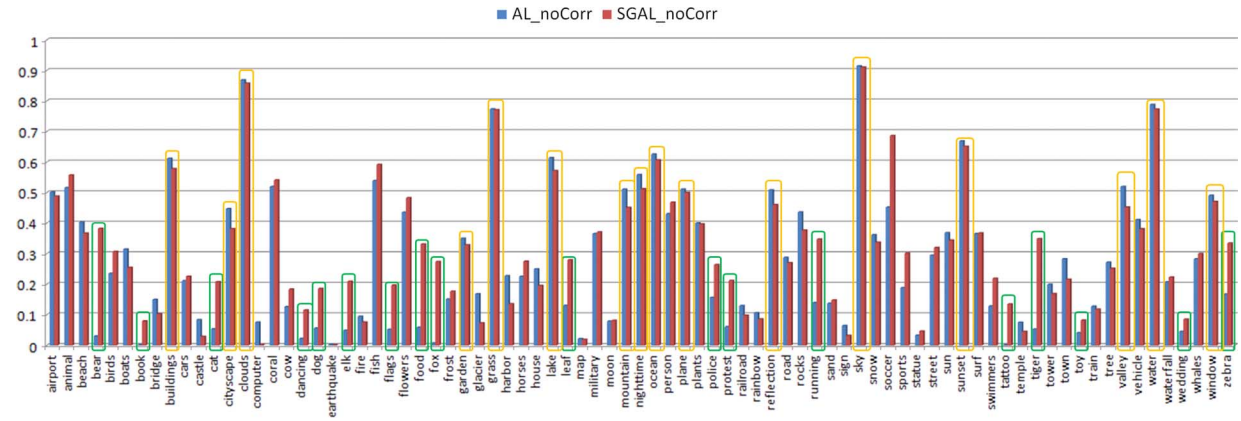

Fig. 3. Comparison of APs of the 81 concepts, obtained by traditional active learning and semantic-gap-oriented active learning.

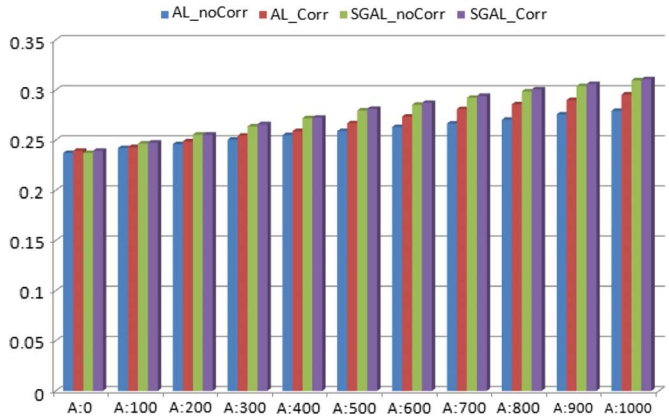

Fig. 4. Comparisons of semantic-gap-oriented active learning and traditional active learning, with and without semantic correlation. The x axis indicates the number of sampled images, and the y axis indicates the MAP.

learning with correlation), SGAL_noCorr (semantic-gap-oriented active learning without correlation), and SGAL_Corr (semantic-gap-oriented active learning with correlation). Fig. 3 compares the average precisions (APs) obtained by AL_noCorr and SGAL_noCorr with 1,000 selected samples. We can see that, although SGAL_noCorr's performance is a bit worse than that of AL_noCorr on the concepts with small semantic gap, such as those marked by the yellow rectangles, it outperforms the AL_noCorr significantly on the concepts with large semantic gap, such as those marked by green rectangles. Thus, SGAL_noCorr can significantly improve the AL_noCorr on average. Fig. 4 compares the average results obtained by the aforementioned four methods. We can see that, by integrating the semantic gap into the sample selection strategy, the performance of active learning significantly improves. For example, using 500 sampled images, the SGAL_noCorr achieves a mean average precision (MAP) of 0.2797, which has an improvement of 7.91% compared with AL_noCorr. We can also see that the utilization of semantic correlation can still improve the inference performance further. For example, the SGAL_Corr has an improvement of 1.04% over SGAL_noCorr using 500 sampled images.

Since we conduct the multilabel annotation, a feedback on one image will result in multiple feedback on labels. Thus, for each round of feedback in different approaches, we may conduct different times of user feedback. Table III compares the numbers of feedback on both images and labels for the four approaches. With the 1000 feedback on images, SGAL_Corr conducted 2998 feedback on labels, which are much less than the 4227 label feedback of AL_Corr. To this end, we can conclude that integrating the semantic gap can improve the active learning performance while reducing the manual efforts of user feedback.

TABLE III
COMPARISONS OF THE NUMBERS OF USER FEEDBACK

	AL_noCorr	AL_Corr	SGAL_noCorr	SGAL_Corr
Image Feedbacks	1000	1000	1000	1000
label Feedbacks	3592	4227	3013	2998

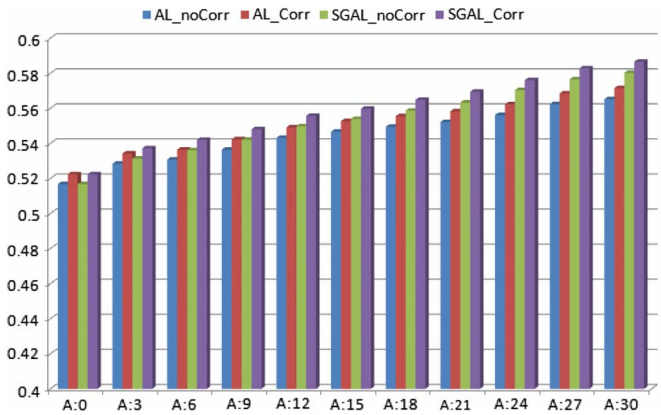

Fig. 5. Comparisons of aforementioned four approaches on Corel data set.

We also evaluate the semantic-gap-oriented active learning strategy on the Corel data set [10], on the 70 most frequent labels provided in [10], using the aforementioned three kinds of features. The obtained MAPs of the four approaches are compared in Fig. 5. We can see that the semantic-gap-oriented sample selection strategy can consistently improve the traditional sample selection for about 2.5% and integrating the semantic correlation can also consistently get an approximate 1.2% further improvement. The improvements from the semantic-gap-oriented sample selection strategy here are not as significant as those in NUS-WIDE-Lite data set. It is due to that the labels evaluated in this

data set are the most frequent ones, which may have similar quantities of low semantic gaps.

VII. Conclusion and Future Work

The size of the semantic gap is an important factor that affects the performance of active learning. This paper presented a semantic-gap-oriented active learning method, which brings the semantic gap measure into the information minimization strategy to account for the effect of semantic gap in the sample selection strategy. We extended the sparse-graph-based semisupervised learning method to multilabel setting by integrating the semantic correlation as the basic learning model. Extensive experiments conducted on two benchmark image data sets have demonstrated that integrating the semantic gap measure into the sample selection strategy can significantly improve the active learning effectiveness while reducing the manual efforts of user feedback. The results also found that the semantic correlation is helpful to improve the sparse-graph-based learning method. In the future work, we will integrate the other criteria such as diversity and relevance into the semantic-gap-oriented active learning framework to further improve the performance.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable suggestions.

References

[1] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua, "Inferring semantic concepts from community contributed images and noisy tags," in *Proc. ACM Multimedia*, 2009, pp. 223–232.

[2] X.-J. Wang, L. Zhang, X. Li, and W.-Y. Ma, "Annotating images by mining image search results," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1919–1932, Nov. 2008.

[3] J. Tang, H. Li, G.-J. Qi, and T.-S. Chua, "Image annotation by graph-based inference with integrated multiple/single instance representations," *IEEE Trans. Multimedia*, vol. 12, no. 2, pp. 131–141, Feb. 2010.

[4] T. Huang, C. Dagli, S. Rajaram, E. Chang, M. Mandel, G. Poliner, and D. Ellis, "Active learning for interactive multimedia retrieval," *Proc. IEEE*, vol. 96, no. 4, pp. 648–667, Apr. 2008.

[5] M. Wang and X.-S. Hua, "Active learning in multimedia annotation and retrieval: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 2, Feb. 2011, Art. 10.

[6] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[7] Y. Lu, L. Zhang, J. Liu, and Q. Tian, "Constructing lexica of high-level concepts with small semantic gap," *IEEE Trans. Multimedia*, vol. 12, no. 4, pp. 288–299, Jun. 2010.

[8] J. Tang, R. Hong, S. Yan, T.-S. Chua, G.-J. Qi, and R. Jain, "Image annotation by kNN-sparse graph-based label propagation over noisily-tagged web images," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 2, Feb. 2011, Art. 14.

[9] T.-S. Chua and J. Tang *et al.*, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. ACM CIVR*, Santorini, Greece, 2009.

[10] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 97–112.

[11] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *J. Artif. Intell. Res.*, vol. 4, no. 1, pp. 129–145, Jan. 1996.

[12] N. Roy and A. Mccallum, "Toward optimal active learning through Monte Carlo estimation of error reduction," in *Proc. Int. Conf. Mach. Learn.*, 2001, pp. 441–448.

[13] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proc. ACM Multimedia*, 2001, pp. 107–118.

[14] J. He, M.-J. Li, H.-J. Zhang, H. Tong, and C. Zhang, "Mean version space: A new active learning method for content-based image retrieval," in *Proc. Workshop Multimedia Inf. Retrieval*, 2004, pp. 15–22.

[15] Y. Wu, I. Kozintsev, J. yves Bouguet, and C. Dulong, "Sampling strategies for active learning in personal photo retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2006, pp. 529–532.

[16] A. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2009, pp. 2372–2379.

[17] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Batch mode active learning and its application to medical image classification," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 417–424.

[18] C. Dagli, S. Rajaram, and T. Huang, "Leveraging active learning for relevance feedback using an information theoretic diversity measure," in *Proc. Int. Conf. Image Video Retrieval*, 2006, pp. 123–132.

[19] C. Zhang and T. Chen, "Annotating retrieval database with active learning," in *Proc. IEEE Int. Conf. Image Process.*, 2003, pp. II-595–II-598.

[20] H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," in *Proc. 21st In. Conf. Mach. Learn.*, 2004, p. 79.

[21] S. Ayache and G. Quénot, "Evaluation of active learning strategies for video indexing," *Image Commun.*, vol. 22, no. 7/8, pp. 692–704, Aug. 2007.

[22] P. H. Gosselin and M. Cord, "A comparison of active classification methods for content-based image retrieval," in *Proc. 1st Int. Workshop CVDB*, 2004, pp. 51–58.

[23] M. Wang, X.-S. Hua, Y. Song, J. Tang, and L.-R. Dai, "Multiconcept multi-modality active learning for interactive video annotation," in *Proc. Int. Conf. Semantic Comput.*, 2007, pp. 321–328.

[24] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang, "Two-dimensional multilabel active learning with an efficient online adaptation model for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1880–1897, Oct. 2009.

[25] S. Vijayanarasimhan and K. Grauman, "What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2009, pp. 2262–2269.

[26] M. E. Hellman and J. Raviv, "Probability of error, equivocation and the Chernoff bound," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 4, pp. 368–372, Jul. 1970.

[27] C. Elkan, "Using the triangle inequality to accelerate k-means," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 147–153.

[28] R. Rao, B. Olshausen, and M. Lewicki, *Probabilistic Models of the Brain: Perception and Neural Function*. Cambridge, MA: MIT Press, 2002.

[29] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[30] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," *Commun. Pure Appl. Math.*, vol. 59, no. 6, pp. 797–829, 2006.

[31] ℓ_1-magic. [Online]. Available: http://users.ece.gatech.edu/~justin/l1magic/

[32] D. Mount and S. Arya, "Ann: A library for approximate nearest neighbor searching," in *Proc. CGC 2nd Annu. Fall Workshop Comput. Geom.*, Durham, NC, 1997.

[33] J. Tang, X.-S. Hua, M. Wang, Z. Gu, G.-J. Qi, and X. Wu, "Correlative linear neighborhood propagation for video annotation," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 2, pp. 409–416, Apr. 2009.

[34] F. Wang and C. Zhang, "Label propagation through linear neighborhoods," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 1, pp. 55–67, Jan. 2008.

[35] Z.-J. Zha, T. Mei, J. Wang, Z. Wang, and X.-S. Hua, "Graph-based semi-supervised learning with multiple labels," *J. Vis. Commun. Image Representations*, vol. 20, no. 2, pp. 97–103, Feb. 2009.

[36] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*. New York: Academic, 1987.