Asymmetric Semi-Supervised Boosting Scheme for Interactive Image Retrieval

Jun Wu and Ming-Yu Lu

Support vector machine (SVM) active learning plays a key role in the interactive content-based image retrieval (CBIR) community. However, the regular SVM active learning is challenged by what we call "the small example problem" and "the asymmetric distribution problem." This paper attempts to integrate the merits of semi-supervised learning, ensemble learning, and active learning into the interactive CBIR. Concretely, unlabeled images are exploited to facilitate boosting by helping augment the diversity among base SVM classifiers, and then the learned ensemble model is used to identify the most informative images for active learning. In particular, a bias-weighting mechanism is developed to guide the ensemble model to pay more attention on positive images than negative images. Experiments on 5000 Corel images show that the proposed method yields better retrieval performance by an amount of 0.16 in mean average precision compared to regular SVM active learning, which is more effective than some existing improved variants of SVM active learning.

Keywords: Interactive image retrieval, support vector machines, semi-supervised learning, active learning, boosting.

I. Introduction

To narrow the gap between low-level visual features and high-level semantic concepts, human interactive techniques have been introduced into content-based image retrieval (CBIR), which is drawing substantial research attention in recent years [1], [2]. Unlike early CBIR, which adopted automatic strategies, recent approaches focus on the interaction between the user and the search engine. Concretely, in a relevance feedback (RF) loop, the user has the option to label a few images returned as either positive or negative in terms of whether they are relevant to the query concept or not. Labeled images are then given to the system as complementary queries so that the search engine can be refined.

Up to the present, various interactive schemes have been proposed which evolved from earlier heuristic methods to probability modeling approaches, and recent classification/clustering-based techniques. Among the various approaches, support vector machine (SVM)-based RF schemes represent the state-of-the-art techniques for improving CBIR performance [3]-[9]. In the RF loop, prompting the user to label images is an important step, but this is a very burdensome task for the user. Active learning plays a key role in alleviating the burden of labeling in the RF loop [10]. The main idea is to actively select the most informative unlabeled images for the user to label, with the aim of greatly improving the retrieval performance. Tong and others proposed the popular active learning-based RF technique called SVM active learning (SVM-AL) [5], which aims to search data points that can maximally reduce the size of version space. They proved that this goal can be approximately achieved by selecting points near the SVM boundary, and thus unlabeled images close to the boundary were regarded as the most informative data points.

Manuscript received Feb. 9, 2010; revised July 21, 2010; accepted Aug. 4, 2010.

This work was partially supported by Natural Science Foundation of China (No. 60773084, No. 60973067).

Jun Wu (phone: +86 134 7860 0994, email: wujunas8@gmail.com) and Ming-Yu Lu (lumingyu@tsinghua.org.cn) are with the School of Information Science & Technology, Dalian Maritime University, Liaoning, China.

doi:10.4218/etrij.10.1510.0016

However, SVM-AL has two main drawbacks. First, SVM may fail to learn an accurate classification model from a small number of labeled examples. Given a limited number of training data labeled by the user, directly applying SVM model may not significantly improve the retrieval accuracy, although it enjoys excellent generalization performance. Second, unlike the traditional pattern classification problem, the relevant and irrelevant classes in an image database are highly imbalanced (there are fewer positive examples than negative ones), thus the learned SVM boundary may be biased toward the negative side. We refer to these two problems as the small example problem and the asymmetric distribution problem, respectively.

This paper aims to improve SVM-AL by following two strategies regarding these two problems. First, to tackle the small example problem, we focus on improving the SVM model. Using semi-supervised boosting technique and enhanced SVM boundary is more helpful to identify the informative unlabeled images for active learning. Second, to attack the asymmetric distribution problem, a bias-weighting mechanism is used in our solution so that positive images are paid more attention than negative images. Our empirical study shows encouraging results in comparison to some existing SVM-based active learning approaches.

The rest of this paper is organized as follows. Section II presents the problem formulation and our solution. Section III shows experimental evaluations. Section IV discusses related works. Finally, section V concludes this paper.

II. Proposed Algorithm

1. Preliminaries

Given a query image, it is natural that the image database can be divided into two classes: one is relevant (positive) in semantic content to the query, and the other is irrelevant (negative). Hence, the learning problem in RF is essentially reduced to a binary classification problem.

Let $\mathbf{DB}=\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n\}$ denote the entire image database, including both labeled image example set and unlabeled image example set. Suppose the first *nl* examples are labeled, given by $\mathbf{y}_l=\left[y_1^l,y_2^l,\cdots,y_{nl}^l\right]$, where each class label $y_i^l\in\{+1,-1\}$. Similarly, class labels of unlabeled examples can be denoted as $\mathbf{y}_u=\left[y_1^u,y_2^u,\cdots,y_{nu}^u\right]$, where $nu=n-nl$. Therefore, labels for the entire image dataset can be denoted as $\mathbf{y}=\left[\mathbf{y}_l;\mathbf{y}_u\right]$. The goal of our solution is to iteratively update the class labels of unlabeled examples and then construct a new classifier using both labeled and pseudo-labeled examples in order to enhance the generalization ability of the learning system. The enhanced classification model is used to identify the most informative examples which can improve the retrieval accuracy most efficiently.

Since SVM-based RF schemes have shown many promising results [3]-[9], our study focuses on applying SVM as the base classifier of proposed learning system. Here, we briefly review SVM. The key idea of SVM is to learn an optimal hyperplane that separates the training examples with the maximal margin by solving the following optimization problem [11]:

$$\max_{\lambda\geq 0}\left(\sum_{i=1}^{nl}\lambda_i-\frac{1}{2}\sum_{i,j=1}^{nl}\lambda_i\lambda_j y_i y_j\, K\left(\mathbf{x}_i,\mathbf{x}_j\right)\right), \qquad (1)$$

$$\text{s.t.}\sum_{i=1}^{nl}\lambda_i y_i=0,$$

where λ_i is Lagrange multiplier, $K(\cdot)$ is a kernel function which can project examples from the original data space to a Hilbert inner product space. For a given kernel function, the decision function of an SVM classifier is given by

$$f(\mathbf{x})=\sum_{i=1}^{nl}\lambda_i y_i\, K\left(\mathbf{x}_i,\mathbf{x}\right)+b\,. \qquad (2)$$

Let $\text{Abs}(\boldsymbol{a})$ denote the function used to produce the absolute value of $\boldsymbol{a}$. In general, when $\text{Abs}(f(\mathbf{x}))$ for a given pattern is high, the corresponding prediction confidence will be high. Meanwhile, a low $\text{Abs}(f(\mathbf{x}))$ of a given pattern means that the pattern is close to the decision boundary and its corresponding prediction confidence will be low. As a result, the decision function $f(\mathbf{x})$ has been used to measure the dissimilarity between a given pattern and the query image in many SVM-based RF schemes.

2. Asymmetric Semi-Supervised Boosting for SVM Active Learning

To generate strong learning systems, ensemble learning [12] tries to mine the complementary information of multiple classifiers, while semi-supervised learning [13] aims to benefit from the unlabeled data. As indicated by Zhou [14], however, ensemble learning and semi-supervised learning are actually mutually beneficial. A key element is that exploiting unlabeled data in an ensemble is helpful to augment the diversity among individual classifiers. Therefore, combing the advantages of ensemble learning and semi-supervised learning has been an appealing research theme. Some successful studies have been reported [15]-[20], most of which are semi-supervised boosting (SemiBoost) techniques [15], [16], [19], [20]. Specifically, the empirical study in [15] showed that a SemiBoost technique can effectively improve SVM performance. Based on the discussion above, we introduce a SemiBoost technique into our solution and modify it for the purpose of CBIR, that is, ensemble learning with asymmetry under the semi-supervised setting.

Similar to boosting, the main idea of SemiBoost is to train an ensemble classifier iteratively [15]. At each round of iterations, the pseudo-labels of the unlabeled examples are predicted using existing ensemble and the pairwise similarity between examples, and then a few confidently pseudo-labeled examples in conjunction with all labeled ones are used to train a new classifier. Finally, all of the learned classifiers will be combined to form the final ensemble.

Let $\mathbf{S}=[\boldsymbol{S}_{i,j}]^{n\times n}$ denote the symmetric similarity matrix, where $\boldsymbol{S}_{i,j} \geq 0$ represents the similarity between example $\mathbf{x}_i$ and $\mathbf{x}_j$. Let $h^{(t)}(\mathbf{x})\text{:}\ \boldsymbol{X} \to \{0,1\}$ denote the individual classifier learned at the t-th iteration (SVM is considered in this paper). Let $H(\mathbf{x}) : \boldsymbol{X} \to \boldsymbol{R}$ denote the ensemble model learned after T iterations. It is computed as a linear combination of T individual classifiers, that is, $H(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t h^{(t)}(\mathbf{x})$, where α_t is the combination weight. At the (T+1)st iteration, SemiBoost aims to find a new component classifier h($\mathbf{x}$) and the combination weight α by solving the following optimization problem:

$$\begin{aligned}
&\underset{h(\mathbf{x}),\alpha}{\mathbf{argmin}}\left\{F(\mathbf{y},\mathbf{S}) = F_l\left(y,\mathbf{S}^{lu}\right) + CF_u\left(\mathbf{y}_u,\mathbf{S}^{uu}\right)\right\} \\
\Leftrightarrow\ &\underset{h(\mathbf{x}),\alpha}{\mathbf{argmin}}\left\{\sum_{i=1}^{nl}\sum_{j=1}^{nu} \boldsymbol{S}_{i,j}^{lu} \exp\left(-2y_i^l\left(H_j + \alpha h_j\right)\right)\right. \\
&\left.\qquad + C\sum_{i,j=1}^{nu} \boldsymbol{S}_{i,j}^{uu} \exp\left(H_i - H_j\right)\exp\left(\alpha\left(h_i - h_j\right)\right)\right\}, \qquad (3) \\
&\text{s.t.}\quad h_i = y_i^l, i = 1\cdots nl,
\end{aligned}$$

where $H_i \equiv H(\mathbf{x}_i)$, $h_i \equiv h(\mathbf{x}_i)$, and the constant C=nl/nu is introduced to weigh the importance between the labeled and the unlabeled data. The objective function of SemiBoost F is a combination of two terms: F_l measures the inconsistency between labeled and unlabeled examples, and F_u measures the inconsistency among unlabeled examples.

To simplify the computation, the above optimization problem can be transformed into a simple format. More details can be found in [15].

$$F_1 = \sum_{i=1}^{nu} \exp(-2\alpha h_i)\boldsymbol{p}_i + \exp(2\alpha h_i)\boldsymbol{q}_i, \qquad (4)$$

$$\boldsymbol{p}_i = \sum_{j=1}^{nl} \boldsymbol{S}_{i,j}^{ul}\exp(-2H_i)\delta\left(y_j,1\right) + \frac{C}{2}\sum_{j=1}^{nu} \boldsymbol{S}_{i,j}^{uu}\exp(H_j - H_i), \quad (5)$$

$$\boldsymbol{q}_i = \sum_{j=1}^{nl} \boldsymbol{S}_{i,j}^{ul}\exp(2H_i)\delta\left(y_j,-1\right) + \frac{C}{2}\sum_{j=1}^{nu} \boldsymbol{S}_{i,j}^{uu}\exp(H_i - H_j), \quad (6)$$

where $\delta(x,y) = 1$ when $x = y$ and 0 otherwise. The quantities $\boldsymbol{p}_i$ and $\boldsymbol{q}_i$ can be interpreted as the confidence in classifying the unlabeled example $\mathbf{x}_i$ into the positive class and the negative class, respectively. Since F_1 is difficult to optimize, its upper bound F_2 is then constructed.

$$\begin{aligned}
F_1 \leq F_2 = &\sum_{i=1}^{nu}\left(\boldsymbol{p}_i + \boldsymbol{q}_i\right)\left(\exp(2\alpha) + \exp(-2\alpha) - 1\right) \\
&-\sum_{i=1}^{nu} 2\alpha h_i\left(\boldsymbol{p}_i - \boldsymbol{q}_i\right). \qquad (7)
\end{aligned}$$

Obviously, F_2 is minimized when $h_i = \text{sign}\left(\boldsymbol{p}_i - \boldsymbol{q}_i\right)$ for maximum value of $\text{Abs}(\boldsymbol{p}_i - \boldsymbol{q}_i)$. Therefore, to minimize F_2, the optimal pseudo-label, z_i, for the example $\mathbf{x}_i$ is $\mathrm{z}_i = \text{sign}(\boldsymbol{p}_i - \boldsymbol{q}_i)$, and its corresponding prediction confidence is $\text{Abs}(\boldsymbol{p}_i - \boldsymbol{q}_i)$. Also, by differentiating F_2 with regard to α and setting it to 0, the optimal α that minimizes the objective function is

$$\alpha = \frac{1}{4}\ln\frac{\sum_{i=1}^{nu}\boldsymbol{p}_i\delta\left(h_i,1\right) + \sum_{i=1}^{nu}\boldsymbol{q}_i\delta\left(h_i,-1\right)}{\sum_{i=1}^{nu}\boldsymbol{p}_i\delta\left(h_i,-1\right) + \sum_{i=1}^{nu}\boldsymbol{q}_i\delta\left(h_i,1\right)}. \qquad (8)$$

For any given query, only a small number of images in the database are positive while most images are negative, that is, the relevant and irrelevant classes are highly imbalanced. However, the SemiBoost fails to take this class-imbalance problem into account. Learning algorithms that do not consider class-imbalance tend to be overwhelmed by the majority class and ignore the minority class [21]. A few methods have been proposed to tackle the class-imbalance learning, such as asymmetric boosting [22] and easy-ensemble [23]. The former raises the minority class examples' weights in the boosting process, while the latter splits the majority class into several subsets in order to train each weak classifier on a balanced number of positive and negative examples within a ensemble framework. However, in these approaches, since the minority examples are overemphasized by every individual classifier, combining these classifiers will have a high probability of suffering from overfitting when the number of minority class examples is limited. Also, most of current class-imbalance learning methods are ensemble strategies, and thus embedding them into SemiBoost, that is, 'nested ensemble' structure, will require much training time.

In the CBIR context, the user is more interested in positive images rather than negative images. Hence, individual classifiers with a high *true positive rate* should be emphasized. Considering this, we modify the weighting strategy for the purpose of CBIR, and the new weighting strategy is termed the bias-weighting mechanism:

$$\begin{aligned}
&\alpha = \frac{\eta}{4}\ln\frac{\sum_{i=1}^{nu}\boldsymbol{p}_i\delta\left(h_i,1\right) + \sum_{i=1}^{nu}\boldsymbol{q}_i\delta\left(h_i,-1\right)}{\sum_{i=1}^{nu}\boldsymbol{p}_i\delta\left(h_i,-1\right) + \sum_{i=1}^{nu}\boldsymbol{q}_i\delta\left(h_i,1\right)} + (1-\eta)\exp\left(\boldsymbol{tpr}\right), \\
&\boldsymbol{tpr} = \Pr\left[h\left(\mathbf{x}_j\right) = y_j\ \&\ y_j = +1\right],\quad j = 1,\ldots,nl+m, \qquad (9)
\end{aligned}$$

Algorithm 1. Proposed ASB-SVM-AL algorithm**.**

```
Input: $\mathbf{DB} = \boldsymbol{L} \cup \boldsymbol{U}$ : image database ($\boldsymbol{L}$ and $\boldsymbol{U}$ are the labeled
          and unlabeled example sets, respectively);
       $T$: number of iteration in boosting;
       $\sigma$ : sampling scale in iterations.
1. for $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbf{DB}$ do $\boldsymbol{S}_{i,j} \leftarrow \exp\left(-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right)$;
   % compute similarities between any two images in DB
In each round of relevance feedback:
2. $\boldsymbol{L} \leftarrow \boldsymbol{L} \cup \boldsymbol{L}_{\text{new}}$ ; $\boldsymbol{U} \leftarrow \boldsymbol{U} - \boldsymbol{L}_{\text{new}}$ ;
3. Repeat $t$ = 1 to $T$
    3.1. for $\mathbf{x} \in \boldsymbol{U}$ compute $\boldsymbol{p}$ and $\boldsymbol{q}$ using (5) and (6), and
         its pseudo label $z = \text{sign}(\boldsymbol{p} - \boldsymbol{q})$ ;
    3.2. $\boldsymbol{L}^* \leftarrow \text{Sampling}(\boldsymbol{U}, \text{Abs}(\boldsymbol{p}-\boldsymbol{q}), \sigma)$ , $|\boldsymbol{L}^*| = \text{fix}(\sigma|\boldsymbol{U}|)$ ;
         % $\boldsymbol{L}^*$stores a few pseudo-labeled examples sampled
         % from $\boldsymbol{U}$ with weight Abs$(\boldsymbol{p}-\boldsymbol{q})$
    3.3. $h^{(t)} \leftarrow \text{SVM\_Train}(\boldsymbol{L} \cup \boldsymbol{L}^*)$ ;
    3.4. Compute $\alpha_t$ using (9);
    3.5. $H \leftarrow H + \alpha_t h(\text{t})$ ; % update the ensemble classifier
4. End Repeat
5. $H \leftarrow \text{Normalize}(H)$ ; % normalize $H$ to $(-1,1)$
6. for $\mathbf{x} \in \boldsymbol{U}$ do $\mathbf{Pool} \leftarrow \text{Sort}_{\text{Asc}}(\text{Abs}(H(\mathbf{x})))$ ;
7. for $\mathbf{x} \in \mathbf{DB}$ do $\mathbf{Result} \leftarrow \text{Sort}_{\text{Dsc}}(H(\mathbf{x}))$ ;
Output: Pool, Result
```

where ***tpr*** denotes the true positive rate of classifier $h(\mathbf{x})$ learned from a mixture of nl labeled examples and m pseudo-labeled examples. exp(***tpr***) is used to augment the relative contribution of ***tpr***. In (9), the first item reflects the general learning performance from both labeled and unlabeled data of an individual classifier, while the second item reflects the ability of an individual classifier on detecting positive images. Given several individual classifiers with the same general performance, bias-weighting will emphasize the individuals with the strong ability of detecting positive images. In other words, under the influence of bias-weighting, the ensemble classification model pays more attention to positive images than negative ones. $\eta \in (0, 1]$ is used to control the relative contribution of each component. When $\eta = 1$, it means that the bias item is ignored: the smaller η is, the more the bias item contributes.

Furthermore, since the classifier learned from RF is not strong, especially in the early rounds, the pseudo-labels they assign to the unlabeled examples may be incorrect. Hence, the pseudo-labeled examples are only temporarily used as training examples, and in the next round of RF, they will be treated as unlabeled data again. In this way, the influence of the possible mistakes made by the classifiers can be limited.

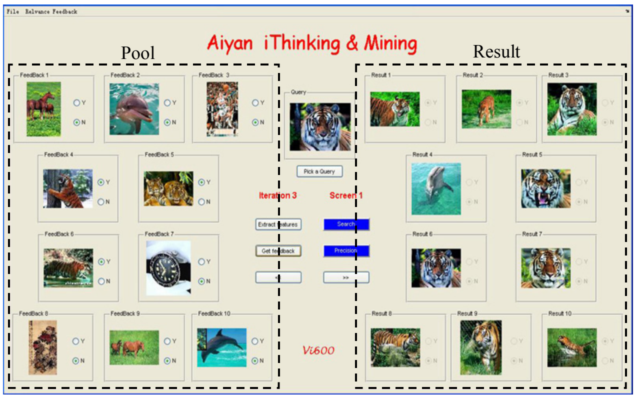

Fig. 1. User interface of prototype system.

In summary, inspired by SVM-AL, the proposed solution asymmetric SemiBoost-based SVM active learning (ASB-SVM-AL) is presented in algorithm 1, where $|\bullet|$ denotes the size of a set, and $\text{fix}(\bullet)$ denotes the mantissa rounding operator. Similar to other active RF schemes, ASB-SVM-AL can actively put the most informative unlabeled images into a "pool" for the user to label, while the retrieval "result" is separated from the pool for feedbacks. The user interface of a prototype system is shown in Fig. 1. Note that the pairwise similarity between any two images in the database can be computed off-line. Thus, ASB-SVM-AL can be quite efficient in processing online queries.

III. Experiments

1. Comparison Methods

To evaluate the performance of the proposed ASB-SVM-AL algorithm, we compare it with three other previous well-known active learning approaches:

- SVM active learning (SVM-AL) [5]: the baseline method that directly learns an SVM model from the labeled examples and then selects the unlabeled examples closest to the decision boundary for labeling,
- Boost SVM active learning (BSVM-AL) [6]: a modification of SVM-AL that enhances SVM performance using the AdaBoost technique,
- Transductive SVM active learning (TSVM-AL) [7]: another modification of SVM-AL that improves SVM performance exploiting unlabeled data within the transductive learing framework.

In the interactive CBIR community, ensemble learning and semi-supervised learning are two popular methods used to improve active learning performance. Therefore, BSVM-AL and TSVM-AL are selected in the comparison. The library for SVM (LIBSVM) software [24] was used for all methods to solve the SVM optimization problem, and the radial basis function (RBF) kernel is used in SVM.

Furthermore, in order to study whether the bias-weighting mechanism is useful to address the asymmetric distribution problem, a degenerated variant of ASB-SVM-AL, that is, SemiBoost-based SVM active learning (SB-SVM-AL) is evaluated for comparative purposes. Roughly speaking, the SB-SVM-AL is almost the same as ASB-SVM-AL except that the former does not consider the asymmetry distribution problem. Specifically, the parameter η is set differently in the two approaches: ASB-SVM-AL aims at emphasizing the importance of positive examples by setting $\eta = 0.3$, while SB-SVM-AL regards the positive and negative examples equally, that is, $\eta = 1$. The reason for setting $\eta = 0.3$ is illustrated in Fig. 2. We set $\eta \in \{0.1, 0.2, \ldots, 1.0.\}$. After tuning all the values in this pool, we found ASB-SVM-AL performs best when $\eta = 0.3$.

2. Configurations

To form the testing image dataset, 50 semantic categories picked from the COREL database are used. Each category contains 100 images, and there are 5,000 images in total. We use three types of features to describe the images:

- Color: the color features are derived using a 4×4×4 bin histogram in HSV space.
- Texture: the texture features are derived using a 3-level pyramidal wavelet transform from the Y component in YCbCr space. Then, the mean and variance calculating in each of 9 high-frequent sub-bands is used to form an 18-dimension vector.
- Shape: the edge direction histogram (EDH) is employed to capture the spatial distribution of edges as a shape figure. EDH is calculated upon the Y component in YCbCr space using a Sobel detector and quantized into five bins; namely, horizontal, 45 diagonal, vertical, 135 diagonal, and isotropic.

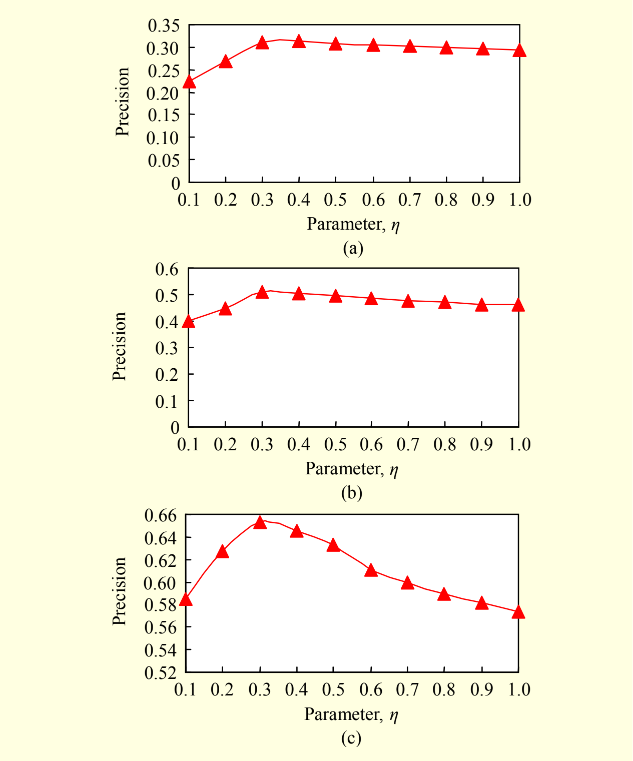

Fig. 2. Performance (P@Top50) of proposed algorithm with various η at the first, third, and fifth feedback.

To evaluate the average performance, 250 queries were randomly selected from the image set. At the beginning of retrieval, images in the database were ranked according to their Euclidean distances to the query, and the top 10 images were labeled as the set of initially labeled images for the learning system. Then, various methods were applied to rerank the images in the database. For each compared method, after obtaining initial labeled images, five rounds of feedback were performed.

In many interactive CBIR systems, the user is required to label 20 to 40 images in each round of feedback, which is not practical because few users are patient to label so many images. In our system, 10 images, judged to be the most informative ones, are put into the pool at each round of feedback (see Fig. 1). In particular, only the positive images are required to be marked by the user, and all the other images are automatically marked as negative by the system. In general, less than 50% of the images in the pool are positive. Thus, only about five images in each round of RF are required to be labeled by the user, which is more practical than many previous interactive CBIR systems.

We adopted an experimental design technique to select optimal values of parameters T and σ. The feasible values of them are set to {5, 10, 15, 20} and {5%, 10%, 15%, 20%}, respectively. In the experiment, we found that, with T and σ growing, the performance of ASB-SVM-AL improved slowly, while the computational time increased quickly. Considering the tradeoff between effectiveness and complexity, T and σ were set to 5 and 5%, respectively.

As a measure of retrieval accuracy, we used the precision at top N retrieval results (P@Top$N$) [25]. The precision-and-recall graph (PR-graph) is a well-known measure for information retrieval systems, but it fails to reflect the changes of the performance caused by feedback directly. P@Top$N$ describes the relationship between precision and feedback iteration at the

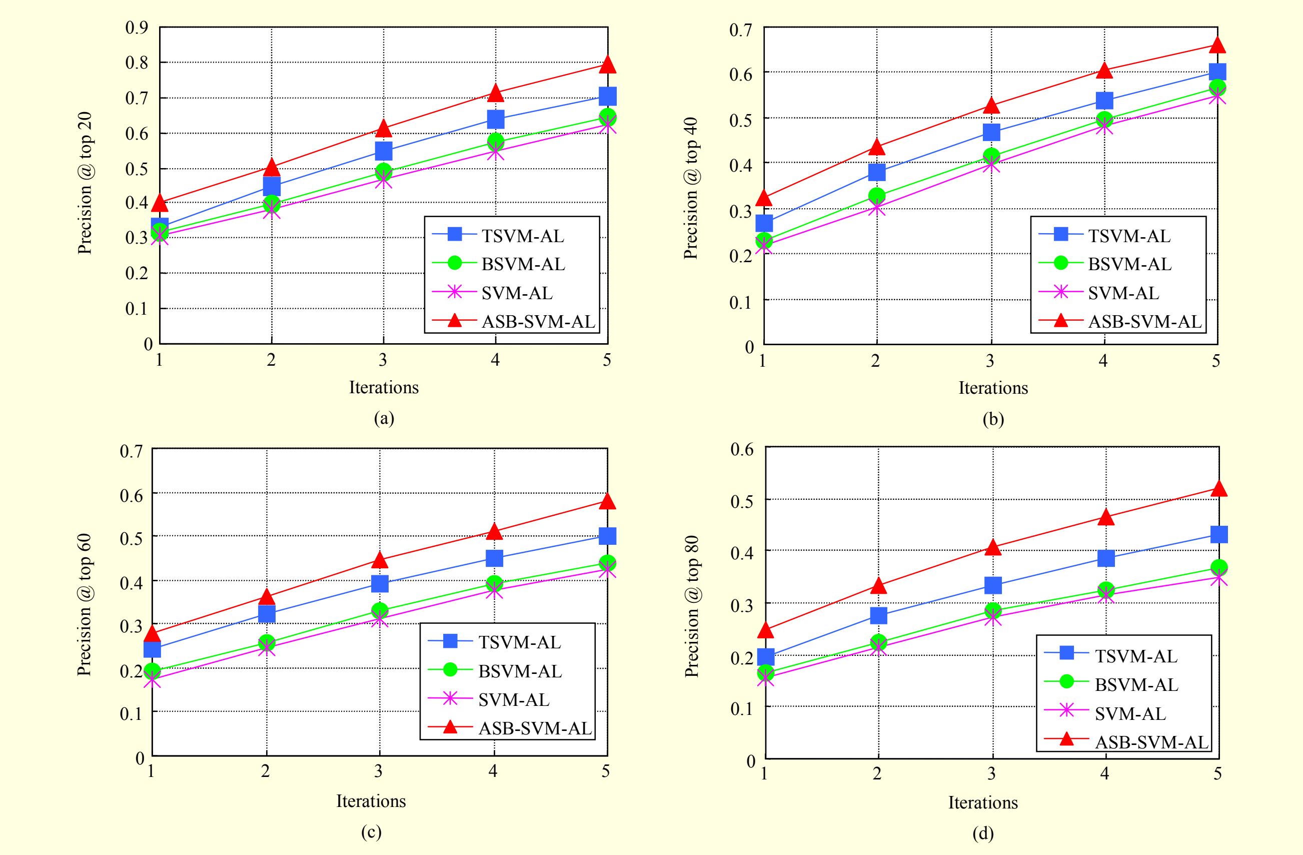

Fig. 3. Precision at top 20, 40, 60, and 80 retrieval results of proposed algorithm compared with some existing methods.

top *N* retrieval results. Hence, P@Top*N* is a more reasonable choice than a PR-graph to evaluate the retrieval performance of interactive CBIR systems.

3. Results

Here we compare the performance of ASB-SVM-AL, SVM-AL, BSVM-AL, and TSVM-AL. Figure 3 shows the precision curves of the different methods at the top 20, top 40, top 60, and top 80 retrieval images. The detailed final testing result, that is, the mean average precision (MAP), is also shown in Table 1. Several observations can be drawn from the experimental results. First, by examining the results of all methods, we found that the BSVM-AL is only marginally better than the baseline method SVM-AL. The main reason is that AdaBoost can hardly boost the performance of strong classifiers such as SVM since the base SVM classifiers learned from the limited number of labeled examples are similar to each other. Consequently, the boosting method degenerates to a single strong classifier. Second, two semi-supervised learning solutions, ASB-SVM-AL and TSVM-AL, outperform the other two supervised learning methods. Finally, comparing the two semi-supervised learning algorithms, we found that the proposed ASB-SVM-AL achieves significantly better performance than TSVM-AL. This result demonstrates that the

Table 1. Average precisions of different algorithms at top *N* retrieval results after five rounds of feedback.

	Top 20	Top 40	Top 60	Top 80	MAP
ASB-SVM-AL	0.79	0.66	0.58	0.52	0.64+11.7%
TSVM-AL	0.70	0.59	0.51	0.43	0.56+11.5%
BSVM-AL	0.64	0.56	0.44	0.37	0.50+12.1%
SVM-AL	0.62	0.55	0.41	0.34	0.48+12.8%

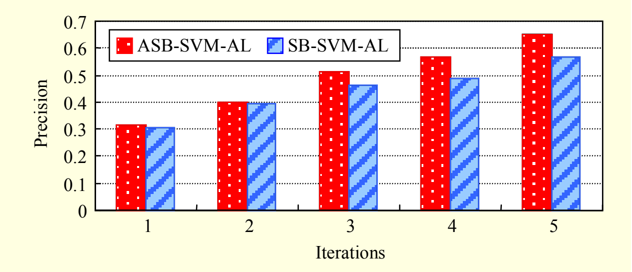

Fig. 4. Performance (P@Top50) comparison between ASB-SVM-AL and SB-SVM-AL.

semi-supervised ensemble paradigm is more efficient than the conventional semi-supervised learning method.

Furthermore, in order to study the effectiveness of the bias-weighting mechanism employed by our solution, ASB-SVM-AL is compared with SB-SVM-AL. Figure 4 shows the comparison results of the two algorithms. As can been seen, with growing rounds of feedback, ASB-SVM-AL increasingly outperforms SB-SVM-AL. It is conjectured that the labeled positive and negative examples are nearly equal in the early rounds of RF. By gradually adding the user's feedback, the labeled positive and negative examples become unequal. Thus, the bias-weighting mechanism is increasingly helpful to ASB-SVM-AL.

IV. Related Works

To improve the learning efficiency of RF, active learning paradigms have been studied in recent years. SVM-AL [5] is a well-known and pioneering work that plays an important role in the CBIR community. Its limitations have been addressed by research efforts. For the small example problem, Jiang and others [6] tried to improve the performance of SVM-AL by using the AdaBoost technique. However, as mentioned before, directly using AdaBoost hardly improves SVM performance. This was also pointed out by Tao and others [9], so they focused on improving SVM by using the bagging technique and feature selection strategy. Wang and others [7] proposed TSVM-AL that validates the SVM boundary by using unlabeled data within the transductive learning framework. Similarly, Hoi and others [8] proposed a better solution in which a kernel is first learned for SVM from a mixture of labeled and unlabeled data. The kernel is then used to identify the informative examples for active learning. However, the solutions only using ensemble learning or semi-supervised learning may not improve the RF performance significantly. There is a key difference between our proposed solution and the previous methods. This paper deals with an SVM ensemble under the semi-supervised setting. Since unlabeled data is exploited in the boosting framework, the diversity among the SVM classifiers is augmented. As a result, the performance of the SVM model is efficiently boosted.

Furthermore, the proposed algorithm is closely related to the asymmetric bagging-based SVM (AB-SVM) approach proposed by Tao and others [9]. For tackling the asymmetric distribution problem, AB-SVM under-samples the negative example set in order to train each SVM classifier on a balanced number of positive and negative examples, and then combines them using the bagging technique. AB-SVM belongs to the family of traditional supervised ensemble learning, and thus it cannot work well with very limited training data. According to [9], AB-SVM requires the user to label 40 images in each round of RF. Generally, the user can hardly accept this heavy labeling burden. In contrast, the proposed solution uses a very simple mechanism, termed "bias-weighting," to attack the asymmetry between the positive and negative examples. This can work well with only 10 images labeled by the user in each round of feedback.

V. Conclusion

In this paper, we proposed a novel RF scheme that integrates the merits of semi-supervised learning, ensemble learning, and active learning to address the small example problem. In particular, a bias-weighting strategy is used in our framework to address the asymmetric distribution problem. The empirical results showed the advantages of the proposed solution compared to some existing methods.

In future work, we will study more efficient solutions to reduce the redundancy among the informative examples by using the clustering technique.

References

[1] X. Zhou and T.S. Huang, "Relevance Feedback in Image Retrieval: A Comprehensive Review," *Multimedia Syst.*, vol. 8, no. 6, 2003, pp. 536-544.

[2] Y. Liu et al., "A Survey of Content-Based Image Retrieval with High-Level Semantics," *Pattern Recog.*, vol. 40, no. 1, 2007, pp. 262-282.

[3] L. Zhang, F. Lin, and B. Zhang, "Support Vector Machine Learning for Image Retrieval," *Proc. IEEE ICIP*, 2001, pp. 721-724.

[4] D.H. Kim et al., "Support Vector Machine Learning for Region-Based Image Retrieval with Relevance Feedback," *ETRI J.*, vol. 29, no. 5, 2007, pp. 700-702.

[5] S. Tong and E. Chang, "Support Vector Machine Active Learning for Image Retrieval," *Proc. ACM Multimedia*, 2001, pp. 107-118.

[6] W. Jiang, G. Er, and Q. Dai, "Boost SVM Active Learning for Content-Based Image Retrieval," *Proc. Asilomar Conf. Signals, Syst., Comput.*, 2003, pp. 1585-1589.

[7] L. Wang, K.L. Chan, and Z. Zhang, "Bootstrapping SVM Active Learning by Incorporating Unlabelled Images for Image Retrieval," *Proc. IEEE, CVPR*, vol. 1, 2003, pp. 629-634.

[8] S.C.H. Hoi et al., "Semi-supervised SVM Batch Mode Active Learning for Image Retrieval," *Proc. IEEE CVPR*, 2008, pp. 1-7.

[9] D.C. Tao et al., "Asymmetric Bagging and Random Subspace for Support Vector Machines-Based Relevance Feedback in Image Retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, 2006, pp. 1088-1099.

[10] T.S. Huang et al., “Active Learning for Interactive Multimedia Retrieval,” *Proc. IEEE*, vol. 96, no. 4, 2008, pp. 648-667.

[11] V.N. Vapnik, *Statistical Learning Theory*, Wiley, 1998.

[12] Z.H. Zhou, “Ensemble Learning,” *Encyclopedia of Biometrics*, Springer, 2009, pp. 116-123.

[13] O. Chapelle, B. Scholkipf, and A. Zien, *Semi-supervised Learning*, MIT Press, 2006.

[14] Z.H. Zhou, “When Semi-supervised Learning Meets Ensemble Learning,” *Proc. Int. Workshop Multiple Classifier System*, *LNCS 5519*, vol. 1, 2009, pp. 529-538.

[15] P.K. Mallapragada et al., “SemiBoost: Boosting for Semi-Supervised Learning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, 2009, pp. 2000-2014.

[16] L. Zheng et al., “Information Theoretic Regularization for Semi-Supervised Boosting,” *Proc. ACM SIGKDD*, 2009, pp. 1017-1026.

[17] Z.H. Zhou and M. Li, “Tri-Training: Exploiting Unlabeled Data Using Three Classifiers,” *IEEE Trans. Knowledge and Data Engineering*, vol. 17, no. 11, 2005, pp. 1529-1541.

[18] M. Li and Z.H. Zhou, “Classifier Ensemble with Unlabeled Data,” *CORR,* vol. abs/0909.3593, 2009.

[19] K. Bennett, A. Demiriz, and R. Maclin, “Exploiting Unlabeled Data in Ensemble Methods,” *Proc. ACM SIGKDD*, 2002, pp. 289-296.

[20] F. d’Alche-Buc, Y. Grandvalet, and C. Ambroise, “Semi-Supervised MarginBoost,” *NIPS*, vol. 14, 2002, pp. 553-560.

[21] N.V. Chawla, N. Japkowicz, and A. Kolcz, “Editorial: Special Issue on Learning from Imbalanced Data Sets,” *Proc. ACM SIGKDD Explorations*, vol. 6, no. 1, 2004, pp. 1-6.

[22] P. Viola, and M. Jones, “Fast and Robust Classification Using Asymmetric AdaBoost and a Detector Cascade,” *NIPS*, vol. 14, 2002, pp. 1311-1318.

[23] X.Y. Liu, J.X. Wu, and Z.H. Zhou, “Exploratory Undersampling for Class-Imbalance Learning,” *IEEE Trans. Syst., Man., Cybern. B, Cybern.*, vol. 39, no. 2, 2009, pp. 539-550.

[24] C. Chang, and C. Lin, “LIBSVM: A Library for Support Vector Machines,” 2001. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[25] D.P. Huijsmans and N. Sebe, “How to Complete Performance Graphs in Content-Based Image Retrieval: Add Generality and Normalize Scope,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, 2005, pp. 245-251.

Jun Wu received his BS from the North University for Ethnics, Yinchuan, China, in 2004, and his MS from Liaoning Normal University, Dalian, China, in 2007. Currently, he is a PhD candidate in the School of Information Science and Technology, Dalian Maritime University, Dalian, China. His research focuses on machine learning and data mining for multimedia retrieval and information security.

Ming-Yu Lu received his BS from Heilongjiang University, Harbin, China, in 1985, and his MS and PhD from Tsinghua University, Beijing, China, in 1988 and 2002, respectively. In 2005, he joined the Faculty of Dalian Maritime University, Dalian, China, where he is presently a professor in the School of Information Science and Technology. He is now the Director of the Intelligent Technology Research Center (ITReC). His research interests include data mining and multimedia content analysis.