2012 IEEE International Conference on Multimedia and Expo

Improving Relevance Feedback for Image Retrieval with Asymmetric Sampling

Biao Niu, Jian Cheng, Hanqing Lu
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
Beijing, China
{*bniu, jcheng, luhq*}*@nlpr.ia.ac.cn*

***Abstract*—Relevance feedback is a quite effective approach to improve performance for image retrieval. Recently, active learning method has attracted much attention due to its capability of alleviating the burden of labeling in relevance feedback. However, most of the traditional studies focus on single sample selection in each feedback which needs heavy computational cost in practice. In this paper, we presents a novel batch mode active learning method for informative sample selection. Inspired by graph propagation, we consider the certainty of labels as asymmetric propagation information on graph, and formulate the correlation between labeled samples and unlabeled samples in an united scheme. Extensive experiments on publicly available data sets show that the proposed method is promising.**

***Keywords*-image retrieval; active learning; selective sampling; semi-supervised learning;**

I. Introduction

The problem of retrieving images has received increasing attention. Content-based image retrieval [1], [2], which facilitates users to access the explosive increasing of the image data, becomes an increasingly active field. The shortage of training examples always remains a challenge for the supervised learning algorithms, especially within the context of semantic gap between low-level features and high-level semantic interpretation. In the past decade, relevance feedback [3] has been proved to be a useful approach to bridge the semantic gap by human-computer interaction. To alleviate the human burden for labeling in the loop of retrieval, active learning (or more precisely, selective sampling) [4] has been introduced which is proven to be effective when unlabeled data are abundant. Most current active learning methods [5]–[7] in image retrieval, typically select the most informative samples for user to label to improve the performance of the classifier greatly. Most active learning methods are selecting only the most informative sample for manually labeling in one iteration which is inefficient since updating classifier usually occupies numerous time in feedback.

To save the training time, several batch mode active learning methods have been proposed [8]–[11]. The key of batch mode active learning is to ensure little redundancy among the selected examples but enough representativeness for the data distribution. For instance, in *uncertainty* sampling framework, the least certain instances have been selected for labeling and *density* criteria aims to select the instances from the dense unlabeled regions and *diversity* criterion aims to select the instances among which the overlap in information content is least. Somewhat surprisingly, the correlation between labeled instances and unlabeled instances during the process has not been fully-exploited for selecting the optimal batch instances. They consider part of the criteria *uncertainty*, *diversity* and *density* and try to properly assemble their optimal query batch in different ways. Instead, there are often only one or two criteria introduced above considered for samples selection, which could partly limit the performance of the active learning in batch mode. Methods selecting most uncertainty instances usually ignore the distribution of the unlabeled instances and lead to serious sample bias. Methods considering the diversity may select outlier and methods only requesting measurement of density usually need selecting a relatively large number of instances before the optimal classifier is found.

In this paper, we present a novel batch mode active learning based on asymmetric propagation by modeling the criteria *uncertainty, diversity,* and *density* with the selection scheme. To this end, we build asymmetric propagation mechanism to unify a formulation explicitly around a scheme of degree of certainty which indicates our degree of "confidence" that the label is 1. In each subprocedure we update the rest unlabeled data's degree of certainty during a round degree of certainty propagation then we select the next most informative sample for soliciting. We repeat this subprocedure several times as one iteration and add these just labeled examples to the training set to retrain the classifier model.

Our contributions are summarized as follows:

- We incorporate the criteria: *uncertainty, diversity,* and *density* to unify a batch mode active learning formulation;
- A scheme of *asymmetric propagation* has been proposed to model these criteria.

Extensive experiments conducted on the real-world data sets show that the proposed method achieve encouraging results.

The rest of the paper is structured as follows: the related work is presented in Section II; Section III introduces the proposed batch mode active learning method; we present the

978-0-7695-4711-4/12 $26.00 © 2012 IEEE
DOI 10.1109/ICME.2012.127

experimental results in Section IV; finally, we conclude this paper in Section V.

II. Related work

In the scenario "pool-based" active learning a single informative sample is typically selected for manually labeling in each iteration. Once the sample has been added to the training set, the model retrains immediately and selects another single sample. For example [5] regards the task of learning concept of users' queries as learning SVM binary classifiers. An SVM classifier capture the query concept by separating the relevant images from irrelevant images with a hyperplane in a projected space. The projected points on one side of the hyperplane are considered relevant to the query concept and the rest is irrelevant. They learn an SVM classifier on the current labeled data and choose the next instance for querying that comes closest to the hyperplane.

Active learning methods in batch mode have been increasingly introduced recently. In each round of relevance feedback, [8] proposes an approach for SVM that explicitly incorporates a diversity measure that considers the angles between the induced hyperplane. Finally, in order to combine both requirements, viz. minimal distance to the classifier and diversity of these angles, they build the convex combination of both measure. Active learning incorporating the information density framework presented by [10] is proposed as a density-weighting technique. Then the Fisher information [9] avoid such traps implicitly, by utilizing the unlabeled pool U when estimating ratios. In [6] they propose active learning approach by querying informative and representative examples. They provide a systematic way for measuring and combining the informativeness and the representativeness. Their method is based on the min-max view [12] of active learning. In most setting these batch mode active learning are more efficient than the mode in which only a single sample is selected in one iteration.

III. Algorithm

In this section we present our batch mode active learning scheme in detail: in each iteration, we select and manually label the data one by one, using the last labeled data point and the distribution of the input dataset to decide the selection of the next; after a batch of data has been labeled and added to the training set, we perform training once to get a better performing classifier. The brief architecture of our proposed batch mode active learning during a single iteration is shown in Fig. 1.

A. Analysis uncertainty and approximation for retraining

In order to obtain the conditional probability estimates of assigning a class label y to the example we follow the initializing mode of [13]. Suppose there is an SVM classifier f trained on the given labeled data. Assume $X = \{x_1, x_2, \ldots, x_n\} = U \cup L$ to denote the input features, where U and L denote the unlabeled and labeled dataset respectively. For each data point $x_i \in X (1 \leqslant i \leqslant n)$, $y_i \in \{0, 1\}$ is its corresponding class label for negative and positive respectively. If $x_i \in U$, y_i is unknown and we use the sign of $f(x_i)$ to estimate y_i, for example if $f(x_i) < 0$ then we regard the label of x_i as 0, where f denotes the current trained classifier on L. We use $f(x_i)$ to represent the distance with sign from an unlabeled data instance x_i to the current decision boundary of SVM. We employ a Sigmoid function to normalize the distance metric into probability label metric within $[0, 1]$,as shown in Eq. 1.:

$$p_i = p(y_i = 1 | x_i, L) = \frac{1}{1+\exp(-f(x_i))} \tag{1}$$

We use this probability to denote an unlabeled example's initial degree of certainty in each iteration.

Inspired by graph propagation, we simply formulate the concept of certainty propagation as follows:

$$p_u^{+(x_k, y_k)} = p_u + (y_k - p_k) w_{ku} \tag{2}$$

where p_k and p_u denote degree of certainty of the last labeled sample and the unlabeled sample respectively, y_k is the true label of x_k. w_{ku} denotes the similarity between unlabeled data x_u and the last selected sample x_k. We define it as follows: We assume a connected graph $G = (V, E)$ to depict the correlation of the data, where the node set V is the input X. The edges E are represented by an $n \times n$ weight matrix W which is given. For example W can represent the pairwise relationship between data points with the radial basis function (RBF):

$$w_{ij} = \exp(-\frac{\|x_i - x_j\|^2}{\alpha}) \tag{3}$$

It is obvious that nearby points in Euclidean space are assigned large edge weights with the same α. Then we update all the rest unlabeled data's degree of certainty according to their correlation to the last selected data point. This procedure will be repeated until a certain number of samples have been labeled. We will find that after simplifying the formulation there is more flexible to implement the sample selection algorithm.

B. Selection scheme

We implement the sample selection algorithm based on our proposed method *asymmetric propagation* by controlling the magnitude of α. We show how to define the parameter α by incorporating the criteria *diversity* and *density* in detail as follows.

1) Uncertainty: This sampling criterion aims at selecting the unlabeled samples that can add most information to the current model. Generally, the most "uncertain" sample in the classification process is selected, such as the sample closest to the hyperplane in SVM [5]. Under probability label metric we infer the same result. The degree of certainty closer to 0.5 an unlabeled data point has, the more uncertainty it is.

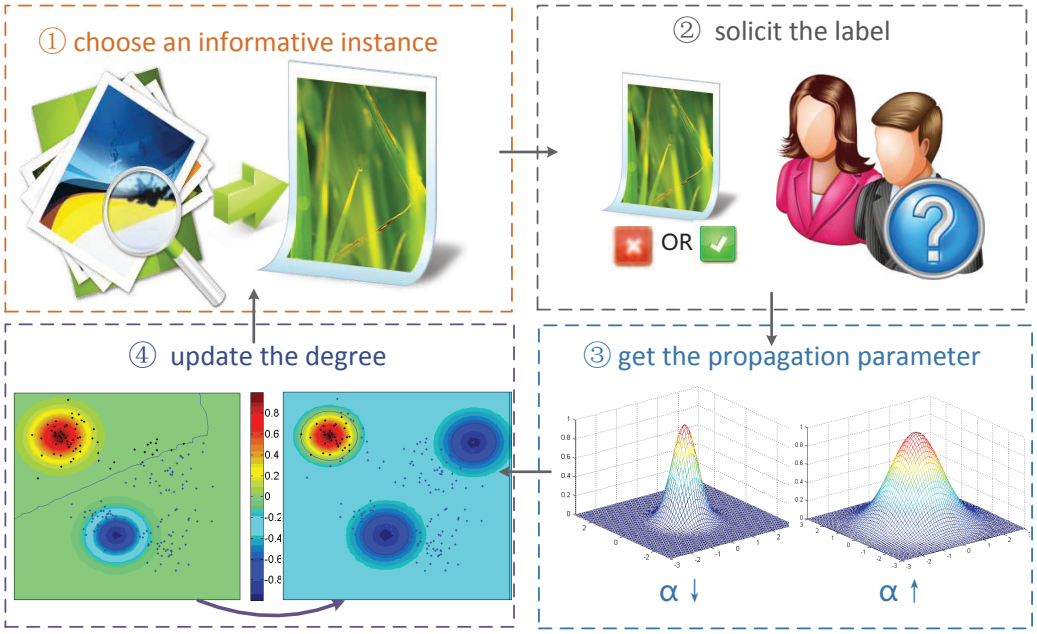

Figure 1. The brief architecture of our method. **(1):** select an instance whose degree of certainty is closest to 0.5; **(2):** query the label of the selected instance; **(3):** accord the label of the instance to obtain the propagation parameter α refereing to Eq. 7; **(4):** update the degree of certainty of rest instances according to Eq. 2.

Every time we only select one single data point with the degree of certainty p_i closest to 0.5 to label.

2) Density and Diversity: There is one parameter, α, in our scheme according to Eq. 3, which controls the influential radius of the correlated data points and reflects our batch model sample selection criteria. In order to ensure the generalization of our new scheme, we do not fix the parameter α empirically. Instead, we evaluate them adaptively according to the distribution of the input itself. According to Kernel Density Estimation(KDE) [14], we estimate the distribution of the input and adapt the parameter α on the basis of the density of the last selected sample. The probability density function $\hat{p}(x)$ can be estimated by

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} K(x, x_i) \tag{4}$$

where $K(x, x_i) = \exp(-\frac{\|x_i - x\|^2}{\beta})$ is a kernel function. β is the parameter of RBF kernel which we set as large as in SVM training. Then the density measure of the selected example can be defined by normalizing to $[0, 1]$ as follows:

$$density(x_s) = \frac{\sum_{j=1}^{n} K(x_s, x_j)}{\max_{x_i} \sum_{j=1}^{n} K(x_i, x_j)} \tag{5}$$

Here x_s is the last selected sample. Observe that the term on the right-hand side is in proportion to $\sum_{j=1}^{n} K(x_s, x_j)$. The kernel function is equivalent to the expression below:

$$\begin{aligned} \log(K(x, x_i)) &= -\frac{\|x_i - x\|^2}{\beta} \\ &\propto \|x_i - x\|^2 \\ &= \|x_i\|^2 + \|x\|^2 - 2\|x_i\|\|x\|\cos(\theta) \end{aligned} \tag{6}$$

where θ is the angle between x_i and x; this follows from the definition of inner product. And Eq. 6 can be used to estimate the diversity measure [8].

Density is a large mutual distance for points in the sample set in which we estimate the density [15]. We only have to compute the density of the instance and then we get the diversity measure of the subset simultaneously. So here the information of *density* include the measure of *diversity*.

3) Asymmetric propagation: The distribution of samples is usually in a very broad domain. We should to model the distribution of examples and learn the classifier with as few samples as possible. Thus the selected samples should be diversity as much as possible. Finally according this idea we can define the parameter α incorporating with measurement of *density*, as shown in Eq. 7:

$$\alpha = \frac{c^{|\ell(x_s) - \hat{\ell}(x_s)|}}{density(x_s)} \tag{7}$$

Here c is a constant positive coefficient which is less than one. $\ell(x_s) \in \{0, 1\}$ is the true label of the last selected sample which is queried by the user. $\hat{\ell}(x_s) \in \{0, 1\}$ is

the predicted label of the last selected sample. When the label is predicted wrong by the classifier in other words $\hat{\ell}(x_s) \neq \ell(x_s)$, then $|\ell(x_s) - \hat{\ell}(x_s)| = 1$. Then Eq. 7 becomes $\alpha = c/density(x_s)$. This means when a sample is classified wrong by the current classifier, we can decrease the parameter α to decrease the propagation radius because c is less than one and tend to select the next sample near the last labeled sample which is predicted wrong. On the other hand, when the label is predicted right by the classifier, in other words $\hat{\ell}(x_s) = \ell(x_s)$, then $|\ell(x_s) - \hat{\ell}(x_s)| = 0$. Then Eq. 7 becomes $\alpha = 1/density(x_s)$. This means when a sample is classified right by the current classifier, we can increase the parameter α to increase the propagation radius and choose the next unlabeled sample far from the last selected one.

IV. Experiment

In this section, we carry out all the experiments on Handwritten Digits Categorization and Content Based Image Retrieval and we compare our proposed batch mode active learning approach with other state-of-the-art methods to validate the effectiveness of the proposed algorithm in improvement of performance. In the following section, we describe the datasets and experimental setup, then we provide results.

A. Dataset

We conduct experiments on the standard dataset: the USPS dataset and the widely used benchmark Corel dataset for Content Based Image Retrieval.

The USPS dataset which contains grayscale handwritten digit images is scanned from envelopes by the U.S. Postal Service. There are 10 categories of "0" through "9". Each category contains 1100 images, 11000 images in all. These images are of size 16×16, with pixel values of 8-bit grayscale. We directly employed pixel values to represent these images. In total, a 256-dimensional vector is used to represent each image.

Corel dataset is one of the most used datasets by many groups in the research area of image retrieval. In total, we use 10,200 images to form our testbed from 102 different image categories which are randomly selected from the Corel image CDs with different semantic meanings, such as tiger, antelope, butterfly, car, cat, dog, horse and lizard, etc. Each category in the collection contains 100 images exactly. In our experiments, the main purpose is to verify if the learning mechanisms of our proposed method are useful, so for each image in the dataset three kinds of features are extracted: 125-dimensional color histogram vector, 6-dimensional color moment vector in RGB and 20-dimensional texture feature vector. The texture features extracted using 3-level discrete wavelet transformation (DWT). The mean and variance averaging on each of 10 subbands are arranged. In total, a 151-dimensional feature vector was extracted from each image.

B. Experiment Setup

In our experiments, we randomly choose 10 images as query images from each category. So on the USPS dataset 100 images and on Corel dataset 1020 images in total are chosen. We query an image and return the top-k images which are most relative to simulate Content Based Image Retrieval procedure every time. The kernel width in the SVM classifier in our experiments is learnt by cross-validation approach. All of them are fixed identically. In our experiments the penalty parameter C of SVM is set to 100 (or $\lambda = 0.01$), and the number of initial labeled images and the batch size k are set with the same constant. The parameter c in Eq. 7 is set to 0.9 empirically.

C. Compared Schemes

To evaluate the effectiveness of our batch mode active learning algorithm, we compare our sample selection strategy to the following traditional algorithms:

- SVM Active Learning: the baseline method for the original SVM active learning algorithm that simply choose the samples closest to the current decision boundary [5], denoted by SVM_{al}.
- SVM Active Learning with Diversity: the baseline method for batch mode SVM active learning by incorporating diversity measure among selected samples [8], denoted by SVM_{al}^{div}.
- Batch Mode Active Learning for Kernel Logistic Regression: the state-of-the-art kernel version of batch mode active learning using the kernel logistic regression [9], denoted by KLR_{bmal}.
- Representative Sampling with Certainty Propagation: batch mode active learning method based on [6], [16], selecting the samples both representative and informative, denoted by $QUIRE$.

When both the image and the query example belong to the same category we judge the image is relevant to the query in our experiments. This is the traditional definition that has been widely used. In order to evaluate the performance from each round, we depict the top-N accuracy vs. scope curves of the five algorithms after several rounds.

D. Results on USPS dataset

Table. I and Table. II show the accuracy vs. scope curves after one and two rounds of relevance feedback respectively. The number of initial labeled images and the batch size k are set to 10, where scope $= x$ means the accuracy is calculated within top x returned images. We can see that most of the advanced batch mode active learning algorithms, KLR_{bmal}, $QUIRE$ and our proposed method have better performance than the baseline method SVM_{al}. SVM_{al}^{div} has the worst performance here. It shows that only considering diversity has negative improvement under such circumstances. KLR_{bmal} has the almost same performance

with our proposed method which has the best performance here.

Table I
THE COMPARISON OF THE FIVE ALGORITHM AT THE FIRST ITERATION WHEN THE INITIAL AND BATCH SIZE IS 10 ON USPS DATASET

model	10	20	30	50	100
SVM_{al}	0.992	0.970	0.932	0.890	0.826
SVM_{al}^{div}	0.992	0.963	0.907	0.846	0.775
KLR_{bmal}	0.999	**0.996**	0.989	0.972	0.934
$QUIRE$	0.999	0.978	0.932	0.885	0.828
$ours$	**1**	0.994	**0.990**	**0.975**	**0.948**

Table II
THE COMPARISON OF THE FIVE ALGORITHM AT THE SECOND ITERATION WHEN THE INITIAL AND BATCH SIZE IS 10 ON USPS DATASET

model	10	20	30	50	100
SVM_{al}	1	0.984	0.972	0.938	0.881
SVM_{al}^{div}	1	0.982	0.955	0.921	0.870
KLR_{bmal}	0.999	0.998	0.995	0.986	0.959
$QUIRE$	1	0.999	0.994	0.983	0.960
$ours$	**1**	**0.999**	**0.997**	**0.994**	**0.985**

We also show the performance of different algorithms under the different initial labeled images and the batch size in Table. III. It shows at the first iteration the average precision of top-20 returned images under the number of initial labeled images and the batch size are set to 10, 15, 20 respectively. From Table. III, we can see that the proposed algorithm can achieve comparable performance to the state-of-the-art algorithms. With batch size of 15 and 20, our algorithm is superior to the other algorithms.

Table III
MAP@20 WITH DIFFERENT INITIAL AND BATCH SIZE LABELED IMAGES AT THE FIRST ITERATION ON USPS DATASET

model	10	15	20	MAP
SVM_{al}	0.970	0.984	0.993	0.982
SVM_{al}^{div}	0.963	0.983	0.993	0.980
KLR_{bmal}	**0.996**	0.998	**1**	**0.998**
$QUIRE$	0.978	0.996	0.999	0.991
$ours$	0.994	**0.999**	**1**	**0.998**

E. Results on Corel dataset

Fig. 2 and Fig. 3 show the accuracy vs. scope curves after 3 and 4 rounds of relevance feedback and the number of initial labeled images and the batch size k are set to 20. Several observations can be drawn from the experiment result in the figures. First of all, on the fourth iteration the performance is significantly better than on the third iteration. This observation matches our intuition that with more training samples the classifier will have better performance. We can see that most of the advanced batch mode active learning algorithms, SVM_{al}^{div}, $QUIRE$ and our proposed method have better performance than the baseline method SVM_{al} on the third iteration in Fig. 2. Here KLR_{bmal} is slightly

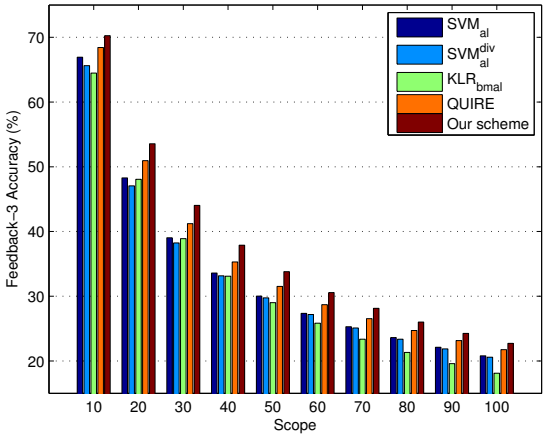

Figure 2. The accuracy of image retrieval with 3 feedbacks

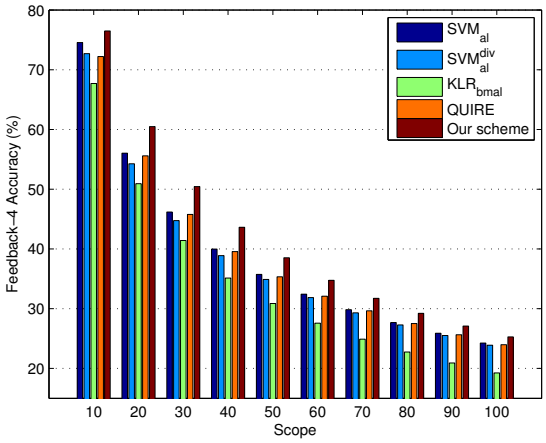

Figure 3. The accuracy of image retrieval with 4 feedbacks

better than SVM_{al}, but their difference is smaller for the top 40 ranked result. As shown in Fig. 3 KLR_{bmal} fails to improve the baseline method SVM_{al}. The fixed number of top eigens and the limited number of labeled samples restrict the performance of the method on the later iteration although KLR_{bmal}'s performance on the first iteration is the best. Second, when the number of iteration increases the difference between the baseline method SVM active learning and the rest advanced active learning methods decreases especially comparing the methods SVM_{al} and $QUIRE$. We found that their performances are similar on the fourth iteration in Fig. 3. Finally, in both figures we see that our proposed method achieve the best performances comparing to other algorithms.

To validate the effectiveness of the proposed algorithm, we also show the result of five algorithms under the different initial labeled images and the batch size in Table. IV. It

shows at the fourth iteration the average precision of top-20 returned images when the number of initial labeled images and the batch size are set to 10, 15, 20 respectively. From Table. IV, we can see that the proposed algorithm has the best performance here.

Table IV
MAP@20 WITH DIFFERENT INITIAL AND BATCH SIZE LABELED IMAGES AT THE FOURTH ITERATION ON COREL DATASET

Model	10	15	20	MAP
SVM_{al}	0.389	0.479	0.560	0.476
SVM_{al}^{div}	0.382	0.470	0.543	0.465
KLR_{bmal}	0.334	0.500	0.555	0.463
$QUIRE$	0.432	0.476	0.509	0.472
$ours$	**0.446**	**0.536**	**0.604**	**0.528**

V. CONCLUSION

This paper presents a novel active learning algorithm for selective sampling in relevance feedback, Asymmetric Propagation based Batch Mode Active Learning. Unlike the traditional batch mode active learning methods, which only considers the correlation among the labeled samples [8], or only estimate the distribution of the unlabeled data [10], our proposed method takes both of them into consideration explicitly. Based on the degree of certainty asymmetric propagation scheme, the proposed approach provides a new way to incorporate the criteria: *uncertainty, diversity* and *density* simultaneously. Experimental results show that after adopting our novel scheme in the selective sampling, the classification performance can be obviously improved on real-world dataset.

ACKNOWLEDGMENT

This work is supported by 973 program (Grant No. 2010CB327905), National Natural Science Foundation of China (Grant No. 61170127, 60975010, 60833006, 61070104).

REFERENCES

[1] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Analysis and Machine Intelligent*, vol. 22, pp. 1349–1380, December 2000.

[2] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Computing, Communications, and Applications*, vol. 2, pp. 1–19, February 2006.

[3] Yong Rui, T.S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool for interactive content-based image retrieval," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 644–655, September 1998.

[4] Meng Wang and Xian-Sheng Hua, "Active learning in multimedia annotation and retrieval: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 10:1–10:21, February 2011.

[5] Simon Tong and Edward Chang, "Support vector machine active learning for image retrieval," in *Proceedings of the 9th ACM International Conference on Multimedia*, 2001, pp. 107–118.

[6] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou, "Active learning by querying informative and representative examples," in *Advances in Neural Information Processing Systems 24*, Vancouver, Canada, 2010, pp. 1–9.

[7] Jian Cheng and Kongqiao Wang, "Active learning for image retrieval with co-svm," *Pattern Recognition*, vol. 40, no. 1, pp. 330–334, 2007.

[8] Klaus Brinker, "Incorporating diversity in active learning with support vector machines," in *Proceedings of the 20th International Conference on Machine Learning*. 2003, pp. 59–66, AAAI Press.

[9] Steven C.H. Hoi, Rong Jin, and Michael R. Lyu, "Batch mode active learning and its applications to text categorization and image retrieval," *IEEE Trans. Knowledge and Data Engineering (TKDE)*, vol. 21, no. 9, pp. 1233–1248, September 2009.

[10] Burr Settles and Mark Craven, "An analysis of active learning strategies for sequence labeling tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, 2008, pp. 1070–1079, Association for Computational Linguistics.

[11] Xiaoyu Zhang, Jian Cheng, Hanqing Lu, and Songde Ma, "Selective sampling based on dynamic certainty propagation for image retrieval," in *Advances in Multimedia Modeling*, vol. 4903 of *Lecture Notes in Computer Science*, pp. 425–435. Springer Berlin / Heidelberg, 2008.

[12] Steven C.H. Hoi, Rong Jin, Jianke Zhu, and Michael R. Lyu, "Semi-supervised svm batch mode active learning for image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition, 2008.*, June 2008, pp. 1–7.

[13] Steven C.H. Hoi and Michael R. Lyu, "A semi-supervised active learning framework for image retrieval," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2005, vol. 2, pp. 302–309.

[14] Emanuel Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, September 1962.

[15] Charlie Dagli, Shyamsundar Rajaram, and Thomas Huang, "Leveraging active learning for relevance feedback using an information theoretic diversity measure," in *Image and Video Retrieval*, vol. 4071 of *Lecture Notes in Computer Science*, pp. 123–132. Springer Berlin / Heidelberg, 2006.

[16] Jian Cheng, Biao Niu, Yikai Fang, and Hanqing Lu, "Representative sampling with certainty propagation for image retrieval," in *Proceedings of the 18th International Conference on Image Processing*, 2011, pp. 2541–2544.