Leveraging Active Learning for Relevance Feedback Using an Information Theoretic Diversity Measure

Charlie K. Dagli, Shyamsundar Rajaram, and Thomas S. Huang

Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign
Urbana, IL 61801
{dagli, rajaram1, huang}@ifp.uiuc.edu

Abstract. Interactively learning from a small sample of unlabeled examples is an enormously challenging task. Relevance feedback and more recently active learning are two standard techniques that have received much attention towards solving this interactive learning problem. How to best utilize the user's effort for labeling, however, remains unanswered. It has been shown in the past that labeling a diverse set of points is helpful, however, the notion of diversity has either been dependent on the learner used, or computationally expensive. In this paper, we intend to address these issues by proposing a fundamentally motivated, information-theoretic view of diversity and its use in a fast, non-degenerate active learning-based relevance feedback setting. Comparative testing and results are reported and thoughts for future work are presented.

1 Introduction

An enormous challenge in interactive image and video retrieval is correlating the user-dependent interpretation of image-content with low-level visual descriptors, closing the so-called *semantic gap*. Relevance feedback has garnered much attention in the past decade in attempting to reach this goal [1][2][3][4]. At its heart, relevance feedback suffers from the *small sample learning* problem [5]. Rankers or classifiers must learn in high-dimensional feature spaces with only a handful of labeled training examples. Consequently, many potential discriminating observations go unlabeled. As an additional practical consideration, because these systems have a user in the loop, they must also be quick and robust to change. In recent years, attention has been given to systems that employ *active learning* to address these challenges.

Active learning is a paradigm that proposes ways to incrementally learn from unlabeled data, provided the system has available to it an *oracle*, an entity which knows the correct labeling of all examples [6][7]. Given an initial weak ranker or classifier, the oracle labels a set of points the systems deems to be most informative, the *pool query set*. The information provided from this labeling can then be used to update the system and this process can be repeated indefinitely to improve the accuracy of those points in the returned or *resultant set*. Traditional

H. Sundaram et al. (Eds.): CIVR 2006, LNCS 4071, pp. 123–132, 2006.

relevance feedback can be seen as a degenerate case of active learning as the set of top-k returned points serves both as the returned *and* pool query sets. Using a unique pool query set, however, has been shown to improve performance [8].

Whether we use the traditional or active learning-based paradigm, the user is often asked to label examples which are quite similar to one another, often times as a result of examples clustering in the same area of the feature space. In a small-sample setting, especially when the users and systems effort is at a premium, it makes more sense for the user to label a *diverse* set of points for each pool-query rather than many similar points which are, in comparison, much less informative. There have been a handful of techniques which have addressed this issue.

In the traditional relevance feedback scenario, NECs PicHunter [9] cast diversity for image retrieval as matter of *exploration* versus *exploitation.* Utilizing Bayesian relevance feedback techniques, they ask users to label images with low-posterior probability in addition to those with high probability. This notion of diversity relies on the probabilistic modeling for it's calculation, however, which tends to limit its general application. The CLUE system of [10] also realizes that images tend to be semantically clustered in the vicinity of query images. To this end, they propose a local-neighborhood based clustering approach to more efficiently present diverse information for labeling by the user. This clustering must occur for every round of feedback, however, and the computational load of doing so may in some cases outweigh the improved retrieval results.

From a purely active learning viewpoint, one of the first works to incorporate diversity sampling was [11] and subsequently [12] where the notion of angular diversity was investigated for support vector machines (SVMs). The idea of using angular diversity in particular, however, was motivated specifically by the version-space reduction requirements inherent in SVM active learning. It is not clear whether this specialized measure of diversity is suitable for general problems.

In this work, we motivate and introduce a more general notion of diversity based on information-theoretic concepts, and apply it to a fast, *non-degenerate* active learning scheme for relevance feedback based on query-point refinement which, to our knowledge, is a scenario that has not received much attention in the past.

The rest of the paper is organized as follows: Section 2 motivates our measure of diversity and leads into Section 3 which presents the active learning based algorithm built around it. Experimental results and thoughts for future study are presented in Sections 4 and 5 respectively.

2 Information-Theoretic Diversity

To motivate the discussion of our active learning framework, we will first define a basic diversity measure, based on Shannon's entropy [13]. It's intuitively attractive to associate high entropy with diversity, as entropy is essentially a measure of randomness. For any continuous random variable, $\mathbf{X}$, which takes a particular value $\mathbf{x}$, entropy is defined as

$$h(\mathbf{x}) = -\int p(\mathbf{x}) \log\left(p(\mathbf{x})\right) = -E\left[\log\left(p(\mathbf{x})\right)\right] \tag{1}$$

where $E[\cdot]$ is the expectation (mean) operator. Calculating entropy in practice involves density estimation, as the underlying probabilities are usually not known. Given a set of points,$\{\mathbf{x}_i\}_{i=1}^N$, we can approximate the expectation in (1) by the sample mean so that

$$h(\mathbf{x}) \approx -\frac{1}{N}\sum_{i=1}^{N} \log\left(p(\mathbf{x}_i)\right) \tag{2}$$

which by the Law of Large Numbers approaches the actual mean in the limit as $N \to \infty$. We are still left with the problem of estimating the distribution, $p(x)$. Through Parzen density estimation

$$p(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{K}(\mathbf{x}, \mathbf{x}_i) \tag{3}$$

where $\mathcal{K}$ is the Parzen window.

Substituting this estimate of the density into (2), we now define a new quantity called *empirical entropy*

$$h_e(\{\mathbf{x}_i\}_{i=1}^N) = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{1}{N}\sum_{j=1}^{N} \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)\right) \tag{4}$$

Expanding this equation, we arrive at the final expression for empirical entropy:

$$h_e(\{\mathbf{x}_i\}_{i=1}^N) = \frac{1}{N}\log(N) - \frac{1}{N}\sum_{i=1}^{N} \log\left(\sum_{j=1}^{N} \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)\right) \tag{5}$$

For all testing and experimentation, we used the Gaussian radial basis kernel

$$\mathcal{K}(\mathbf{x}, \mathbf{x}_i) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{x}_i)^T \Sigma^{-1}(\mathbf{x}-\mathbf{x}_i)\right)}{(2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}} \tag{6}$$

with isotropic covariance matrix $\Sigma = \sigma\mathbf{I}$.

Before utilizing empirical entropy in an active learning algorithm, it is worth examining this quantity in a bit more detail. To gain better insight into entropy as a general measure of diversity, we present the following theorem.

Theorem 1. *If $h_e(\{\mathbf{x}_i\}_{i=1}^N)$ denotes the empirical entropy of $\{\mathbf{x}_i\}_{i=1}^N$, then a large upper bound on this value corresponds to large distances and angles between mutual paris of points in the sample set $\{\mathbf{x}_i\}_{i=1}^N$.*

Proof. By definition of empirical entropy in Eqn. 4,

$$h_e(\{\mathbf{x}_i\}_{i=1}^N) = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{1}{N}\sum_{j=1}^{N}\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j)\right)$$

The logarithm function is concave and applying Jensen's Inequality, we obtain

$$h_e(\{\mathbf{x}_i\}_{i=1}^N) \leq -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\log\left(\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j)\right)$$

Assuming $\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j)$ is a Gaussian Kernel as in Equation (6) we obtain,

$$h_e(\{\mathbf{x}_i\}_{i=1}^N) \leq \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(d\log(2\pi) + \log|\Sigma| + (\mathbf{x}_i-\mathbf{x}_j)^T\Sigma^{-1}(\mathbf{x}_i-\mathbf{x}_j)$$

Observe that the third term on the right-hand side in the above equation is equivalent to the canonical inner product of

$$\langle(\mathbf{x}_i^{'},\mathbf{x}_j^{'})\rangle = (\mathbf{x}_i-\mathbf{x}_j)^T\Sigma^{-1}(\mathbf{x}_i-\mathbf{x}_j) = M(\mathbf{x}_i,\mathbf{x}_j)$$

where $\mathbf{x}_i^{'} = \Sigma^{-1/2}\mathbf{x}_i$, $\mathbf{x}_j^{'} = \Sigma^{-1/2}\mathbf{x}_j$ and the (non-unique) existence of $\Sigma^{-1/2}$ follows from the symmetric, positive semi-definite property of covariance matrix Σ. In the new space induced by $\Sigma^{-1/2}$, the distance between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$\|(\mathbf{x}_i^{'}-\mathbf{x}_j^{'})\|^2 = \|\mathbf{x}'_i\|^2 + \|\mathbf{x}'_j\|^2 - 2\|\mathbf{x}_i^{'}\|\|\mathbf{x}_j^{'}\|\cos(\theta)$$

where θ is the angle between $\mathbf{x}_i^{'}$ and $\mathbf{x}_j^{'}$; this follows from the definition of inner product. This expression is largest when $\theta = \pi$, the largest angle between two vectors. □

The major implication of this analysis is that a large bound on empirical entropy depends on a large mutual distance for points in a sample set, which by the definition of the canonical inner product also implies large mutual angles. In this way, entropic diversity in a general setting is able to capture both these notions of uniqueness, and therefore, in the limit, diversity.

3 Algorithm for Pool-Query Selection

A general active learning algorithm chooses both a resultant and pool-query set to present to the user at each step. We assume that the algorithm narrows down the set of all unlabeled points at each step to a candidate pool-query set. In the case of SVM active learning, these are the unlabeled points which lie in the version space. In a query-point refinement algorithm, one can choose from a large number of points in the neighborhood of the query centroid. We don't want to arbitrarily choose the most diverse points, however. We must keep proper perspective and ensure that these points are still close to the query centroid. To this end, the cost function we minimize is a convex sum of both query point distance and the negative of entropic diversity.

At each round of feedback, then, we must choose from the set of M points (where $M >> k$) closest to the query centroid, $\mathbf{x}_c$, an N point subset known as the *pool-query* set for oracle labeling. Using brute force results in $\binom{M}{N}$ unique N point subsets for which we must compute our empirical entropy diversity measure. Even for moderate sample sizes, however, this number quickly becomes computationally intractable. Instead, we use an adaption of the greedy algorithm in [11] used for calculating angular diversity. Starting with the point closest to the query centroid, $\mathbf{x}_{min}$, at each step we add to the pool-query set PQ that point which most decreases the cost function C.

```
$\mathbf{x}_{\text{min}} = \mathbf{x}_{\underset{i}{\text{argmin}} \|\mathbf{x}_i - \mathbf{x}_c\|}$
$\text{PQ} = \{\mathbf{x}_{\text{min}}\}$
**do**
    $h_{max} = \max_i (\text{PQ} \cup \{\mathbf{x}_i - \mathbf{x}_c\})$
    $h_{min} = \min_i (\text{PQ} \cup \{\mathbf{x}_i - \mathbf{x}_c\})$
    $C(i) = \alpha||\mathbf{x}_i - \mathbf{x}_c|| + (1-\alpha)\left[-\frac{h_e(\text{PQ} \cup \{\mathbf{x}_i - \mathbf{x}_c\})}{h_{max} - h_{min}}\right]$
        $\forall i : i \notin \text{PQ}$
    $PQ \cup \{\mathbf{x}_{\underset{j \notin PQ}{\text{argmin}}} C(j)\}$
**while** $|PQ| \leq N$
```

The mixing parameter, α, allows us to scale up or down the influence of empirical entropy to the cost function. When $\alpha = 1$, the pool-query technique defaults to a nearest neighbour regime completely discounting any diversity information. When $\alpha = 0$, the cost function becomes purely an entropic diversity measure. We will explore the effects of mixing later.

3.1 Biased Discriminant Analysis

In the evaluation of our proposed diversity-framework, we chose to use a small-sample learner especially suited for information-retrieval problems, Biased Discriminant Analysis [5]. BDA casts the problem of relevance feedback from a two-class (positive and negative) to a one-to-many class (one positive, multiple negative) problem. The goal is to find a transformation of the feature space which closely clusters positive examples while pushing away negative ones.

The optimal transformation is obtained by maximizing the following objective function

$$\underset{\mathbf{W}}{\operatorname{argmax}} \left| \frac{\mathbf{W}^T \mathbf{S}_{PN} \mathbf{W}}{\mathbf{W}^T \mathbf{S}_P \mathbf{W}} \right| \tag{7}$$

where $\mathbf{S}_P$ is the intra-class-scatter matrix for all the training examples and $\mathbf{S}_{PN}$ is the inter-class scatter matrix between positive examples and negative training examples, treating each negative example as an individual class.

The optimal value of $\mathbf{W}$ is the solution to the generalized eigenvalue problem presented by the Rayleigh Coefficient in (7). The optimal transformation then becomes

$$\mathbf{A} = \mathbf{\Phi} \mathbf{\Lambda}^{1/2} \tag{8}$$

where $\mathbf{\Lambda}$ is the diagonal eigenvalue matrix and $\mathbf{\Phi}$ is the corresponding eigenvector matrix and $\mathbf{W} = \mathbf{A}\mathbf{A}^T$. The distance between two points in the new space can be computed using the standard Euclidean measure, or in the original space using the distance metric

$$\text{distance}(\mathbf{x} - \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \mathbf{A} (\mathbf{x} - \mathbf{y}) \tag{9}$$

Although BDA can also be used for feature reduction, in its full form it is essentially a query point refinement algorithm. Each round of user feedback yields a new transformation of the feature space which results in the centroids of both the positive and negative examples being moved.

4 Image Retrieval Experimentation

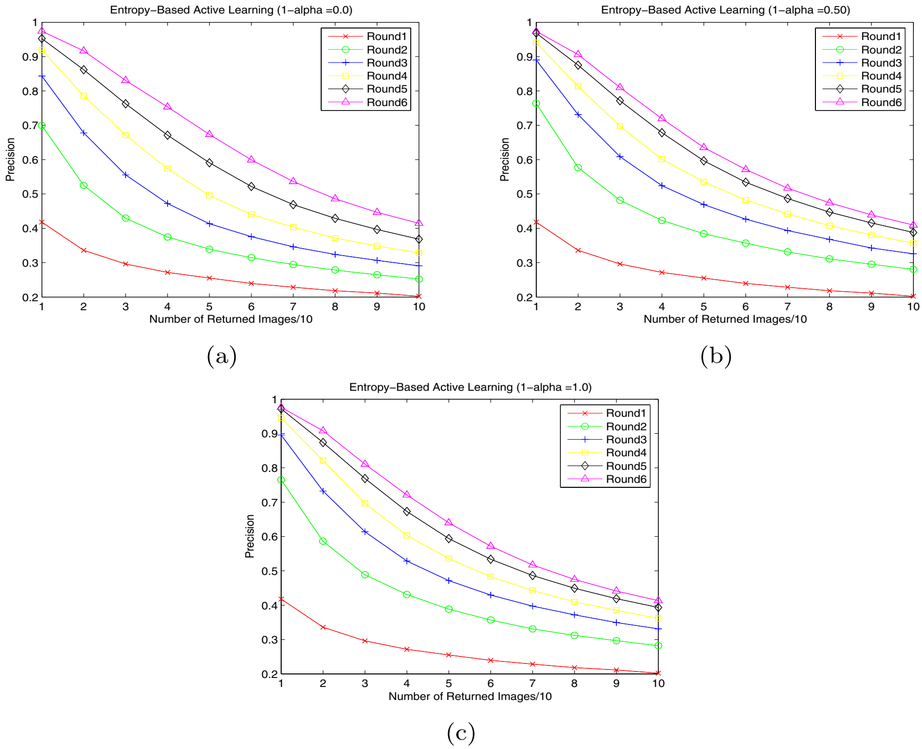
4.1 Image Features and Testing Procedure

To explore the practical performance of our entropy-based active learning system, we performed extensive tests using a 5000 image subset of the COREL image database. To appropriately model the small-sample scenario, only 1400 images were used for target sets. The target set consisted of 13 unique query concepts.

The first, second and third moments in each channel of the HSV color space, first and second wavelet sub-band moments at three levels of decomposition and a Waterfilling algorithm were used for color, texture and shape features respectively. In total, a 47-dimensional feature vector was extracted from each image.

Initially, the user (oracle) is presented with 20 randomly chosen images. After each round of retrieval, they mark which images they deem to be relevant. The remaining are assumed to be irrelevant. From this information, the systems adjusts its understanding of the query concept using BDA, and returns both the k most similar images and the pool-query set of images to label for active learning.

For each query class, we conducted 25 random feedback sessions of 6 rounds each. In total, there were 325 user-guided sessions, with 1950 total rounds of



(a)

(b)

(c)

Fig. 1. Effect of α across Feedback Round for (a) $\alpha = 1$, (b) $\alpha = 0.5$, (c) $\alpha = 0$

feedback for each method tested. The size of the pool-query was chosen to reflect a reasonable number of image a human operator could label at one time, 20. Results were averaged across all 325 tests.

4.2 Temporal Role of Diversity

It is reasonable to suspect that diversity is not equally important at every stage of the learning process. Intuition would suggest that at lower feedback rounds, the diversity strategy would be more important as the system has the smallest amount of knowledge and thus needs the most diverse set of labeled examples to learn its query concept. In higher rounds, diversity should be less important as the system should have utilized the information from prior diverse labelings to localize the area in the feature space were most of the relevant images are.

Investigating this intuitive idea, we performed tests using mixing parameter α values of 0, 0.25, 0.5, 0.75 and 1 corresponding to, at the extremes, pure diversity and pure distance (k nearest neighbors). Precision values for oracle-query sets of size 20 were calculated. Figure 1 shows the values of precision versus number of images returned for increasing rounds of feedback, marginalized across query class.

Table 1. Percentage Change in Precision between (a) Entropic Diversity Sampling versus Pure Distance (k-nearest neighbours), α-varying Entropic Diversity versus (b) Angular SVM and (c) Angular Diversity Sampling Techniques

	α			
R	0.75	0.5	0.25	0.0
1	0	0	0	0
2	11.84	11.74	12.00	12.59
3	10.28	11.22	13.10	12.27
4	7.34	6.90	8.21	7.46
5	2.97	2.69	3.42	2.83
6	-0.87	-2.76	-2.49	-2.42

(a)

	α			
R	0.75	0.5	0.25	0.0
1	100	100	100	100
2	42.42	42.32	42.59	43.17
3	29.66	30.60	32.47	31.65
4	21.81	21.37	22.68	21.93
5	15.16	14.89	15.61	15.03
6	9.65	7.77	8.04	8.11

(b)

	α			
R	0.75	0.5	0.25	0.0
1	0	0	0	0
2	3.61	3.22	1.73	0.63
3	2.82	2.46	2.84	0.91
4	2.05	1.78	2.63	0.54
5	2.40	1.50	2.46	1.17
6	2.61	0.31	0.00	0.62

(c)

Trivially, the first round of feedback yields the same precision for all values of α. As the round number increases, however, we can begin to resolve the performance of different values of the mixing parameter. Comparing the purely distance strategy ($\alpha = 1$, Figure 1 (a)) with the first introduction of diversity ($\alpha = 0.5$, Figure 1 (b)) we see marked improvement in precision for Rounds 2-3. (Round 1, initial labeling, is the same for all values.) This improvement increases as we decrease α and weight more toward diversity ($\alpha = 0$, Figure 1(c)). Conversely, the addition of diversity in the final rounds is not as beneficial as there is only slight improvement in Round 5 and a decrease in improvement when using diversity in Round 6. This agrees with our intuition that in higher rounds diversity is not as important.

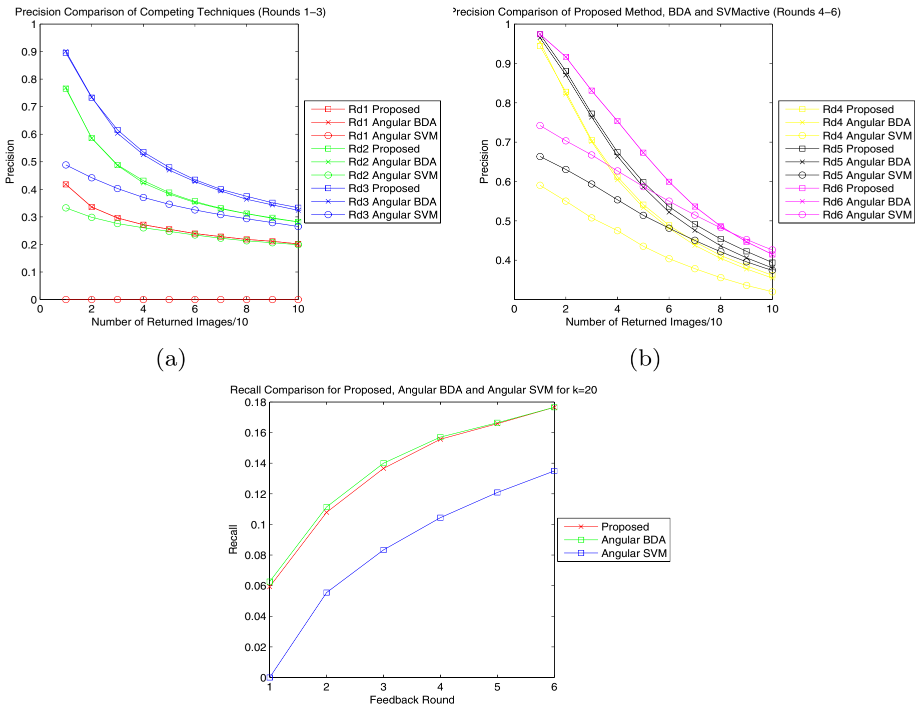
The exact percentage change in precision between $\alpha = 1.0$, pure distance, and differing degrees of diversity marginalized over different number of returned images can be seen in Table 1 (a).

We can now adjust our entropic diversity pool-query selection scheme to reflect the temporal effect of the mixing parameter. In Round 2, we set α= 0 because it yields the largest percent increase, 12.59%. Rounds 3-5 we tune diversity down by changing α to 0.25. Finally, in the last round, we move to a purely distance-based strategy, setting α= 1.

4.3 Comparing with Other Relevance Feedback Algorithms

To further investigate the performance of this α-varying, entropic diversity framework, testing with other active learning-based relevance feedback techniques was performed. Comparisons between the proposed system, angular diversity BDA of [14] and a basic version of the angular-diversity SVM framework of [8] and [12] were conducted. (Comparison with plain BDA was unnecessary as it was done already in Section 5.2.) As before, the same batch of 325 user-guided tests using pool-query size of 20 were done. The results of these tests can be summarized in Figure 2.

Looking at Figure2(a)-(b), we can seen that entropy-based diversity yields substantially better results than angular SVM and better results than purely



(a)

(b)

(c)

Fig. 2. Precision Comparison for Entropic Diversity BDA, Angular Diversity BDA and Angular SVM for (a) Rounds 1-3 and (b) Rounds 4-6, (c) Recall Comparison for all algorithms (k=20)

angular-based BDA. The recall curves of Figure2(c) also show similar behavior. (Since we are comparing the *algorithms* themselves, the first round of angular-based SVM is zero, since there are two rounds of labeling before the user begins to see results.) Tables 1(a) and 1(b) show specific quantitative results.

As the results show, the incorporation of an information-based diversity measure into a general active learning framework can, indeed, improve performance. Given a weak learner, BDA, we have seen marked improvement over the degenerate case (BDA with nearest neighbors) of at most 12-13% in the lower rounds. In addition, these empirical results also coincide with our notion that an entropic diversity measure should perform as well or better than angular diversity.

5 Summary

In this work, we have proposed a fundamentally motivated, information theoretic view of diversity and incorporated it into a non-degenerate, query-point refinement scheme for relevance feedback. Our results support entropic diversity's viability as a useful tool for pool-query selection in active learning. In the future, we plan to investigate the viability of these types of measures in more general active learning scenarios, as well as looking into these ideas for video mining and active ranking for collaborative filtering.

References

1. Y. Rui and T.S. Huang, "Relevance Feedback Techniques in Image Retrieval," in *Principles of Visual Information Retrieval*, M.S. Lew, Ed. London: Springer-Verlag, 2001.
2. B.C. Ko and H. Byun, Probabilistic Neural Networks Supporting Multi-Class Relevance Feedback in Region-Based Image Retrieval," in *International Conference on Pattern Recognition*, 2002.
3. A. Dong and B. Bhanu, "Active Concept Learning for Image Retrieval in Dynamic Databases," in *International Conference on Computer Vision*, 2003.
4. X.S. Zhou, Y. Rui and T.S. Huang, Exploration of Visual Data, Kluwer Academic Publishers, 2003.
5. X. Zhou and T.S. Huang, "Small Sample Learning during Multimedia Retrieval using BiasMap," in *IEEE Conference Computer Vision and Pattern Recognition*, 2001.
6. D.A. Cohn, Z. Ghahramani and M.I. Jordan, "Active Learning with Statistical Models", *Advances in Neural Information Processing Systems*, Vol. 7, G. Tesauro, D. Touretzky and J. Alspector, Eds. Cambridge: MIT Press, 1995.
7. N. Roy and A. McCallum, "Toward Optimal Active Learning through Monte Carlo Estimation of Error Reduction," in *International Conference on Machine Learning*, 2001.
8. S. Tong and E. Chang, "Support Vector Machine Active Learning for Image Retrieval," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2001.
9. I.J. Cox, T.P. Minka, T.V. Papathomas and P. N. Yianilos, "Pichunter: Bayesian Relevance Feedback for Image Retrieval," in *International Conference on Pattern Recognition*, 1996.
10. Y. Chen, J.Z. Wang and R. Krovetz "CLUE: Cluster-based Retrieval of Images by Unsupervised Learning," IEEE Transactions on Image Processing, Vol. 14, No. 8, pp. 1187-1201, 2005.
11. K. Brinker, "Incorporating Diversity in Active Learning with Support Vector Machines," in *International Conference on Machine Learning*, 2003.
12. K. Goh, E. Y. Chang, and W.-C. Lai, "Concept-dependent Multimodal Active Learning for Image Retrieval", in *ACM International Conference on Multimedia*, 2004.
13. C.E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, Vol. 27, pp. 379-423 and 623-656, 1948.
14. C.K. Dagli, S. Rajaram and T.S. Huang, "Combining Diversity-Based Active Learning with Discriminant Analysis in Image Retrieval," in *IEEE International Conference on Information Technology and Applications*, 2005.