Stretching Bayesian Learning in the Relevance Feedback of Image Retrieval

Ruofei Zhang and Zhongfei (Mark) Zhang

Department of Computer Science
State University of New York at Binghamton, Binghamton, NY 13902, USA
{rzhang,zhongfei}@cs.binghamton.edu

Abstract. This paper is about the work on user relevance feedback in image retrieval. We take this problem as a standard two-class pattern classification problem aiming at refining the retrieval precision by learning through the user relevance feedback data. However, we have investigated the problem by noting two important unique characteristics of the problem: small sample collection and asymmetric sample distributions between positive and negative samples. We have developed a novel approach to stretching Bayesian learning to solve for this problem by explicitly exploiting the two unique characteristics, which is the methodology of **BA**yesian **L**earning in **A**symmetric and **S**mall sample collections, thus called **BALAS**. Different learning strategies are used for positive and negative sample collections in **BALAS**, respectively, based on the two unique characteristics. By defining the relevancy confidence as the relevant posterior probability, we have developed an integrated ranking scheme in **BALAS** which complementarily combines the subjective relevancy confidence and the objective feature-based distance measure to capture the overall retrieval semantics. The experimental evaluations have confirmed the rationale of the proposed ranking scheme, and have also demonstrated that **BALAS** is superior to an existing relevance feedback method in the current literature in capturing the overall retrieval semantics.

1 Introduction

This paper is on Content-Based Image Retrieval (CBIR). Since 1990's, CBIR has attracted significant research attention [8]. Early research focused on finding the "best" representation for image features. The similarity between two images is typically determined by the distances of individual low-level features and the retrieval process is performed by a *k-nn* search in the feature space [1]. In this context, high level concepts and user's perception subjectivity cannot be well modelled. Recent approaches introduce human-computer interaction (HCI) into CBIR. The interaction mechanism allows a user to submit a coarse initial query and continuously refine his(her) searching via relevance feedback. This approach greatly reduces the labor required to precisely compose a query and easily captures the user's subjective retrieval preference.

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3023, pp. 355–367, 2004.

However, most approaches to relevance feedback were based on heuristic formulation of empirical parameter adjustment and/or feature component reweighing, which is typically *ad hoc* and not systematic, and thus cannot be substantiated well. Some of the recent work [16,14,4] investigated the problem from a more systematic point of view by formulating the relevance feedback problem as a general classification or learning problem and used optimization methods to address it. These learning methods are all based on the assumption that both positive and negative samples confirm either implicitly or explicitly a well formed distribution. We note that without further exploiting the unique characteristics of training samples in the relevance feedback of image retrieval, it is difficult to map the image retrieval problem to a general two-class (i.e., relevance vs. irrelevance) classification problem in realistic applications. Consequently before we design a specific relevance feedback methodology, two unique characteristics of the relevance feedback problem in image retrieval must be noted and addressed. The first is the small sample collection issue. In relevance feedback of image retrieval, the number of training samples is usually small (typically < 20 in each round of interaction) relative to the dimensionality of the feature space (from dozens to hundreds, or even more), whereas the number of image classes or categories is usually large for many real-world image databases. The second characteristic is the asymmetric training sample issue. Most classification or learning techniques proposed in the literature of pattern recognition and computer vision, such as discriminant analysis [6] and Support Vector Machine(SVM) [15], regard the positive and negative examples interchangeably and assume that both sets are distributed approximately equally. However, in relevance feedback, while it is reasonable to assume that all the positive samples confirm to a specific class distribution, it is typically not valid to make the same assumption for the negative samples, as there may be an arbitrary number of semantic classes for the negative samples to a given query; thus, the small, limited number of negative examples is unlikely to be representative for all the irrelevant classes, and this asymmetric characteristic must be taken into account in the relevance feedback learning.

In this paper, we investigate the relevance feedback problem in image retrieval using Bayesian learning. Specifically, we stretch Bayesian learning by explicitly exploiting the two unique characteristics through developing a novel user relevance feedback methodology in image retrieval — **BA**yesian **L**earning in **A**symmetric and **S**mall sample collections, called **BALAS**. In **BALAS**, we introduce specific strategies to estimate the probabilistic density functions for the positive and negative sample collections, respectively. It is shown that an optimal classification can be achieved when a scheme for measuring the relevancy confidence is developed to reflect the *subjective* relevancy degree of an image w.r.t. a query image. The relevancy confidence is integrated with the measure of feature-based distance, which reflects the *objective* proximity degree between feature vectors, to order the ranking of the retrieved images from a database.

2 BALAS Methodology

Given a query image, a "good" relevance feedback method would, after learning, allow as many as relevant images to be retrieved and reject as many as irrelevant images from being retrieved. Given a feature space in which each image is represented as a feature vector, we apply Bayesian theory to determine the degree in which an image in the database is classified as a relevant or an irrelevant one to the query image. It is proven that Bayesian rule is optimal in the expectation of misclassification aspect [6].

We define the notations as follows. We always use boldface symbols to represent vectors or matrices, and non-boldface symbols to represent scalar variables. Given a query image, Let R and I be the events of the relevancy and irrelevancy for all the images in the image database to a query image, respectively, and let Img_i be the ith image in the image database. We use $P()$ to denote a probability, and use $p()$ to denote a probability density function (pdf). Thus, $P(R)$ and $P(I)$ are the prior probabilities of relevancy and irrelevancy for all the images in the image database to the query image, respectively; and $p(Img_i)$ is the pdf of the ith image in the image database. Based on the Bayes' rule the following equations hold:

$$P(R|Img_i) = \frac{p(Img_i|R)P(R)}{p(Img_i)}, \quad P(I|Img_i) = \frac{p(Img_i|I)P(I)}{p(Img_i)} \tag{1}$$

where $i = 1, \ldots, M$ and M is the number of images in the database.

Definition 1. *Given a specific image Img_i in the image database, for any query image, the relevancy confidence of this image to the query image is defined as the posterior probability $P(R|Img_i)$. Similarly, the irrelevancy confidence of this image to the query image is defined as the posterior probability $P(I|Img_i)$. Obviously, the two confidences are related as $P(R|Img_i) + P(I|Ing_i) = 1$.*

The relevancy confidence and irrelevancy confidence of an image are used to indicate the *subjective* relevance and irrelevance degree quantitatively to the query image, respectively. From Eq. 1, the problem of determining whether an image Img_i is (ir)relevant to the query image and the corresponding (ir)relevancy confidence is reduced to estimating the conditional pdfs $p(Img_i|R)$ and $p(Img_i|I)$, respectively, the prior probabilities $P(R)$ and $P(I)$, respectively, and the pdf $p(Img_i)$ in the continuous feature space. These probabilities and pdfs may be estimated from the positive and negative samples provided by the user relevance feedback, as we shall show below.

Since in CBIR, each image is always represented as a feature vector or a group of feature vectors (when each feature vector is used to represent a region or an object in the image) in a feature space, to facilitate the discussion we use a feature vector to represent an image in this paper. Consequently, in the rest of this paper, we use the terminologies vector and image interchangeably. Due to the typical high dimensionality of feature vectors, it is safe and desirable to perform vector quantization before the pdf estimations to ease the computation intensity.

As a preprocessing, we apply uniform quantization to every dimension of feature vectors and each interval is represented by its corresponding representative value.

It is straightforward to estimate the pdf $p(Img_i)$ by statistically counting the percentage of the quantized feature vectors in the feature space of the whole image database. Note that this estimation is performed offline and for each image it is only required to be computed once, resulting in no complexity for online retrieval. For image databases updated with batch manner(most practical databases are updated in this way), the content of databases does not change during the online search session, and periodically updating $p(Img_i)$ along with the database updating is feasible.

Since it is well observed that all the positive (i.e., the relevant) samples "are alike in a way" [19]. In other words some features of the class-of-interest usually have compact support in reality. We assume that the pdf of each feature dimension of all the relevant images to a given query image satisfies the Gaussian distribution.

$$p(x_k|R) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp[-\frac{(x_k - m_k)^2}{2\sigma_k^2}] \tag{2}$$

where x_k is the k^{th} dimension of the feature vector of an image, m_k is the mean value of the x_k of all relevant images to the query image, and σ_k is the variance of the corresponding x_k.

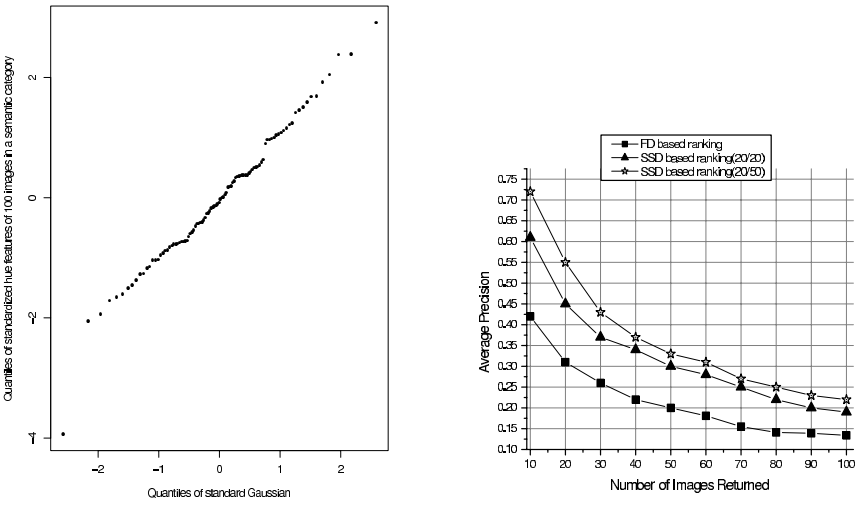
To verify this model for positive samples, we tested on images of several predefined semantic categories. The experiment confirms that the model is practically acceptable. Fig. 1(a) shows a quantile-quantile test [9] of the standardized *hue* feature of 100 images in one predefined semantic category. It is shown that the quantile of the standardized feature dimension and the quantile of the standard Gaussian distribution are similar, which means that the feature dimension of the 100 images in this semantic category can be approximated as a Gaussian.

Assume that $L = \{l^1, l^2, \ldots, l^N\}$ is the labelled relevant sample set. Applying the maximum-likelihood method [2], we obtain the following unbiased estimations of the mean vector m_k and the variance σ_k:

$$\widehat{m_k} = \frac{1}{N}\sum_{i=1}^{N} l_k^i, \quad \widehat{\sigma_k} = \frac{1}{N-1}\sum_{i=1}^{N}(l_k^i - \widehat{m_k})^2 \tag{3}$$

In order to ensure that these estimations are close to the true values of the parameters, we must have sufficient relevant samples. However, the number of relevant samples in each relevance feedback iteration is typically limited. Hence, we develop a cumulative strategy to increase the number of relevant samples. Specifically, the relevant samples in each iterations in a query session are recorded over the iterations; when we estimate the parameters using Eq. 3, we not only use the relevant samples labelled by the user in the current iteration, but also include all the relevant samples recorded in the previous iterations to improve the estimation accuracy.

It is notable that not every feature dimension of relevant images conforms to a Gaussian distribution equally well. It is possible that, for one semantic category,



(a) (b)

Fig. 1. (a): Quantile-quantile test of a standardized feature dimension for images in one semantic category. (b): Average precisions vs. the numbers of the returned images with and without **BALAS** enabled.

some feature dimensions are more semantically related than other dimensions such that these dimensions appear to conform to a Gaussian model better, while other dimensions' distributions in the feature space are jumbled, and thus do not conform to Gaussian well. To describe the difference of conformity degrees and compensate the corresponding effect we introduce a measure, called *trustworthy degree*, for every feature dimension. The trustworthy degree depicts the importance weight for every feature dimension. It is defined as $w_k = \frac{\sigma_k^{-1}}{\max_{k=1}^{T} \sigma_k^{-1}}$, where T is the number of dimensions in one image feature. If the variance of the relevant samples is high along a dimension k, we deduce that the values on this dimension are not very relevant to the query image and thus the Gaussian distribution might not be a good model for this dimension because the features are not centered with a prominent mean. Consequently a low trustworthy degree w_k is assigned. Otherwise, a high trustworthy degree w_k is assigned. Note that the $\max w_k = 1$ for $k = 1 \ldots T$.

It is reasonable to assume all dimensions of one feature are independent (raw features *per se* are independent, e. g., color and texture features, or we can always apply K-L transform [5] to generate uncorrelated features from raw features; in this way the support to independency is strengthened), thus the pdf of positive samples is determined as a trustworthy degree pruned joint pdf:

$$p(\boldsymbol{x}|R) = \prod_{\substack{k=1 \\ w_k \geq \delta}}^{T} p(x_k|R) \tag{4}$$

where δ is a threshold for incorporating only high trustworthy dimensions (conforming to the Gaussian model well) to determine $p(\boldsymbol{x}|R)$. Those dimensions that do not conform to the Gaussian distribution well would result in inaccurate pdf estimations, and consequently are filtered out.

In order to correctly and accurately estimate the conditional pdf distribution for the negative samples, we assume that each negative sample represents a unique potential semantic class, and we apply the kernel density estimator [12] to determining the statistical distribution function of this irrelevance class. In case two negative samples happen to come from the same semantic class, it is supposed that they would exhibit the same distribution function, and thus this assumption is still valid. Consequently, the overall pdf for the negative samples is the agglomeration of all the kernel functions.

We choose the kernel function in the estimator as an isotropic Gaussian function (assuming all the feature vectors have been normalized). The window of the estimation is a hyper-sphere centered at each negative sample $\boldsymbol{x_j}, j = 1, 2, \ldots, N$, assuming that there are N negative samples in total. Let the radius of the jth hyper-sphere be r_j, which is called the *bandwidth* of the kernel density estimation in the literature [3]. Typically it is practical to assume that $r_j = r$ for all the different j, where r is a constant bandwidth. Hence, the conditional pdf to be estimated for the sample $\boldsymbol{x_i}$ in the feature space is given by

$$p(\boldsymbol{x_i}|I) = \sum_{j=1}^{N} kernel(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sum_{j=1}^{N} \exp\{-\frac{\|\boldsymbol{x_i} - \boldsymbol{x_j}\|_2^2}{2r_j^2}\} \tag{5}$$

where $\|\boldsymbol{x_i} - \boldsymbol{x_j}\|_2$ is the Euclidian distance between the neighboring sample $\boldsymbol{x_j}$ and the center feature vector $\boldsymbol{x_i}$.

The choice of the bandwidth r has an important effect in the estimated pdfs. If the bandwidth is too large, the estimation would suffer from low resolution. On the other hand, if the bandwidth is too small, the estimation may be locally overfitted, hurting the generalization of the estimation. In this consideration, the optimal Parzen window size has been studied extensively in the literature [13]. In practice, the optimal bandwidth may be determined by minimizing the *integrated squared error* (ISE), or the *mean integrated squared error* (MISE). Adaptive bandwidth is also proposed in the literature [13]. For simplicity, we choose a constant bandwidth r based on the maximum distance from all the negative samples to their closest neighbor D defined as $r = \lambda D = \lambda \max_{\boldsymbol{x_k}}[\min_{\boldsymbol{x_l}}(\|\boldsymbol{x_k} - \boldsymbol{x_l}\|_2)]$, where λ is a scalar. We find in our experiments that with well-normalized feature vectors, a λ between 1 and 10 often gives good results.

The computational overhead in estimating conditional pdf with Eq. 5 is tractable due to the limited number of negative samples and utilization of dimensionality reduction techniques, such as PCA [5], on the low-level features, while the estimation accuracy is acceptable.

Since negative samples may potentially belong to different semantic classes, and since each such semantic class only has a very limited number of samples thus far in one typical relevance feedback iteration, we must "generate" a sufficient number of samples to ensure that the estimated pdf for the negative

samples is accurate. To solve for this "scarce sample collection" problem, we actually generate additional negative samples based on the kernel distributions for each semantic classes defined in Eq. 5. These generated additional samples are the hypothetical images. For the sake of discussion, we call the original negative samples provided by the user in the relevance feedback iterations as the *labelled* samples, and the generated samples as the *unlabelled* samples. To ensure that the number of generated samples is sufficiently large, for each labelled negative sample in one relevance feedback iteration, we generate q additional unlabelled negative samples based on Eq. 5, where q is a parameter. To ensure a "fair sampling" to the kernel function in Eq. 5, the generation of the unlabelled samples follows a probability function defined by the following Gaussian pdf function $p(\boldsymbol{y}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{\|\boldsymbol{y}-\boldsymbol{x_i}\|_2^2}{2\sigma^2}\}$, where $d = \|\boldsymbol{y}-\boldsymbol{x_i}\|_2$ is the Euclidian distance between the unlabelled sample $\boldsymbol{y}$ and each labelled sample $\boldsymbol{x_i}$, and σ is the standard deviation, which is set to the average distance between two feature vectors in the labelled negative feature space defined as $\sigma = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \|\boldsymbol{x_i} - \boldsymbol{x_j}\|_2$.

Hence, an unlabelled sample is more likely to be selected if it is close to a labelled negative sample. The probability density defined in above decays when the Euclidian distance to the labelled sample increases.

Consequently, an algorithm, called SAMPLING, is designed to perform the unlabelled sample selection based on roulette wheel selection strategy [2]. SAMPLING implements a roulette wheel sampling strategy to select unlabelled samples. The unlabelled samples with smaller distances to a labelled sample have larger probabilities to be selected as the additional samples. On the other hand, those potential unlabelled samples farther away from a labelled sample are not completely eliminated from being selected, though their chances of being selected are small. With the extended number of the negative samples, the accuracy of the pdf estimation defined in Eq. 5 is significantly improved. Similarly, the cumulative learning principle adopted in the estimation of the conditional pdf for the positive samples described above is also applied in the estimation of the conditional pdf for the negative samples to further improve the estimation accuracy.

In order to determine the relevancy and irrelevancy confidences defined as the posterior probabilities in Eq. 1, we must solve for the prior probabilities $P(R)$ and $P(I)$ first. Unlike the typical approach in the classical pattern classification problems in which a prior probability is usually estimated from the supervised training samples, in the problem of the relevance feedback in image retrieval the relevancy or irrelevancy of an image is subject to different query images and different users' subjective preferences. Thus, the relevancy and irrelevancy of an image vary in different queries and in different query sessions. Consequently, it is impossible to estimate the prior probabilities in advance. In other words, these prior probabilities must be estimated online also in solving for the relevance feedback problem.

Given a query image, for each image Img_i in the image database, we have

$$p(Img_i) = p(Img_i|R)P(R) + p(Img_i|I)P(I) \tag{6}$$

and for the query image we also have

$$P(R) + P(I) = 1 \tag{7}$$

Combining Eqs. 6 and 7, we immediately have:

$$P(R) = \frac{p(Img_i) - p(Img_i|I)}{p(Img_i|R) - p(Img_i|I)} \tag{8}$$

From Eq. 8, it is clear that since we have already developed methods to determine $p(Img_i|R)$, $p(Img_i|I)$, and $p(Img_i)$, the prior probability $P(R)$ may be uniquely determined immediately. Thus, $P(I)$ may also be immediately determined from Eq. 7. This reveals that for each query image given, the *overall* relevancy and irrelevancy of *all* the images in the image database may be uniquely determined by *any individual* image Img_i in the image database. In other words, any individual image Img_i in the image database may be used to determine the prior probabilities, and given a query image, the prior probabilities are independent of the selection of any of the images in the database. The experimental results have verified this conclusion. Nevertheless, due to the noise in the data, in practice, the estimated prior probabilities based on different individual images in the database may exhibit slight variations. In order to give an accurate estimation of the prior probabilities that are not subject to the bias towards a specific image in the database, we denote $P_i(R)$ as the prior probability determined in Eq. 8 using the individual image Img_i, and $P(R)$ is the average from all the images in the database, i.e., $P(R) = \frac{1}{M}\sum_{i=1}^{M} P_i(R)$. The prior probability $P(I)$ is thus determined accordingly.

Given a query image in a query session, for each image Img_i in the database, there is a corresponding relevancy confidence $P(R|Img_i)$, which represents the relevancy degree of this image to the query image learned from the user's subjective preference through the relevance feedback. Hence, this relevancy confidence captures the *subjective* relevancy degree of each image in the database to a query. On the other hand, for any CBIR system, there is always a feature-based distance measure used for image retrieval. The feature-based distance measure typically does not incorporate the user relevance preferences, and thus, only captures the *objective* proximity degree in the feature space of each image in the database to a query. Consequently, in order to design a ranking scheme in image retrieval that "makes best sense", it is natural to consider to integrate the subjective relevancy confidence and the objective distance measure together through taking advantage of labelled sample image set to define an comprehensive ranking scheme.

Note that the relevancy confidence and the feature-based distance measure are complementary to each other. Exploiting this property explicitly, we define a unified ranking scheme, called *Session Semantic Distance* (SSD), to measure the relevance of any image Img_i within the image database in terms of both relevancy confidence $P(R|Img_i)$, irrelevancy confidence $P(I|Img_j)$, and feature-based distance measure $FD(Img_i)$.

The SSD for any image $SSD(Img_i)$ is defined using a modified form of the Rocchio's formula [10] as follows:

$$\begin{aligned} SSD(Img_i) &= \log(1 + P(R|Img_i))FD(Img_i) \\ &+ \beta\{\frac{1}{N_R}\sum_{k\in D_R}[(1 + P(R|Img_k))D_{ik}]\} \\ &- \gamma\{\frac{1}{N_I}\sum_{k\in D_I}[(1 + P(I|Img_k))D_{ik}]\} \end{aligned} \tag{9}$$

where N_R and N_I are the sizes of the positive and negative labelled sample set D_R and D_I, respectively, in the feedback. D_{ik} is the feature-based distance between the image Img_i and Img_k. We have replaced the first parameter α in Rocchio's formula with the logarithm of the relevancy confidence of the image Img_i. The other two parameters β and γ are assigned a value of 1.0 in our current implementation of the system for the sake of simplicity. However, other values can be given to emphasize the different weights between the last two terms.

With this definition of the $SSD(Img_i)$, the relevancy confidence of Img_i, the relevancy confidence of images in the labelled relevant set, the irrelevancy confidence of images in the labelled irrelevant set, and the objective feature distance measure are integrated in a unified way. The (ir)relevancy confidences of images in the labelled sample set act adaptively as weights to correct the feature-based distance measure. In the ranking scheme, an image is ranked high in the returned list if it is similar, in relevancy confidence measure and/or feature-based distance measure, to the query image and images in the labelled relevant image set and it is dissimilar to images in the labelled irrelevant image set in both relevancy confidence and feature-based distance measure; otherwise, its rank is low. Thus, the robustness and accuracy of the semantic distance measure is improved, resulting in lower false-positives, by using both subjective and objective similarity measures to form a more accurate measure for semantic similarity.

3 Experiments and Discussions

The focus of this paper is on user relevance feedback in image retrieval rather than on a specific image indexing and retrieval method. The relevance feedback methodology we have developed in this paper, **BALAS**, is independent of any specific image indexing and retrieval methods, and in principle, may be applied to any such image indexing and retrieval methods. The objective of this section is to demonstrate that **BALAS** can effectively improve the image retrieval relevancy through the user relevance feedback using any specific CBIR system.

For the evaluation purpose, we implemented an image indexing and retrieval prototype system. Many types of low-level features may be used to describe the content of images. In the current implementation, we use color moment. We extract the first two moments from each channel of CIE-LUV color space, and the simple yet effective *L2* distance is used to be the feature-based ranking metric. Since the objective is to test the relevance feedback learning method

rather than to evaluate features, the feature we use is not as sophisticated as those used in some existing CBIR systems [17,18].

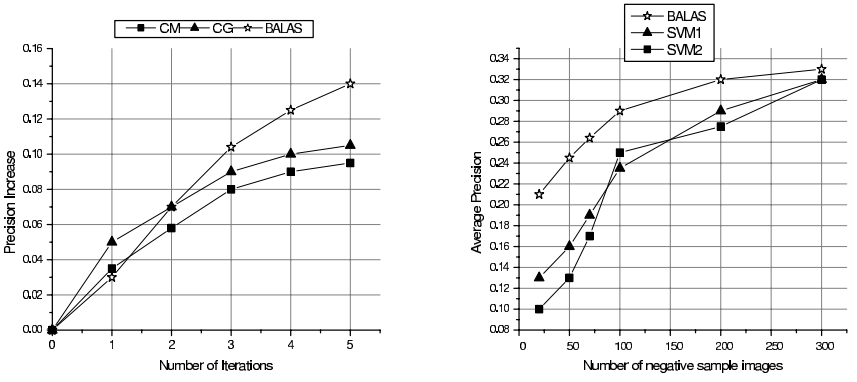
The following evaluations are performed on a general-purpose color image database containing 10,000 images from the COREL collection with 96 categories. 1,500 images were randomly selected from all the categories of this image database to be the query set. A retrieved image is considered semantics-relevant if it is in the same category of the query image. We note that the category information in the COREL collection is only used to ground-truth the evaluation, and we do not make use of this information in the indexing and retrieval procedures.

In order to evaluate the semantics learning capability of **BALAS**, we implemented the **BALAS** methodology on the prototype CBIR system, which we also call **BALAS** for the purpose of the discussion in this paper. The threshold δ in Eq. 4 was empirically set as 0.7 in the prototype. Since user relevance feedback requires subjective feedback, we invite a group of 5 users to participate the evaluations. The participants consist of CS graduate students as well as lay-people outside the CS Department. We ask different users to run **BALAS** initially without the relevance feedback interaction, and then to place their relevance feedbacks after the initial retrievals. For the evaluation purpose, we define the retrieval precision as the ratio of the number of relevant images retrieved to the total number of retrieved images in each round of the retrieval in a query session. For the comparison purpose, we have recorded the retrieval precisions in the initial retrieval, i.e., without the **BALAS** relevance feedback capability and purely based on the similarity measure, the retrieval precisions after every rounds of relevance feedback using **BALAS** *only* based on the relevancy confidence, and the retrieval precisions after every rounds of relevance feedback using **BALAS** based on the session semantic distance, respectively. All the reported data are the averages of the whole group of users. The average time for each round of retrieval after the relevance input is about 5 seconds on a $PentiumIV$ 2GHz computer with 512MB memory.

We ran the implemented CBIR system with **BALAS** for the 1,500 query image set with varied number of truncated top retrieved images and plotted the curves of the average retrieval precision vs. the number of truncated top retrieved images. Fig. 1(b) shows the average precision-scope plot for the system with and without **BALAS** enabled. In other words, for ranking scheme one test is based solely on feature-based distance FD and another test is based on session semantic distance SSD with different numbers of provided sample images. The notation (m/n) in the figure legend denotes the number of positive sample images vs. number of negative sample images for the learning. It is clear that the **BALAS** relevance feedback learning capability enhances the retrieval effectiveness substantially.

For performance comparison, we used the same image database and the query set to compare **BALAS** with the relevance feedback method developed by Yong and Huang [11], which is a combination and improvement of its early version and MindReader [7] and represents the state-of-the-art relevance feedback research in the literature. Two versions of [11] are implemented. The first uses the color moments (CM) computed in the same way as described above and the other

uses the correlogram (and thus is called CG here). The overall comparison evaluations are documented in Fig. 2(a). The average precision in this evaluation is determined based on the top 100 returned images for each query out of the 1,500 query image set. From the figure, it appears that during the first two iterations, the CG version of [11] performs noticeably better than **BALAS** while the CM version of [11] performs comparably with **BALAS**. After the second iteration, **BALAS** exhibits a significant improvement in performance over that of [11] in either of the two versions, and as the number of iterations increases, the improvement of the performance of **BALAS** over [11] appears to increase also. This also confirms with the cumulative learning strategy employed in **BALAS** and the fact that when more iterations of relevance feedback are conducted, more learning samples are given, and thus more accurate density estimation may be expected from **BALAS**.



(a) (b)

Fig. 2. (a): Retrieval precision comparison using relevance feedback between **BALAS** and CM and CG. (b): Average precision in top 100 images returned. Number of positive sample images =20. SVM1 denotes SVM classifier with $\sigma = 50$ and SVM2 denotes SVM classifier with $\sigma = 100$.

To evaluate the effectiveness of explicitly addressing the asymmetry property of CBIR, we compared **BALAS** with SVM [15] classification method. SVM classifier adopts the two-class assumption and treats positive and negative samples equally, which is not valid in CBIR as is discussed above. In addition, there is no satisfactory method to optimally select kernel function and its parameters other than empirically testing yet. In the comparison experiment, the RBF kernel $K(x, y) = \exp^{-\|x-y\|^2/2\sigma^2}$ with different σs were tested for SVM classifier. The original SVM classifier only gives a decision boundary without providing confidence of each object belonging to each class. To utilize SVM classifiers in image retrieval, a ranking scheme is needed. In the comparison, *Larger margin first* retrieval scheme [15] is adopted for SVM to determine the rank of the

retrieved images. A query set was composed of randomly selected 100 images, which was applied to **BALAS** and SVM, respectively; the average precisions in the top 100 images were recorded for different number of negative sample images with the number of positive samples images fixed. SVM with two different σs were tested; $\sigma = 50$ and $\sigma = 100$. Fig. 2(b) shows the result. We see that the performance of SVM is affected by σ in some degree but **BALAS** outperforms SVM consistently. The unsatisfactory performance of SVM is partially due to the false assumption that the two classes are equivalent and the negative samples are representative of the true distributions. With this invalid assumption, we found that the positive part "spills over" freely into the part of the unlabelled areas of the feature space by the SVM classification. The result of this "spillover" effect is that after the user's feedback, the machine returns a totally different set of images, with most of them likely to be negative. In **BALAS**, this phenomenon did not occur due to the asymmetric density estimations.

4 Conclusions

This paper focuses work on user relevance feedback in image retrieval. We take this problem as a standard two-class pattern classification problem aiming at refining the retrieval precision by learning through the user relevance feedback data. However, we have investigated the problem by noting two important unique characteristics: small sample collection and asymmetric sample distributions between positive and negative samples. We have developed a novel approach to stretching Bayesian learning to solve for this problem by explicitly exploiting the two unique characteristics, which is the methodology of **BA**yesian **L**earning in **A**ymmetric and **S**mall sample collections, thus called **BALAS**. Different learning strategies are used for positive and negative sample collections in **BALAS**, respectively, based on the two unique characteristics. By defining the relevancy confidence as the relevant posterior probability, we have developed an integrated ranking scheme in **BALAS** which complementarily combines the subjective relevancy confidence and the objective similarity measure to capture the overall retrieval semantics. The experimental evaluations have confirmed the rationale of the proposed ranking scheme, and have also demonstrated that **BALAS** is superior to an existing relevance feedback method in the literature in capturing the overall retrieval semantics.

References

1. A. D. Bimbo. *Visual Information Retrieval.* Morgan kaufmann Pub., San Francisco, CA, 1999.
2. G. Blom. *Probability and Statistics: Theory and Applications.* Springer Verlag, London, U. K., 1989.
3. S.-T. Chiu. A comparative review of bandwidth selection for kernel density estimation. *Statistica Sinica*, 16:129–145, 1996.

4. I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos. The bayesian image retrieval system, pichunter: Theory, implementation and psychophysical experiments. *IEEE Trans. on Image Processing*, 9(1):20–37, 2000.
5. W. R. Dillon and M. Goldstein. *Multivariate Analysis, Mehtods and Applications.* John Wiley and Sons, New York, 1984.
6. R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* John Wiley and Sons, New York, 1973.
7. Y. Ishikawa, R. Subramanya, and C. Faloutsos. Mindreader: Query databases through multiple examples. In *the 24th VLDB Conference Proceedings*, New York, 1998.
8. M. D. Marsicoi, L. Cinque, and S. Levialdi. Indexing pictorial documents by their content: a survey of current techniques. *Imagee and Vision Computing*, 15:119–141, 1997.
9. B. D. Ripley and W. N. Venables. *Modern Applied Statistics with S.* Springer Verlag, New York, New York, 2002.
10. Rocchio and J. J. Relevance feedback in information retrieval. In *The SMART Retreival System — Experiments in Automatic Document Processing*, pages 313–323. Prentice Hall, Inc, Englewood Cliffs, NJ, 1971.
11. Y. Rui and T. S. Huang. Optimizing learning in image retrieval. In *IEEE Conf. Computer Vision and Pattern Recognition*, South Carolina, June 2000.
12. B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, New York, 1986.
13. G. R. Terrell and D. W. Scott. Variable kernel density estimation. *The Annals of Statistics*, 20:1236–1265, 1992.
14. K. Tieu and P. Viola. Boosting image retrieval. In *IEEE Conf. Computer Vision and Pattern Recognitin Proceedings*, South Carolina, June 2000.
15. V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.
16. Y. Wu, Q. Tian, and T. S. Huang. Discriminant em algorithm with application to image retrieval. In *IEEE Conf. Computer Vision and Pattern Recognition Proceedigns*, South Carolina, June 2000.
17. R. Zhang and Z. Zhang. Addressing cbir efficiency, effectiveness, and retrieval subjectivity simultaneously. In *ACM Multimedia 2003 Multimedia Information Retrieval Workshop*, Berkeley, CA, November 2003.
18. R. Zhang and Z. Zhang. A robust color object analysis approach to efficient image retrieval. *EURASIP Journal on Applied Signal Processing*, 2004.
19. X. S. Zhou and T. S. Huang. Small sample learning during multimedia retrieval using biasmap. In *IEEE Conf. Computer Vision and Pattern Recognition Proceedings*, Hawaii, December 2001.