BALAS: Empirical Bayesian Learning in the Relevance Feedback of Image Retrieval[*]

Ruofei Zhang[†]and Zhongfei (Mark) Zhang
Department of Computer Science
Sate University of New York at Binghamton
Binghamton, NY 13902, U. S. A.
email: {rzhang,zhongfei}@cs.binghamton.edu
Phone: 001-607-777-2935
Fax: 001-607-777-4729

Abstract

This paper is 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 empirical 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**. In **BALAS** different learning strategies are used for positive and negative sample collections, 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 current literature in capturing the overall retrieval semantics.

Keywords: CBIR, relevance feedback, Bayesian learning, relevancy confidence, session semantic distance.

[*] A preliminary version of this work appeared in the proceedings of the 8th European Conference on Computer Vision (ECCV), Prague, Czech Republic, May, 2004.

[†] Corresponding author.

1. Introduction

Very large collections of images have become ever more common than before. From stock photo collections and proprietary databases to the World Wide Web, these collections are diverse and often poorly indexed; unfortunately, image retrieval systems have not kept pace with the collections they are searching. How to effectively index and retrieve semantically relevant images according to users' queries is a challenging task. Most existing image retrieval systems, such as image search engine in Google [1], are textual based. The images are searched by using the surrounding text, captions, keywords, etc. Although the search and retrieval techniques based on textual features can be easily automated, they have several inherent drawbacks. First, textual description is not capable to capture the visual contents of an image accurately and in many circumstances the textual annotations are not available. Second, different people may describe the content of an image in different ways, which limits the recall performance of textual-based image retrieval systems. Third, for some images there is something that no words can convey. Try to imagine an editor taking in pictures without seeing them or a radiologist deciding on a verbal description. The content of images are beyond words. They have to be seen and searched as pictures: by objects, by style, by purpose.

To resolve these problems, Content-Based Image Retrieval (CBIR) has attracted significant research attention [18, 23, 30, 16]. In CBIR, a query image (an image to which a user tries to find similar ones) is imposed to the image retrieval system to obtain the semantics-relevant images. The similarity between the query image and the indexed images in the image database are determined by their visual contents, instead of the textual information. Early research of CBIR focused on finding the "best" representation for image features, e. g., color, texture, shape, and spatial relationships. 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 [2]. In this context, high level concepts and user's perception subjectivity cannot be well modeled. Recent approaches introduce more advanced human-computer interaction (HCI) into CBIR. The retrieval procedure incorporates user's interaction into the loop, which consists of several iterations. In each iteration, the user cast *positive samples* (relevant images) as well as *negative samples* (irrelevant images) for the returned results from the previous iteration. Based on user's feedback, the retrieval system is able to adaptively customize the search results to the user's query preference. This interaction mechanism is called relevance feedback, which allows a user continuously refine his(her) querying information after submitting a coarse initial query to the image retrieval system. This approach greatly reduces the labor required to precisely compose a query and easily captures the user's subjective retrieval preference.

However, most approaches to relevance feedback, e. g., [27, 21, 22, 46], are based on heuristic formulation of empirical parameter adjustment, which is typically ad hoc and not systematic, and thus cannot be substantiated well. Some of the recent work [43, 17, 37, 38, 34] formulates the relevance feedback problem as a classification or learning problem. Without further exploiting the unique characteristics of the 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.

Before we design a specific relevance feedback methodology, two unique characteristics of the relevance feedback problem in image retrieval must be observed and addressed when compared with the general pattern classification problems. The first is the small sample collection issue. In relevance feedback of image retrieval, the number of the training samples is usually small (typically < 20 in each iteration 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 typical image databases. The second characteristic is the asymmetric training sample issue. Most classification or learning techniques proposed in the literature of pattern recognition and machine learning, such as discriminant analysis [10] and Support Vector Machine(SVM) [39] regard the positive and negative examples interchangeably, and assume that both sets are distributed approximately equally. However, in relevance feedback of image retrieval, while it is reasonable to assume that all the positive samples conform to a single 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 asymmetry characteristic must be taken into account in the relevance feedback learning.

In this paper, we investigate the relevance feedback problem in image retrieval using empirical Bayesian learning. Specifically, we apply 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 probability 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 image feature vectors, to order the ranking of the retrieved images from an image database.

The rest of the paper is organized as follows. Beginning with the discussion of the related work in Section 2, we describe **BALAS** methodology in Section 3, in which the probability density estimations for the positive and negative sample distributions are introduced, and the measurement of the relevancy confidence is presented. In Section 4 the ranking scheme of **BALAS**, *session semantic distance*, is defined and developed. **BALAS** is evaluated as a prototype system, and the evaluations are reported in Section 5. Finally, the paper is concluded in Section 6.

2. Related work

The problem of relevance feedback of image retrieval was identified and then received focused attention even in the early days of the CBIR research. Some of the early efforts attempted to learn a new query and the relative importance of different features or feature components [27, 19, 20], while others attempted to learn a linear transformation in the feature space taking into account correlations among feature components [13, 26], or to map the learning problem to a two-class discriminant classification problem [40]. In particular, inspired by term weighting and relevance feedback techniques in textual document retrieval, some of those efforts focused on heuristics-based techniques to adjust parameters empirically, such as [27, 21]. The intuition was to emphasize more on the features that best cluster the positive samples and maximize the separation between the positive and negative samples. To achieve this goal, Kohonen's Learning Vector Quantization (LVQ) algorithm [42] and the tree-structured self-organizing map (TS-SOM) [15] were used for dynamic data clustering during the relevance feedback. For example, Laaksonen et al. [15] applied TS-SOMs to index images along different feature dimensions such as color and texture. Positive and negative examples were mapped to positive and negative impulses and a low-pass filtering was applied to generate a map to implicitly reveal relative importance of different features, as a "good" map always keeps positive examples well clustered while negative examples are scattered away. Similarly, Peng et al. [20] used the same intuition to capture the feature relevance but took the probabilistic approach instead. Santini and Jain [28] proposed a method to optimize a parametric similarity metric according to the feedback from users.

Recently, the problem of the relevance feedback of image retrieval was investigated in the literature from a more systematic point of view by formulating it into an optimization problem. In the work of Ishikawa et al. [13] and Rui and Huang [26], based on the minimization of the total distance of positive examples from the new query, an optimal solution was shown to be the weighted average as the new query and a whitening transform in the feature space. Moreover, Rui and Huang [26] adopted a two-level weighting scheme to cope with the singularity issue due to the small number of training samples. MacArthur et al. [17] cast the relevance feedback problem into a two class learning problem, and used a decision tree algorithm to sequentially "cut" the feature space until all the points in the feature space within a partition are of the same class. The image data set was classified by the resulting decision tree: images that fell into a relevant leaf were collected and the nearest neighbors of the query were returned. Classification techniques from Machine Learning community were extensively used in the recent literature to solve for the relevance feedback problem of image retrieval, such as the BiasMap [47] and the SVM active learning method [38], both using a kernel form to deal with the nonlinear classification boundaries, with the former emphasizing the small sample collection issue while the latter exploring the active learning issue. Active learning in image retrieval has been attracting attentions lately, several methods have been proposed [11, 41]. Different metrics were proposed to select the most informative images for the user's labeling and the objective is to minimize the feedback rounds for a desire retrieval results by quickly learning a boundary that separates the images that satisfy the user's query concept form the rest of the database. Although interesting and reported promising, the evaluation and performances analysis of active leaning methods and traditional learning methods need more investigation.

Alternatively one-class classifiers were proposed, which formalized the relevance feedback in image retrieval as a one-class (relevant images) classification problem [35, 6]. In this formalization, only positive examples are used while the presentation of users' query interests in aspect of irrelevancy is not exploited. Compared to the methodologies using both positive and negative examples, the performance of one-class classifiers in the context of image retrieval is limited [14, 44]. In two-class classification category, Wu et al. [43] developed the D-EM algorithm within the transductive learning framework based on the examples both from the user feedback data (called labeled data) and from other data points (called unlabeled data); the method performs discriminant analysis through the EM (i.e., expectation and maximization) iterations to select a subspace of features, such that the two-class (i.e., the positive and negative sample classes) assumption on the data distribution has a better support. One notable work employing Bayesian relevance feedback is PicHunter [8]. Also using Bayesian reasoning, there are three major differences between PicHunter and BALAS. First, the motivation is different. PicHunter addresses the "target search" problem, in which users seek to find a specific target image, i. e., exactly the same image as the query image. While BALAS attempts to find more semantics-relevant images for a query image, which is harder compared with "target search". Second, PicHunter assumes that the probability of any given image being the target is independent of who the user is, in other words, all users are identical. This assumption does not hold true in the semantics-relevant image retrieval. Third, the BALAS methodology is different from PicHunter's methodology. In PicHunter, a user model is derived from the offline learning and tuning to obtain the correlation between users' action and the displayed image set, whereas in BALAS we determine the relevancy/irrelevancy degree of each image by the online density estimation.

In these learning methods in relevance feedback of image retrieval, they are all based on the assumption that both positive and negative samples conform either implicitly or explicitly to a well formed distribution. Considering the two unique characteristics of the relevance feedback problem of image retrieval that in a typical user relevance feedback scenario, the sample collection is small and the two class sample collections exhibit asymmetric property, this assumption is often hardly valid, especially for negative samples, consequently limiting the performance of these methods. In this paper, we propose **BALAS** methodology that not only generates an optimal classifier but also exploits the two unique characteristics of the problem to arrive at a novel *relevancy confidence* measure solving for the relevance feedback problem in image retrieval. In addition, **BALAS** offers an effective ranking scheme to integrate the *relevancy confidence* measure with a conventional feature-based distance to model the semantic similarity more precisely than the existing methods.

3. BALAS Methodology

Given a query image, it is natural that an indexed image data set can be classified into two classes of images, one is relevant in semantic content to the query and the other is irrelevant. 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. Consequently, this learning problem is reduced to a two-class classification problem in essence.

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 image data set 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 [10]. In other words, no other rule has a lower expected error rate.

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 data set to a query image, respectively, and let Img_i be the ith image in the image data set. 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 indexed data set to the query image, respectively; $p(\text{Img}_i)$ is the pdf of the ith image in the image data set; $P(R|\text{Img}_i)$ and $P(I|\text{Img}_i)$ are the conditional probabilities of the ith image's relevancy and irrelevancy to the query image, respectively; and $p(\text{Img}_i|R)$ and $p(\text{Img}_i|I)$ are the pdfs of the ith image given the relevant and irrelevant classes, respectively, in the image data set to the query image. Based on the Bayes' rule the following equations hold:

$$P(R|\text{Img}_i) = \frac{p(\text{Img}_i|R)P(R)}{p(\text{Img}_i)} \tag{1}$$

$$P(I|\text{Img}_i) = \frac{p(\text{Img}_i|I)P(I)}{p(\text{Img}_i)} \tag{2}$$

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

Definition 1 *Given a specific image Img$_i$ in an image data set, for any query image, the relevancy confidence of this image to the query image is defined as the posterior probability $P(R|\mathit{Img}_i)$. Similarly, the irrelevancy confidence of this image to the query image is defined as the posterior probability $P(I|\mathit{Img}_i)$. Obviously, the two confidences are related as $P(R|\mathit{Img}_i) + P(I|\mathit{Img}_i) = 1$.*

The relevancy confidence and irrelevancy confidence of an image are used to quantitatively describe the *subjective* relevance and irrelevance degrees to the query image, respectively.

From Eqs. 1 and 2, 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 $p(\text{Img}_i|R)$ and $p(\text{Img}_i|I)$, respectively, the prior probabilities $P(R)$ and $P(I)$, respectively, and the pdf $p(\text{Img}_i)$ 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 (e.g., when each feature vector is used to represent a region or an object in the image [4]) 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, we perform vector quantization before the pdf estimations to ease the computation intensity. Typically, as a preprocessing, uniform quantization is applied to every dimension of feature vectors and each interval is represented by its corresponding representative value. In the rest of this paper, all the feature vectors in the image data set are meant to be the quantized feature vectors.

It is straightforward to estimate the pdf $p(\text{Img}_i)$ by statistically counting the percentage of the quantized feature vectors in the feature space of the whole image data set. 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 working periods and periodically updating $p(\text{Img}_i)$ with data set updating is feasible.

While the relevance feedback problem in image retrieval is a typical two class classification problem, due to the asymmetry nature between the positive and negative samples collected in the relevance feedback, the often-used assumption in the literature that the positive and the negative samples both conform to their corresponding well-formed distribution functions is not typically valid. In fact, positive samples typically have a compact, low-dimensional support while negative samples can have arbitrary configurations [47]. Moreover, in relevance feedback, the sample collection, either positive or negative, is typically small. To address these characteristics explicitly and precisely, the **BALAS** methodology employs different strategies to estimate the conditional pdfs $p(\text{Img}_i|R)$ and $p(\text{Img}_i|I)$, respectively.

3.1. Estimating the Conditional pdf of the Positive Samples

It is well observed that all the positive (i.e., the relevant) samples "are alike in a way" [47]. 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.

$$p(x_k|R) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp[-\frac{(x_k - m_k)^2}{2\sigma_k^2}] \tag{3}$$

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 k^{th} dimension of the relevant images.

To verify this model for positive samples, we have tested it on images of several predefined semantic categories. The experiment confirms that the model is practically acceptable. Fig. 1 shows a quantile-quantile test [24] 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 $\boldsymbol{L} = \{\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_N\}$ is the relevant sample set provided by a user. Applying the maximum-likelihood method [3], we obtain the following unbiased estimations of the mean m_k and the standard deviation σ_k for the k^{th} dimension of the features.

$$\widehat{m_k} = \frac{1}{N} \sum_{i=1}^{N} l_{ki} \tag{4}$$

and

$$\widehat{\sigma_k} = \frac{1}{N-1} \sum_{i=1}^{N} (l_{ki} - \widehat{m_k})^2 \tag{5}$$

where l_{ki} denotes the k^{th} dimension of the feature vector $\boldsymbol{l}_i$.

In order to ensure that these estimates 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 available relevant samples. Specifically, the relevant samples in each iterations in a query session are recorded over the iterations; when we estimate the parameters using Eqs. 4 and 5, we not

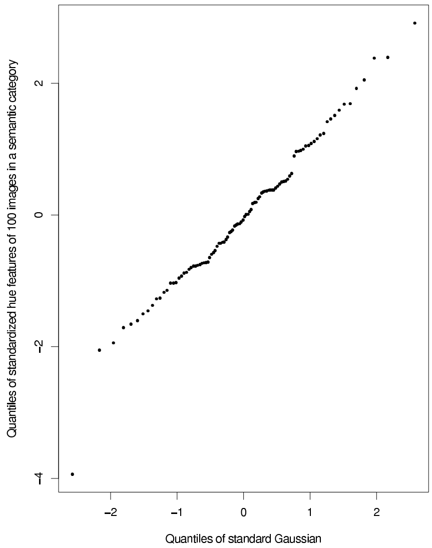


Figure 1: Quantile-quantile test of a standardized feature dimension for images in one semantic category

only use the relevant samples labeled 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 the relevant images conforms to a Gaussian distribution equally well. It is possible that, for one semantic category, 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 a Gaussian model well. To describe the different conformity degrees to Gaussian distributions among different dimensions in the feature space we introduce a measure, called *trustworthy degree*, to each feature dimension. The trustworthy degree depicts the goodness of model fitting on each feature dimension. It is defined as follows:

$$w_k = \frac{\sigma_k^{-1}}{\max_{j=1}^{T} \sigma_j^{-1}} \tag{6}$$

for each dimension k, $k = 1, \ldots, T$, where T is the number of dimensions in one image feature. This is a heuristic measure. The justification follows. If the variance of the relevant samples is large along the dimension k, then we can 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 well around a prominent mean. Consequently, a low trustworthy degree w_k is assigned. Otherwise, a high trustworthy degree w_k is assigned. Note that the concept of the trustworthy degree is always relative to different image databases with different feature distributions. That is why we use the $\max$ function in the denominator, which maps $w_k \in [0, 1]$ for $k = 1, \ldots, T$.

To simplify the estimation of the class probability, we assume that all dimensions of one feature are independent (the raw features *per se* are independent, e. g., color and texture features, or we can always apply K-L transform [9] to generate uncorrelated features from the raw features, resulting in the strengthened support to the assumption); this is just a justifiable design decision. Thus, the pdf of positive samples is determined as a trustworthy degree pruned joint pdf:

$$p(\boldsymbol{x}|R) = \prod_{\substack{k=1 \\ w_k \geq \delta}}^{T} p(x_k|R) \tag{7}$$

where δ is a threshold for incorporating only high trustworthy dimensions (conforming to the Gaussian model well) to determine $p(\boldsymbol{x}|R)$; it is determined empirically as 0.70 in the experiment by using cross-validation. Those dimensions that do not conform to the Gaussian distribution well would result in inaccurate pdf estimations, and consequently are filtered out.

3.2. Estimating the Conditional pdf of the Negative Samples

Unlike the positive samples which may be assumed to conform to a single, well-formed distribution function, negative (i.e., the irrelevant) samples may not be assumed to follow a single distribution function, as there may be many different irrelevant semantics to a query image. While the relevance feedback problem may be tackled as a general two-class classification problem (the relevance class and the irrelevance class), due to the fact that each negative sample is "negative in its own way" [47], samples from the irrelevance class may come from different semantic classes. Consequently, it is neither reasonable nor practical to assume that all negative samples conform to one single distribution function as is assumed for the positive samples.

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 [29] to determine the statistical distribution function of all negative samples. 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 [7]. 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

$$p(\boldsymbol{x_i}|I) = \sum_{j=1}^{N} kernel(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sum_{j=1}^{N} \exp\{-\frac{\|\boldsymbol{x_i} - \boldsymbol{x_j}\|_2^2}{2r_j^2}\} \tag{8}$$

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 might be locally overfitted, hurting the generalization of the estimation. In this consideration, the optimal Parzen window size has been studied extensively in the literature [7, 36]. In practice, the optimal bandwidth may be determined by minimizing the *integrated squared error* (ISE), or the *mean integrated squared error* (MISE) [7]. Adaptive bandwidth is also proposed in the literature [36]. 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 follows:

$$r = \lambda D = \lambda \max_{\boldsymbol{x_k}}[\min_{\boldsymbol{x_l}}(\|\boldsymbol{x_k} - \boldsymbol{x_l}\|_2)] \tag{9}$$

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. 8 is tractable due to the limited number of negative samples and the use of dimensionality reduction techniques as discussed in Section 5, while the estimation accuracy is satisfactory.

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. This problem has been studied in a semi-supervised learning framework in the community, e.g., Szummer and Jaakkola formulated a regularization approach [33] to linking the marginal and the conditional in a general way to handle the partially labeled Data. Although reported effective, the approach is very computation intensive for large scale database. Similarly performed in the semi-supervised learning framework, we address this "scarce sample collection" problem from another perspective. We actually generate additional negative samples based on the kernel distributions for each semantic classes defined in Eq. 8. 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 *labeled* samples, and the generated samples as the *unlabeled* samples. To ensure that the number of the generated samples is sufficiently large, for each labeled negative sample in one relevance feedback iteration, we generate q additional unlabeled negative samples, where q is a parameter. To ensure a "fair sampling" to the kernel function in Eq. 8, the generation of the unlabeled samples follows a probability function defined by the following Gaussian pdf function:

$$p(\boldsymbol{y}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{\|\boldsymbol{y} - \boldsymbol{x_i}\|_2^2}{2\sigma^2}\} \tag{10}$$

where $\|\boldsymbol{y} - \boldsymbol{x_i}\|_2$ is the Euclidian distance between the unlabeled sample $\boldsymbol{y}$ and the corresponding labeled sample $\boldsymbol{x_i}$. σ is the standard deviation, which is set to be the average distance between two feature vectors in the labeled negative feature space defined as follows:

$$\sigma = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \|\boldsymbol{x_i} - \boldsymbol{x_j}\|_2 \tag{11}$$

Eq. 10 may be represented in another form as a function of the Euclidian distance, z, between the unlabeled sample $\boldsymbol{y}$ and the labeled sample $\boldsymbol{x_i}$:

$$p(z) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{z^2}{2\sigma^2}) \tag{12}$$

Based on Eq. 12, a vector is more likely to be selected as unlabeled negative sample if it is closer to a labeled negative sample than if it is farther away from a labeled negative sample. The probability density defined in Eq. 12 decays when the Euclidian distance to the labeled sample increases.

The following algorithm, called SAMPLING, is designed to perform the unlabeled sample selection based on Eq. 12. In the algorithm, NUM is the number of the unlabeled negative samples we intend to generate for each labeled sample. τ is a parameter to adjust the range centered at the positions with a distance d to the labeled sample, in which we select the unlabeled samples. In our experiments, τ is set to d/NUM.

```
input  : $S$, Labeled negative example set
output : $R$, Extended negative sample set
begin
    $R = \{\}$;
    for each labeled negative example $s \in S$ do
        $i = 0$;
        Create an array $A$ with $NUM + 1$ elements, each element $v_k = \frac{k}{NUM} * \frac{1}{\sqrt{2\pi\sigma}}$ , where $k \in [0, NUM]$;
        while $i < NUM$ do
            $d = 0$;
            Generate a random number $t = rand()$ in $[0, \frac{1}{\sqrt{2\pi\sigma}}]$;
            Search array A to find the k with $v_{k-1} < t \leq v_k$;
            Determine the distance d based on the function $v_k = p(d)$ defined in Eq. 12;
            Randomly sample the feature space until a feature V is found with a distance $s$ in $[d - \tau, d + \tau]$;
            $R = R \cup \{V\}$;
            $i = i + 1$;
        end
    end
    return $R$;
end
```

Algorithm 1: Algorithm SAMPLING

In essence, SAMPLING implements a roulette wheel sampling strategy [3] to select unlabeled samples. The vectors with smaller distances to a labeled sample have larger probabilities to be selected as unlabeled samples. However, those potential unlabeled samples farther away from a labeled sample are not completely eliminated from being selected, though their chances of being selected are small. This random selection principle is reasonable. Since in general the feature space is huge and has loose cluster presentation for the distribution of features, the feature vectors selected in this algorithm are statistically possible to be the true negative samples. With the extended number of the negative samples, the accuracy of the pdf estimation defined in Eq. 8 is significantly improved as shown in the experiments. In addition, the cumulative learning principle adopted in the estimation of the conditional pdf for the positive samples described in Section 3.1 is also applied in the estimation of the conditional pdf for the negative samples to further improve the estimation accuracy. While the estimation obtained from this method is still biased towards the given negative samples, it is noticeable that under the circumstances where there is no other information given, this is the best we can do. Compared to the existing approaches to handling unlabeled examples [32], the SAMPLING algorithm is simple and does not use the unlabeled examples explicitly, yet more efficient and scalable.

3.3. Determining the Prior Probabilities

In order to determine the relevancy and irrelevancy confidences defined as the posterior probabilities in Eqs. 1 and 2, 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 user subjective preferences. Thus, the relevancy and irrelevancy of an image vary to 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 also be estimated online in solving for the relevance feedback problem. In **BALAS**, we propose the following method to solve for the prior probabilities.

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

$$p(\text{Img}_i) = p(\text{Img}_i|R)P(R) + p(\text{Img}_i|I)P(I) \tag{13}$$

and for the query image we also have

$$P(R) + P(I) = 1 \tag{14}$$

Combining Eqs. 13 and 14, we immediately have:

$$P(R) = \frac{p(\text{Img}_i) - p(\text{Img}_i|I)}{p(\text{Img}_i|R) - p(\text{Img}_i|I)} \tag{15}$$

From Eq. 15, it is clear that since we have already developed methods to determine $p(\text{Img}_i|R)$, $p(\text{Img}_i|I)$, and $p(\text{Img}_i)$, the prior probability $P(R)$ can be uniquely determined immediately. Thus, $P(I)$ can also be immediately determined from Eq. 14. This reveals that for each given query image, the *overall* relevancy and irrelevancy of *all* the images in the image data set may be uniquely determined by *any individual* image Img_i in the image data set. In other words, any individual image Img_i in the image data set 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 data set. 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 data set 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 data set, we denote $P_i(R)$ as the prior probability determined in Eq. 15 using the individual image Img_i, i.e.,

$$P_i(R) = \frac{p(\text{Img}_i) - p(\text{Img}_i|I)}{p(\text{Img}_i|R) - p(\text{Img}_i|I)} \tag{16}$$

Thus, the final prior probability $P(R)$ is determined by an average of all the images in the data set, i.e.,

$$P(R) = \frac{1}{M}\sum_{i=1}^{M} P_i(R) \tag{17}$$

where M is the number of images in the database, as defined in Eqs. 1 and 2.

The prior probability $P(I)$ is determined accordingly from Eq. 14.

4 Ranking scheme

Given a query image, for each image Img_i in the data set, there is a corresponding relevancy confidence $P(R|\text{Img}_i)$, 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 data set to a query. On the other hand, for any CBIR system, there is always a feature-based distance measure used for similarity comparisons. 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 data set to a query. Consequently, in order to design a ranking scheme in image retrieval that "makes best sense", it is ideal to consider to integrate the subjective relevancy confidence and the objective distance measure together through taking advantage of the labeled sample image set to define an comprehensive ranking scheme.

Noting that the relevancy confidence and the feature-based distance measure are complementary to each other, we define a unified ranking scheme, called *Session Semantic Distance* (SD), to measure the relevance of any image Img_i within the image data set in terms of both the relevancy confidence $P(R|\text{Img}_i)$, the irrelevancy confidence $P(I|\text{Img}_j)$, and the feature-based distance measure $FD(\text{Img}_i)$ (e.g., Euclidean distance in the feature space between the query image and Img_i, $FD(\text{Img}_i) = \|\text{Qu} - \text{Img}_i\|_2$, where Qu is the query image).

The SD for any image $SD(\text{Img}_i)$ is defined using a modified form of the Rocchio's formula [25]. The Rocchio's formula for relevance feedback and feature expansion has proven to be one of the best iterative optimization technique in the field of information retrieval. It is frequently used to estimate "optimal query" $\boldsymbol{Q'}$ with an initial query $\boldsymbol{Q}$ in relevance feedback for sets of relevant documents D_R and irrelevant documents D_I given by the user. The formula is

$$\boldsymbol{Q'} = \alpha\boldsymbol{Q} + \beta(\frac{1}{N_R}\sum_{\boldsymbol{d_j}\in D_R}\boldsymbol{d_j}) - \gamma(\frac{1}{N_I}\sum_{\boldsymbol{d_j}\in D_I}\boldsymbol{d_j}) \tag{18}$$

where α, β, and γ are suitable constants; N_R and N_I are the numbers of documents in D_R and D_I, respectively. Based on the Rocchio's formula, $SD(\text{Img}_i)$ is defined as follows:

$$\begin{aligned} SD(\text{Img}_i) &= \log(1 + P(R|\text{Img}_i))FD(\text{Img}_i) \\ &+ \beta\{\frac{1}{N_R}\sum_{k\in D_R}[(1 + P(R|\text{Img}_k))U_{ik}]\} \\ &- \gamma\{\frac{1}{N_I}\sum_{k\in D_I}[(1 + P(I|\text{Img}_k))U_{ik}]\} \end{aligned} \tag{19}$$

where N_R and N_I are the sizes of the positive and negative labeled sample set D_R and D_I, respectively, in the relevance feedback, and U_{ik} is the feature-based distance between the images Img_i and Img_k (e.g., $U_{ik} = \|\text{Img}_i - \text{Img}_k\|_2$). We have replaced the first parameter α in the original 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 may be given to emphasize the different weights between the last two terms.

With this definition of the $SD(\text{Img}_i)$, the relevancy confidence of Img_i, the relevancy confidence of images in the labeled relevant set, the irrelevancy confidence of images in the labeled irrelevant set, and the objective feature distance measure are integrated in a unified approach. The (ir)relevancy confidences of images in the labeled 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 labeled relevant image set, and it is dissimilar to images in the labeled irrelevant image set in both relevancy confidence and feature-based distance measure; otherwise, its rank is low. Thus, the robustness and precision 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 and unified measure for semantic similarity.

5. 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 a prototype CBIR system.

For the evaluation purpose, we have implemented an image indexing and retrieval prototype system. In such a prototype system, many kinds of low-level features may be used to describe the content of images. In the current implementation, we use color moment, which is shown to be robust and effective [31]. 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 ranking metric. Since the objective is to test

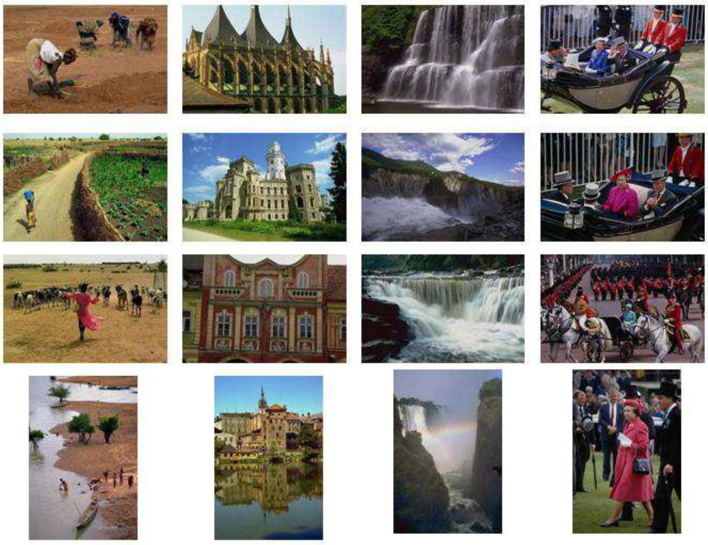


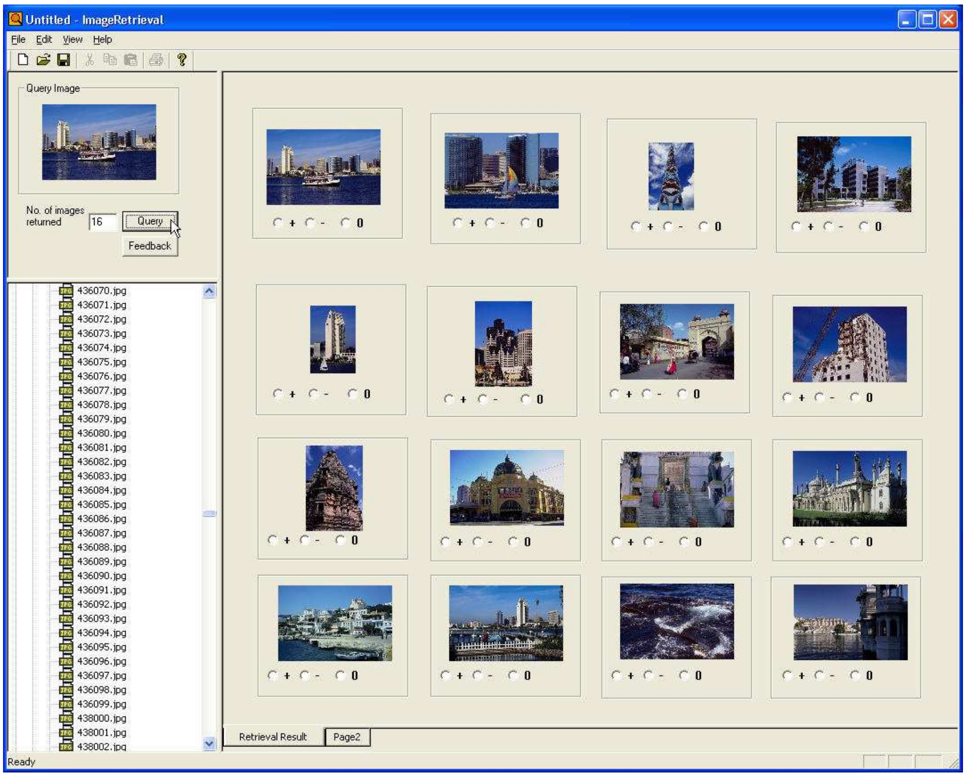
Figure 2: Sample images in the testing image database. The images in each column are assigned to one category. From left to right, the categories are “Africa rural area”, “historical building”, “waterfalls”, “British royal event”, respectively.

the relevance feedback learning method rather than to evaluate the effectiveness of the features, the features we used are not as sophisticated as those used in some existing CBIR systems [45, 8].

The following evaluations are performed on a general-purpose color image database containing 10,000 images from the COREL collection with 96 categories. All images are indexed as described above, and 1,500 images are randomly selected from all categories 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 is only used to ground-truth the evaluation, and we do not make use of this information in the indexing and retrieval procedures. Fig. 2 shows a few samples of the indexed images.

We have implemented the **BALAS** methodology on the prototype CBIR system, which we also call **BALAS** for the purpose of the discussion in this paper. Since user relevance feedback requires subjective feedback, we have invited 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 asked different users to run **BALAS** initially without the relevance feedback interaction, and then to place their relevance feedbacks after the initial retrievals. Fig. 3 is a screen shot of the **BALAS** system. Users check (+) or (-) radio buttons in the interface to place their relevance/irrelevance votes for each returned image. All the reported results are the averages of the whole group of users. The average time for each round of retrieval after the relevance input is about 2 seconds on a Pentium IV 2GHz computer with 512MB memory.

The feature-based distance FD describes the difference between image content details while the relevancy confidence depicts the perceptual aspects, i. e., semantic concepts, of images to users. To compare the capabilities of the two similarity measures as well as the session semantic distance (SD) we have defined to deliver an effective image retrieval, we test the three similarity metrics on images of 10 different semantic categories (each category has 100 images). These 10 categories



No. of images returned 16

Figure 3: A screen shot of **BALAS** system. The query image is "city skyline" and the first page of the retrieved images are shown. In this example, 12 of top 16 returned images are relevant.

Table 1: The precision in top N returned images based on FD, $P(R|\text{Img})$, and SD metrics for the 10 categories. The "Average" column for each distance metric is the average of the corresponding three left columns (Top 20, Top 30, and Top 50, respectively.)

	FD				$P(R\|\text{Img})$				SD			
Category	Top 20	Top 30	Top 50	Average	Top 20	Top 30	Top 50	Average	Top 20	Top 30	Top 50	Average
Africa	0.23	0.20	0.12	0.18	0.41	0.26	0.22	0.30	0.59	0.31	0.25	0.38
beach	0.32	0.25	0.20	0.26	0.62	0.55	0.48	0.55	0.60	0.58	0.50	0.56
buildings	0.38	0.30	0.24	0.31	0.51	0.31	0.29	0.37	0.49	0.38	0.30	0.39
buses	0.62	0.58	0.52	0.57	0.51	0.49	0.50	0.50	0.62	0.60	0.55	0.59
dinosaurs	0.80	0.72	0.69	0.74	0.73	0.65	0.58	0.65	0.81	0.75	0.70	0.75
elephants	0.58	0.50	0.43	0.50	0.61	0.59	0.50	0.57	0.60	0.55	0.51	0.55
flowers	0.81	0.72	0.70	0.74	0.90	0.81	0.82	0.65	0.91	0.86	0.80	0.83
horses	0.91	0.85	0.60	0.79	0.69	0.65	0.61	0.65	0.91	0.86	0.80	0.86
mountains	0.52	0.46	0.30	0.43	0.76	0.69	0.60	0.68	0.80	0.73	0.65	0.73
food	0.41	0.32	0.20	0.31	0.81	0.73	0.69	0.74	0.79	0.75	0.68	0.74

are selected such that no semantic overlap existing among them. The categories are $\{$Africa, beach, buildings, buses, dinosaurs, elephants, flowers, horses, mountains, food$\}$. In the comparison, FD is defined as Euclidean distance metric due to its effectiveness and simplicity. The average precision of top 20(30, 50) returned images based on FD, the relevancy confidence $P(R|\text{Img})$, and the SD for each category are recorded in Table 1. In the experiment, the number of the positive and negative samples are both 30 for the learning relevancy confidence. It shows that for the categories which has clear foreground/background and/or prominent objects, such as "dinosaur", "flowers", "horses", the feature-based distance performs well while the relevancy confidence is comparable. For the categories which do not have clear definitions of the objects in the images, such as "Africa", "beach", "food", the relevancy confidence is noticeably better for capturing the semantics. The integrated distance metric, SD, performs best or is comparable to the better one of FD and $P(R|\text{Img})$ in all the 10 categories.

To evaluate the systematic performance on the 10,000 image database, we have run the implemented CBIR system with **BALAS** for the 1,500 query image set with varied number of truncated top retrieved images and have plotted the curves of the average retrieval precision vs. the number of truncated top retrieved images (called scope). Fig. 4 shows the average precision-scope plot for the system with and without **BALAS** enabled. In other words, one test is based solely on the feature distance FD and the other test is based on the session semantic distance SD with different numbers of provided sample images. The notation (m/n) in the figure denotes the number of positive sample images vs. the number of negative sample images for the learning. In the evaluations, the sampling multiplier for the negative samples (NUM in the algorithm SAMPLING) is set to 5. From this figure, it is clear that the **BALAS** relevance feedback learning capability enhances the

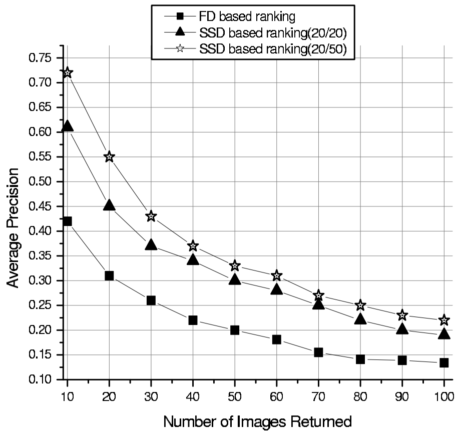


Figure 4: Average precisions vs. the numbers of the returned images with and without **BALAS** enabled

retrieval effectiveness substantially.

In order to compare the performance of **BALAS** with those of the state-of-the-art user relevance feedback methods in the literature, we have used the same image data set and the query set to compare **BALAS** with the relevance feedback method developed by Yong and Huang [26], which is a combination and improvement of its early version and MindReader [13] and represents the state-of-the-art relevance feedback research in the literature. Two versions of [26] are implemented. The first uses the color moments (called CM here) computed in the same way as described above and the other uses the correlogram (called CG here) [12]. For the latter, we consider RGB color space with a quantization of 64 total buckets. The distance set $D = \{1, 3, 5, 7\}$ is used for computing the autocorrelograms, which results in a feature vector of 256 dimensions. The overall comparison evaluations are documented in Fig. 5. 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 [26] performs noticeably better than **BALAS** while the CM version of [26] performs comparably with **BALAS**. After the second iteration, **BALAS** exhibits a significant improvement in performance over that of [26] in either of the two versions, and as the number of iterations increases, the improvement of the performance of **BALAS** over [26] appears to increase also. For example, after five iterations of the feedback, **BALAS** boosts its retrieval precision (14%) more than those of [26] using both CG (10.5%) and CM (9.5%), which means that **BALAS** has more potential. 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 better performance is expected from **BALAS**.

To evaluate the effectiveness of explicitly addressing the asymmetry issue of relevance feedback in CBIR, we compare **BALAS** with the SVM [39] classification method. SVM classifier adopts the two-class assumption and treats positive and negative samples equally, which is not valid in CBIR as we have discussed. In addition, for SVM there is no satisfied method to optimally select kernel function and its parameters yet except empirically testing. In the comparison experiment, the RBF kernel $K(x, y) = \exp^{-\|x-y\|^2/2\sigma^2}$ is used in the SVM classifier and the best σ is determined by using offline

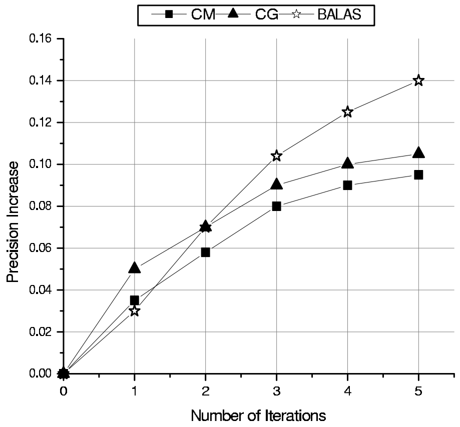


Figure 5: Retrieval precision comparison using relevance feedback between **BALAS** and CM and CG

cross-validation [5][1] The original SVM classifier only give 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 [39] is adopted for the SVM to determine the rank of retrieved images. A smaller query set composed of randomly selected 100 images from the 1,500 image query set is applied to **BALAS** and the SVM, respectively; the average precision in the top 100 images are recorded for different numbers of negative sample images with the number of positive sample images fixed. Fig. 6 shows the comparison. It indicates that **BALAS** outperforms the SVM consistently. The unsatisfactory performance of the SVM is 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 in the SVM learning, we have found that the positive part “spills over” freely into the part of the unlabeled areas in the feature space in 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.

6. Conclusions

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 with investigating 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, called **BALAS**, to empirical Bayesian learning to solve for this problem by explicitly exploiting the two unique characteristics. Different learning strategies are used for positive and negative sample collections, respectively. By defining the relevancy confidence as the relevant posterior probability, we have developed an integrated and unified ranking scheme in **BALAS**

[1] Because the trained SVM classifiers are dynamic (i.e., for each query image, a different SVM classifier is trained), so no validation data are available for online cross-validation. We used offline validation data for emulation purpose and hopefully to obtain a good σ for the online SVM classifiers.

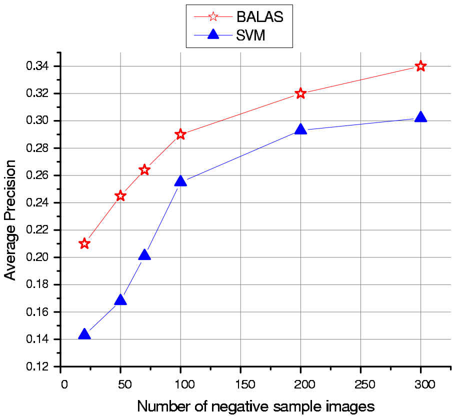


Figure 6: Comparison of BALAS and SVM on the average precision in the top 100 images returned. Number of positive sample images =20.

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 literature in capturing the overall retrieval semantics. We also show that the standard SVM classifier does not perform well in the relevance feedback of CBIR and **BALAS** outperforms the SVM in this problem.

References

[1] http://www.google.com/.

[2] A. D. Bimbo. *Visual Information Retrieval*. Morgan kaufmann Pub., San Francisco, CA, 1999.

[3] G. Blom. *Probability and Statistics: Theory and Applications*. Springer Verlag, London, U. K., 1989.

[4] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Trans. on PAMI*, 24(8):1026–1038, 2002.

[5] C. C. Chang and C. J. Lin. Libsvm: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001.

[6] Y. Chen and X. Z. aand Thomas Huang. One-class SVM for learning in image retrieval. In *Proceedings of the IEEE International Conference on Image Processing 2001*, Thessaloniki, Greece, October 2001.

[7] S.-T. Chiu. A comparative review of bandwidth selection for kernel density estimation. *Statistica Sinica*, 16:129–145, 1996.

[8] 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.

[9] W. R. Dillon and M. Goldstein. *Multivariate Analysis, Mehtods and Applications*. John Wiley and Sons, New York, 1984.

[10] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, New York, 1973.

[11] D. Geman and R. Moquet. A stochastic model for image retrieval. In *Proceedings of RFIA 2000*, Paris, France, February 2000.

[12] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlograms. In *IEEE Int'l Conf. Computer Vision and Pattern Recognition Proceedings*, Puerto Rico, 1997.

[13] Y. Ishikawa, R. Subramanya, and C. Faloutsos. Mindreader: Query databases through multiple examples. In *the 24th VLDB Conference Proceedings*, New York, 1998.

[14] F. Jing, M. Li, H.-J. Zhang, and B. Zhang. Support vector machines for region-based image retrieval. In *Proceedings of the IEEE International Conference on Multimedia & Expo*, Baltimore, MD, 2003.

[15] J. Laakdonen, M. Koskela, and E. Oja. Picsom: Self-organizing maps for content-based image retrieval. In *IJCNN'99 Proceedings*, Washington DC, 1999.

[16] Y. Liu, W. Rothfus, M.D., and T. Kanade. Content-based 3D neuroradiologic image retrieval: Preliminary results. In *IEEE International Workshop on Content-based Access of Image and Video Databases*, pages 91 – 100, January 1998. in conjunction with International Conference on Computer Vision (ICCV98).

[17] S. D. MacArthur, C. E. Brodley, and C. Shyu. Relevance feedback decision trees in content-based image retrieval. In *IEEE Workshop CBAIVL Proceedings*, South Carolina, June 2000.

[18] 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.

[19] C. Nastar, M. Mitschke, and C. Meilhac. Efficient query refinement for image retrieval. In *IEEE Conf. Computer Vision and Pattern Recognition Proceedings*, CA, June 1998. IEEE Comp. Soc.

[20] J. Peng, B. Bhanu, and S. Qing. Probabilistic feature relevance learning for content-based image retrieval. *Computer Vision and Pattern Recognition*, 75:150–164, 1999.

[21] R. W. Picard, T. P. Minka, and M. Szummer. Modeling user subjectivity in image libraries. In *IEEE International Conference on Image Processing Proceedigns*, Lausanne, Switzerland, September 1996.

[22] K. Porkaew, S. Mehrotra, and M. Ortega. Query reformulation for content based multimedia retrieval in MARS. In *IEEE Int'l Conf. Multimedia Computing and Systems*, June 1999.

[23] A. L. Ratan and W. E. L. Grimson. Training templates for scene classification using a few examples. In *IEEE Workshop on Content-Based Access of Image and Video Libraries Proceedings*, pages 90–97, 1997.

[24] B. D. Ripley and W. N. Venables. *Modern Applied Statistics with S*. Springer Verlag, New York, New York, 2002.

[25] J. J. J. Rocchio. 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.

[26] Y. Rui and T. S. Huang. Optimizing learning in image retrieval. In *IEEE Conf. Computer Vision and Pattern Recognition*, South Carolina, June 2000.

[27] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool in interactive content-based image retrieval. *IEEE Trans. on Circuits and Systems for Video Tech*, 8(5):644–655, September 1998.

[28] S. Santini and R. Jain. Integrated browsing and querying for image databases. *IEEE Multimedia*, 7:26–39, 2000.

[29] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York, 1986.

[30] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, October 2003.

[31] M. A. Stricker and M. Oregno. Similarity of color images. In *SPIE Storage and Retrieval of Still Image Video Databases IV*, volume 2420, pages 381–392, 1996.

[32] M. Szummer and T. Jaakkola. Kernel expansions with unlabeled examples. In *Proceedings of the Neural Information Processing Systems (NIPS) 2000*, 2000.

[33] M. Szummer and T. Jaakkola. Information regularization with partially labeled data. In *Proceedings of the Neural Information Processing Systems (NIPS) 2002*, Vancouver, Canada, December 2002.

[34] D. Tao and X. Tang. Random sampling based SVM for relevance feedback image retrieval. In *Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, June 2004.

[35] D. M. J. Tax and R. P. W. Duin. Combining one-class classifiers. In *Proceedings of the Second International Workshop Multiple Classifier systems*, pages 299–308, Berlin, Germany, 2001.

[36] G. R. Terrell and D. W. Scott. Variable kernel density estimation. *The Annals of Statistics*, 20:1236–1265, 1992.

[37] K. Tieu and P. Viola. Boosting image retrieval. In *IEEE Conf. Computer Vision and Pattern Recognitin Proceedings*, South Carolina, June 2000.

[38] S. Tong and E. Chan. Support vector machine active learning for image retrieval. In *ACM Multimedia 2001 Proceedings*, Ottawa, Canada, 2001.

[39] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[40] N. Vasconcelos and A. Lippman. Learning from user feedback in image retrieval. In S. A. Solla, T. K. Leen, and K. R. Muller, editors, *Adv. in Neural Information Processing Systems 12*. MIT Press, 2000.

[41] L. Wang, K. L. Chan, and Z. Zhang. Bootstrapping SVM active learning by incorporating unlabelled images for image retrieval. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Madison, WI, 2003.

[42] M. E. J. Wood, N. W. Campbell, and B. T. Thomas. Iterative refinement by relevance feedback in content-based digital image retrieval. In *ACM Multimedia 98 Proceedings*, Bristol, UK, September 1998.

[43] 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.

[44] R. Yan, A. Hauptmann, and R. Jin. Negative pseudo-relevance feedback in content-based video retrieval. In *Proceedings of the ACM Multimedia 2003*, Berkeley, CA, November 2003.

[45] R. Zhang and Z. Zhang. Addressing CBIR efficiency, effectiveness, and retrieval subjectivity simultaneously. In *the 5th ACM Int'l Workshop on Multimedia Information Retrieval*, Berkeley, CA, November 2003. in conjunction with ACM Multimedia (ACM MM) 2003.

[46] R. Zhang and Z. Zhang. Hidden semantic concept discovery in region based image retrieval. In *IEEE International Conference on Computer Vision and Pattern Recogntion (CVPR) 2004*, Washington, DC, June 2004.

[47] 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.