# Similarity-Based Online Feature Selection in Content-Based Image Retrieval

Wei Jiang, *Student Member, IEEE*, Guihua Er, *Member, IEEE*, Qionghai Dai, *Senior Member, IEEE*, and Jinwei Gu, *Student Member, IEEE*

***Abstract*—Content-based image retrieval (CBIR) has been more and more important in the last decade, and the gap between high-level semantic concepts and low-level visual features hinders further performance improvement. The problem of online feature selection is critical to really bridge this gap. In this paper, we investigate online feature selection in the relevance feedback learning process to improve the retrieval performance of the region-based image retrieval system. Our contributions are mainly in three areas. 1) A novel feature selection criterion is proposed, which is based on the psychological similarity between the positive and negative training sets. 2) An effective online feature selection algorithm is implemented in a boosting manner to select the most representative features for the current query concept and combine classifiers constructed over the selected features to retrieve images. 3) To apply the proposed feature selection method in region-based image retrieval systems, we propose a novel region-based representation to describe images in a uniform feature space with real-valued fuzzy features. Our system is suitable for online relevance feedback learning in CBIR by meeting the three requirements: learning with small size training set, the intrinsic asymmetry property of training samples, and the fast response requirement. Extensive experiments, including comparisons with many state-of-the-arts, show the effectiveness of our algorithm in improving the retrieval performance and saving the processing time.**

***Index Terms*—Boosting, online feature selection, region-based image retrieval, relevance feedback.**

## I. Introduction

CONTENT-BASED image retrieval (CBIR) has been largely explored in the last decade. In the CBIR context, an image is represented by a set of low-level visual features, which have no direct correlation with high-level semantic concepts, and the gap between high-level concepts and low-level features is the major difficulty that hinders further development of CBIR systems [24]. *Relevance Feedback* and *region-based image retrieval* (RBIR) have been proposed to bridge this gap. The relevance feedback mechanism is an iterative learning process, which is generally treated as online supervised learning [12], [14], [22], [28], [31]. During each iteration, the user labels some images to be "relevant" or "irrelevant" to his query concept, and the system uses the labeled images as training samples to successively refine the learning mechanism and gives better retrieval results in the next iteration. The RBIR approaches [3], [4], [6], [15], [21], [36] segment images into several regions and retrieve images with low-level features extracted based on these regions. Since *region-based features* represent images in the object level and better fit the human perception than the *global low-level features* extracted from the entire image, better performance can be expected for RBIR methods.

In all CBIR systems, the online learning process must tackle a fundamental problem: which features are more representative for explaining the current query concept than the others. This refers to the problem of *online feature selection*, which is the issue we mainly address in this paper. Compared with other machine learning problems, CBIR online learning has three challenges. 1) Small size of the training set: The training samples are the labeled images from the user during each query session, which are very few compared with the feature dimensionality and the size of the database. The CBIR learning algorithm usually encounters severe problems due to the curse of dimensionality. 2) Intrinsic asymmetry: The images labeled to be "relevant" during a query session share some common semantic cues, while the "irrelevant" images are different from the "relevant" ones in different ways. Thus, the "relevant" images are more important for the system to grasp the query concept. This asymmetry requirement makes it necessary to treat the "relevant" and "irrelevant" sets unequally with an emphasis on the "relevant" one. 3) Fast response requirement: The system should give out the retrieval results within a tolerable amount of time.

### A. Previous Works

Due to the above three special requirements, most classical feature selection criterions, such as the distribution-based approaches [e.g., mutual information maximization (MMI) method [29] and Kullback–Leibler divergence (K–LD) method [20]] and the conventional boosting approach [27], are not suitable for CBIR online learning. Since the few training samples are usually not representative of the whole dataset, it is difficult for the system to well estimate the samples' distribution. For the same reason, the conventional boosting method will not perform well because of the poor generalization ability due to the training-error-based feature selection criterion. Furthermore, the asymmetry requirement is not considered in these feature selection methods. There is another kind of methods for online feature selection in CBIR, which is called the *discriminant analysis* (DA) approach. *multiple discriminant analysis* (MDA) method [33], *biased discriminant analysis* (BDA) method [37], and *symmetric maximized minimal distance in subspace* (SMMS) method [35] are three

Manuscript received October 3, 2004; revised March 31, 2005. This work was supported in part by the Distinguished Young Scholars of NSFC (60525111) and in part by the key project of NSFC (60432030). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jianying Hu.

The authors are with Tsinghua University, Beijing 100084, China (e-mail: jiangwei98@mails.tsinghua.edu.cn).

Digital Object Identifier 10.1109/TIP.2005.863105

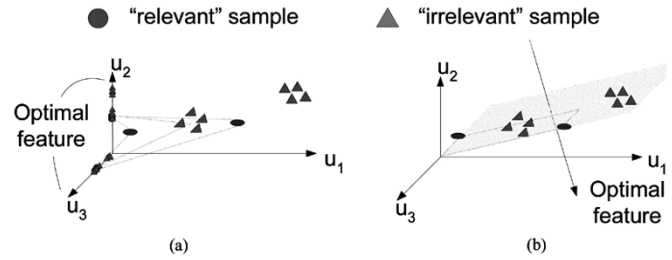


Fig. 1. Toy problem where "relevant" images gather to more than one clusters. The genuine optimal feature axis is $u_1$. (a) By MDA and BDA, $u_2$ or $u_3$ is selected as the first optimal feature axis. (b) SMMS selects a feature axis perpendicular to the subspace spanned by the "relevant" samples.

typical approaches. They generalize linear discriminant analysis (LDA) [11], and assume that the "relevant" images group together as one cluster. To meet the asymmetry requirement they do not assume the one-cluster distribution for "irrelevant" images. MDA assumes that each "irrelevant" image is from a different class, while BDA assumes that the "irrelevant" images come from an uncertain number of classes. From the aspect of computation, MDA and BDA minimize the covariance of the "relevant" set (the within class scatter $S_w$) over the between class distance (the between class scatter $S_b$). SMMS selects the feature subspace which is perpendicular to the subspace spanned by the "relevant" samples.

However, when the "relevant" images actually gather to multiple clusters, the effectiveness of these DA methods will suffer. Fig. 1 gives an example of toy problem for such cases. In the example, $u_1$ is the genuine optimal feature axis. When projected to feature axis $u_2$ or $u_3$, the "relevant" images have a very small $S_w$. Thus, in most cases, $u_2$ or $u_3$ will be selected as the best feature axis by MDA and BDA no matter how the "irrelevant" images distribute. When the projections of the "irrelevant" images on $u_2$ or $u_3$ are close to those of the "relevant" ones, the selected feature axis is not discriminative [Fig. 1(a)]. As for SMMS, when the "irrelevant" samples distribute in the subspace spanned by the "relevant" ones, the selected feature is not discriminative [Fig. 1(b)]. In practical CBIR systems, the similar situations often occur, especially when the "relevant" images are very few and are scattered in the feature space. Fig. 2 gives an example of real-image dataset from Corel gallery [7]. Images from the "Autumn" (red $+$), "Bonsai" (green $*$), and "Bird" (blue $\circ$) categories are represented in the 2-D feature space spanned by the top two principle components of a dataset with 10 000 images (see Section V). Assume that the user wants images from the "Bird" category, i.e., blue $\circ$ are "relevant," red $+$ and green $*$ are "irrelevant." The histograms of the "relevant" and "irrelevant" samples along the selected feature axis are given, and the Bayes error rate is used to measure the separability of "relevant" and "irrelevant" sets. We can see that, along the first optimal feature axes selected by MDA and BDA, the "relevant" samples can not be separated from the "irrelevant" ones, while along the optimal feature axis selected by our method, the "relevant" samples can be much better distinguished.

One way to solve the problem of one-cluster assumption is to map training samples to a higher dimensional space with the kernel method, so that after mapping "relevant" samples may be well gathered into one cluster. BiasMap (kernel BDA)

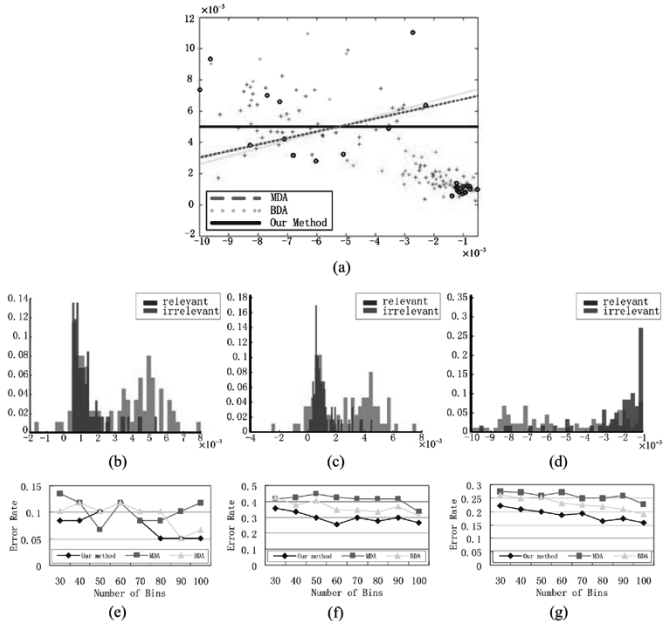


Fig. 2. Example of real images. (a) The first optimal feature axes selected by MDA, BDA, and our method. (b)–(d) Distributions of the "relevant" and "irrelevant" samples along the optimal feature axes selected by MDA, BDA, and our method, respectively (SMMS can not be used because the "relevant" samples cannot be projected into one point along a 1-dimensional feature axis [35]). (e)–(g) Bayes error rates for our method, MDA and BDA with different numbers of histogram bins. Along the feature axis selected by our method, the "relevant" and "irrelevant" samples are best separated.

[37] and SVM-based approaches [5], [13], [28], [32], [34] are two typical methods. The SVM-based methods can be classified into two categories: the traditional SVM approaches [13], [28], [32], [34] which treat CBIR online learning as strict two-class classification problem; and the one-class SVM approach [5] which uses only "relevant" samples with some regularization in training and testing. It is generally accepted that currently in the CBIR context, the SVM-based techniques yield almost the most promising performances. There are, however, several potential problems. First, two-class SVM approaches ignore the intrinsic asymmetry requirement of CBIR online learning, and information of "irrelevant" samples are not fully used in the one-class SVM approach. BiasMap often suffers the matrix singularity problem, and the regularization method [11] which adds small quantities to the diagonal of the singular matrices may lead to ill-posed problem. In addition, it is well known that choosing proper kernel functions and parameters for a specific real database remains challenging, and the number of support vectors that compose the decision function increases dramatically when the decision manifold becomes complicated [20]. Moreover, the over fitting problem may be more severe, that is, training samples may be too few to train a good classifier in a high dimensional space [35].

As for RBIR, as far as the authors know, little work has been done to tackle the problem of online feature selection over the region-based features. This is partly due to the difficulty to represent images in a uniform feature space because different images have different numbers of regions. Instead of online learning or feature selection, most RBIR systems directly calculate the region-to-region similarity or the image-to-image similarity to retrieve images [3], [4], [6], [21]. For example,

the *unified feature matching* (UFM) measurement has been proposed to calculate the image-to-image similarity based on region-based features [6]. Since no learning mechanism is used, the retrieval performance of the straightforward similarity-based methods is usually unsatisfactory. Recently, a binary region-based feature representation has been proposed in [15], [36], which extracts a codebook from the original region-based features and uses the indexes of the codebook to represent images. Online learning can be applied based on the binary features. However, much valuable low-level information is lost in the binary codebook representation, and the effectiveness of these approaches is limited. In addition, online feature selection, which is an important issue for all CBIR systems, is not considered in these approaches.

### B. Overview of Our Approach

In this paper, we address the issue of online feature selection in relevance feedback learning for region-based image retrieval systems. Our contributions are mainly in three areas. 1) A novel feature selection criterion is proposed which accounts for the asymmetry requirement of training samples and does not need distribution estimation or one-cluster assumption for "relevant" images. Since the goal of feature selection is to find the optimal feature subspace where the "relevant" and "irrelevant" image sets can be best separated, the similarity between these two sets is used as the feature selection criterion. Specifically, the *fuzzy feature contrast model* (FFCM) [23] is adopted to calculate the asymmetric similarity between images, and the UFM measurement [6] is used to calculate the similarity between the "relevant" and "irrelevant" image sets. 2) We implement an effective online feature selection algorithm in a boosting manner, which incorporates the proposed feature selection criterion with the Real AdaBoost framework [10] to select the optimal features and combine the incrementally learned classifiers over the selected features into a strong ensemble classifier. 3) To introduce online learning and feature selection into RBIR systems, a novel region-based representation is proposed based on a generative model, which describes images in a uniform feature space with real-valued fuzzy features. We systematically carry out experiments over 10 000 images, including comparisons with classical RBIR approaches, DA feature selection approaches, and kernel-based methods. The results show that the proposed system can significantly improve the retrieval performance and save the processing time.

The rest of this paper is organized as follows. In Section II, we propose a novel similarity-based feature selection criterion. Section III presents the online feature selection algorithm which incorporates the proposed feature selection criterion with the boosting framework. A new region-based image representation is developed in Section IV to apply our feature selection algorithm in the RBIR system. The experimental results are given in Section V. We conclude our work in Section VI.

## II. Criterion for Online Feature Selection

Let $\vec{\mathbf{x}} = [f_1(\mathbf{x}), \ldots, f_d(\mathbf{x})]^T$ denote the feature vector of an image $\mathbf{x}$ represented in a $d$-dimensional feature space $\mathcal{F} = \{f_k, k = 1, \ldots, d\}$. We have a "relevant" image set $\mathcal{R} = \{\mathbf{x}_i^{\mathcal{R}}, i = 1, \ldots, m\}$ of size $m$, and an "irrelevant" image set $\mathcal{IR} = \{\mathbf{x}_j^{\mathcal{IR}}, j = 1, \ldots, n\}$ of size $n$. $\mathcal{W}^{\mathcal{R}} = \{w(\mathbf{x}_i^{\mathcal{R}}), i = 1, \ldots, m\}$ and $\mathcal{W}^{\mathcal{IR}} = \{w(\mathbf{x}_j^{\mathcal{IR}}), j = 1, \ldots, n\}$ are the sample weights. Let $\mathcal{R}_k = \{f_k(\mathbf{x}_i^{\mathcal{R}}), i = 1, \ldots, m\}$ and $\mathcal{IR}_k = \{f_k(\mathbf{x}_j^{\mathcal{IR}}), j = 1, \ldots, n\}$ represent the projections of $\mathcal{R}$ and $\mathcal{IR}$, respectively, along the $k$th feature axis $f_k$. The similarity between $\mathcal{R}_k$ and $\mathcal{IR}_k$ is used as the criterion to determine the discriminative degree of $f_k$.

### A. UFM Similarity Between Image Sets

To calculate the similarity between "relevant" and "irrelevant" image sets, we introduce the UFM measurement [6] as follows.

Assume that $\tilde{\mathcal{A}}_i$ and $\tilde{\mathcal{B}}_j$ are two fuzzy sets. The similarity between them, denoted by $s(\tilde{\mathcal{A}}_i, \tilde{\mathcal{B}}_j)$, can be defined in many ways [6]. Let $\mathcal{A}$ and $\mathcal{B}$ be two collections of fuzzy sets, $\mathcal{A} = \{\tilde{\mathcal{A}}_i, i = 1, \ldots, p\}, \mathcal{B} = \{\tilde{\mathcal{B}}_j, j = 1, \ldots, q\}$. In [6] the similarity for $\tilde{\mathcal{A}}_i$ and $\mathcal{B}$ and that for $\tilde{\mathcal{B}}_j$ and $\mathcal{A}$ are given by

$$l_i^{\mathcal{B}} = s\left(\tilde{\mathcal{A}}_i, \bigcup_{j=1}^{q} \tilde{\mathcal{B}}_j\right)$$

$$l_j^{\mathcal{A}} = s\left(\tilde{\mathcal{B}}_j, \bigcup_{i=1}^{p} \tilde{\mathcal{A}}_i\right)$$

and the similarity between $\mathcal{A}$ and $\mathcal{B}$ is given by

$$\mathcal{S}(\mathcal{A}, \mathcal{B}) = \sum_{i=1}^{p} w(\tilde{\mathcal{A}}_i) l_i^{\mathcal{B}} + \sum_{j=1}^{q} w(\tilde{\mathcal{B}}_j) l_j^{\mathcal{A}} \tag{1}$$

where $w(\tilde{\mathcal{A}}_i)$ and $w(\tilde{\mathcal{B}}_j)$ are the saliency memberships of $\tilde{\mathcal{A}}_i$ and $\tilde{\mathcal{B}}_j$ in $\mathcal{A}$ and $\mathcal{B}$ respectively. The operator $\bigcup$ can take many forms in fuzzy set theory [38]. The conventional maximization operator is adopted in [6], i.e.,

$$l_i^{\mathcal{B}} = \max_{j=1,\ldots,q} s(\tilde{\mathcal{A}}_i, \tilde{\mathcal{B}}_j)$$

$$l_j^{\mathcal{A}} = \max_{i=1,\ldots,p} s(\tilde{\mathcal{B}}_j, \tilde{\mathcal{A}}_i). \tag{2}$$

In our RBIR system, as will be discussed in Section IV-C, each image $\mathbf{x}$ is represented by a set of fuzzy features $\vec{\mathbf{x}} = [f_1(\mathbf{x}), \ldots, f_d(\mathbf{x})]^T$. We treat each image as a fuzzy singleton (a fuzzy set with a single element), and treat $\mathcal{R} = \{\mathbf{x}_i^{\mathcal{R}}, i = 1, \ldots, m\}$ and $\mathcal{IR} = \{\mathbf{x}_j^{\mathcal{IR}}, j = 1, \ldots, n\}$ as collections of fuzzy sets consisting of fuzzy singletons. Along each feature axis $f_k, \mathcal{R}_k = \{f_k(\mathbf{x}_i^{\mathcal{R}}), i = 1, \ldots, m\}$ and $\mathcal{IR}_k = \{f_k(\mathbf{x}_j^{\mathcal{IR}}), j = 1, \ldots, n\}$ can be treated as two collections of singletons with one-dimensional fuzzy features. The similarity between $\mathcal{R}_k$ and $\mathcal{IR}_k$ can be calculated based on the UFM measurement with (1) and (2) as

$$\mathcal{S}(\mathcal{R}_k, \mathcal{IR}_k) = \sum_{i=1}^{m} w\left(\mathbf{x}_i^{\mathcal{R}}\right) \max_{j=1,\ldots,n} s\left(f_k\left(\mathbf{x}_i^{\mathcal{R}}\right), f_k\left(\mathbf{x}_j^{\mathcal{IR}}\right)\right) + \sum_{j=1}^{n} w\left(\mathbf{x}_j^{\mathcal{IR}}\right) \max_{i=1,\ldots,m} s\left(f_k\left(\mathbf{x}_j^{\mathcal{IR}}\right), f_k\left(\mathbf{x}_i^{\mathcal{R}}\right)\right). \tag{3}$$

$s(f_k(\mathbf{x}_i^{\mathcal{R}}), f_k(\mathbf{x}_j^{\mathcal{IR}}))$ is the similarity between images $\mathbf{x}_i^{\mathcal{R}}$ and $\mathbf{x}_j^{\mathcal{IR}}$ along the feature axis $f_k$. To meet the asymmetry requirement and to make an emphasis on the features of the "relevant" set $\mathcal{R}$, the asymmetric FFCM measurement [23] is adopted to calculate $s(f_k(\mathbf{x}_i^{\mathcal{R}}), f_k(\mathbf{x}_j^{\mathcal{IR}}))$.

### B. FFCM Similarity Between Images

The famous *feature contrast model* (FCM) has been proposed by Tversky as a psychological similarity measurement between two objects [30]. Let $a, b$ be two stimuli, instead of considering them as points in a metric space, FCM represents them by binary feature sets, denoted by $\mathbf{A}, \mathbf{B}$, and defines their similarity as

$$\tilde{S}(a,b) = f(\mathbf{A} \cap \mathbf{B}) - \alpha f(\mathbf{A} - \mathbf{B}) - \beta f(\mathbf{B} - \mathbf{A}) \quad (4)$$

where $\mathbf{A} \cap \mathbf{B}$ is the *common feature* contained by $a$ and $b$, and $\mathbf{A} - \mathbf{B}$ is the *distinctive feature* contained by $a$ but not $b$. $f(\cdot)$ is a salient function, whose value increases monotonically when the variable in the bracket increases. If $\alpha \neq \beta$, we emphasize the features in different stimuli unequally. Since the exposition of the similarity is a generalization process, the similarity should in principle depend on both the common and distinctive features of objects [25]. FCM successfully measures the perception similarity between two objects.

In [23] FCM is generalized to the FFCM and is introduced to the computer vision field to represent the similarity between human faces or between textures. Assume that $a$ and $b$ are represented by two fuzzy feature vectors $\mathbf{A} = [A_1, \ldots, A_d]$ and $\mathbf{B} = [B_1, \ldots, B_d]$. FFCM defines operators $\cap$ and $-$ in a traditional way as

$$\mathbf{A} \cap \mathbf{B} = [\min\{A_1, B_1\}, \ldots, \min\{A_d, B_d\}]$$
$$\mathbf{A} - \mathbf{B} = [\max\{A_1 - B_1, 0\}, \ldots, \max\{A_d - B_d, 0\}]$$

and calculates the similarity between $a$ and $b$ by generalizing (4) as

$$\tilde{S}(a,b) = \sum_{i=1}^{d} (\min\{A_i, B_i\} - \alpha \max\{A_i - B_i, 0\} - \beta \max\{B_i - A_i, 0\}). \quad (5)$$

The asymmetry direction depends on the relative "saliency" of the stimuli: if features in $a$ are more salient, $\alpha \geq \beta$.

In our CBIR context, as shown in (3), singletons $f_k(\mathbf{x}_i^{\mathcal{R}})$ and $f_k(\mathbf{x}_j^{\mathcal{IR}})$ can be treated as 1-dimensional fuzzy feature vectors. Thus, we can define $s(f_k(\mathbf{x}_i^{\mathcal{R}}), f_k(\mathbf{x}_j^{\mathcal{IR}}))$ based on (5) as follows:

$$\begin{aligned} & s\left(f_k\left(\mathbf{x}_i^{\mathcal{R}}\right), f_k\left(\mathbf{x}_j^{\mathcal{IR}}\right)\right) \\ &= \tilde{S}\left(f_k\left(\mathbf{x}_i^{\mathcal{R}}\right), f_k\left(\mathbf{x}_j^{\mathcal{IR}}\right)\right) \\ &= \min\{f_k\left(\mathbf{x}_i^{\mathcal{R}}\right), f_k\left(\mathbf{x}_j^{\mathcal{IR}}\right)\} \\ &\quad - \alpha \max\{f_k\left(\mathbf{x}_i^{\mathcal{R}}\right) - f_k\left(\mathbf{x}_j^{\mathcal{IR}}\right), 0\} \\ &\quad - \beta \max\{f_k\left(\mathbf{x}_j^{\mathcal{IR}}\right) - f_k\left(\mathbf{x}_i^{\mathcal{R}}\right), 0\}. \end{aligned} \quad (6)$$

As pointed out before, to grasp the query concept, the features shared by images in $\mathcal{R}$ but not contained by images in $\mathcal{IR}$ are more important than the features shared by images in $\mathcal{IR}$ but not contained by images in $\mathcal{R}$. We use $\alpha \geq \beta$ to meet this asymmetry requirement. The effect of the parameters $\alpha$ and $\beta$ will be investigated in later experiments.

With (3) and (6), the similarity between $\mathcal{R}_k$ and $\mathcal{IR}_k$ can be obtained. The smaller the $\mathcal{S}(\mathcal{R}_k, \mathcal{IR}_k)$, the better the feature axis $f_k$ in discriminating the "relevant" and "irrelevant" sets. Thus, the feature selection criterion can be given as follows:

$$f^* = \arg\min_k \mathcal{S}(\mathcal{R}_k, \mathcal{IR}_k). \quad (7)$$

This feature selection criterion is suitable for CBIR online learning, which accounts for the asymmetry requirement of training samples and does not need one-cluster assumption for the "relevant" image set. It is also more suitable for learning with small sample set in CBIR than the traditional training-error-based feature selection criterion. As for the toy problem described in Fig. 1, the genuine optimal feature axis, $u_1$, can be selected based on our feature selection criterion, because along $u_1$ the system has the smallest similarity between the projected "relevant" and "irrelevant" sets. For the problem with real images shown in Fig. 2, our method can also select the feature axis better than MDA and BDA. Note that, for cases where the "relevant" samples are well gathered into one Gaussian cluster and are far from the "irrelevant" samples, our feature selection criterion can obtain similar results as MDA, BDA, and SMMS. This is because the intuitive motivation of (3) and (6) is similar as the basic idea of the other methods: the system will find the feature axis which makes "relevant" and "irrelevant" sets most dissimilar.

## III. Boosting Feature Selection

Given a set of optimal feature axes, the boosting mechanism gives an effective way for selecting a new added feature axis by re-weighting training samples, and combines the classifiers constructed over the incrementally learned features into an ensemble classifier with a decreased training error. The AdaBoost algorithm [9] has been introduced into CBIR by [27]. As discussed in Section I, since the small size of the training set makes the training error of the weak classifiers unrepresentative for the generalization error (the testing error), especially for large scale databases, the conventional AdaBoost approach which selects features based on the training error is usually over fitted. In this section, we develop a novel boosting feature selection algorithm. Instead of the training error, the feature selection criterion proposed in above section is used, and the Real AdaBoost framework [10] is adopted to implement the CBIR online learning algorithm.

### A. Similarity-Based Boosting Feature Selection

Assume that $t-1$ features $f^i \in \mathcal{F}, i = 1, \ldots, t-1$ have already been selected during the previous $t-1$ boosting iterations, and $t-1$ weak classifiers have been constructed, one for each selected feature. The output of the $i$th classifier is the class label estimation $h^i(\mathbf{x})$ and the class probability estimation $p(h^i(\mathbf{x}) = 1|\mathbf{x})$ for each image $\mathbf{x}$. The current sample weights are $\mathcal{W}^{t-1}$, where for each training sample $\mathbf{x}$, the weight is $w^{t-1}(\mathbf{x})$. Assume that $y_{\mathbf{x}}$ is the true label of $\mathbf{x}$. To select a new feature which is optimal in discriminating the "relevant" set $\mathcal{R}$ and "irrelevant" set $\mathcal{IR}$, the original Real AdaBoost algorithm [10] uses the process described in Fig. 3.

To develop an effective feature selection algorithm suitable for CBIR online learning, we select the optimal feature $f^t$ based

**Input:** feature pool $\mathcal{F}=\{f_k, k=1,\ldots,d\}$, training set $\mathcal{R}$, $\mathcal{IR}$

1) Initialization: set $w^0(\mathbf{x})=\frac{1}{|\mathcal{R}|+|\mathcal{IR}|}$ for all $\mathbf{x}\in\mathcal{R}$ or $\mathcal{IR}$;
2) Repeat for $t=1,\ldots,T$:
   a) For $k=1,\ldots,d$
      - Construct a weak classifier along $f_k$ with weight $\mathcal{W}^{t-1}$, and get the class label estimation $h_k(\mathbf{x})$ and class probability estimation $p(h_k(\mathbf{x})=1|\mathbf{x})$;
      - Get training error $\epsilon_k=\sum_{\mathbf{x}} w^{t-1}(\mathbf{x})|y_{\mathbf{x}}-h_k(\mathbf{x})|$;

   b) Get $f^t$ along which the classifier has the smallest $\epsilon^t=\min_{k=1,\ldots,d}\epsilon_k$, and set $h^t(\mathbf{x})$ and $p(h^t(\mathbf{x})=1|\mathbf{x})$ as the corresponding label and probability estimations;
   c) Set $q^t(\mathbf{x})=\frac{1}{2}\log\frac{p(h^t(\mathbf{x})=1|\mathbf{x})}{1-p(h^t(\mathbf{x})=1|\mathbf{x})}$;
   d) Update $w^t(\mathbf{x}) = w^{t-1}(\mathbf{x})\exp(-y_{\mathbf{x}}q^t(\mathbf{x}))$, and renormalize $w^t(\mathbf{x})$ over $\mathbf{x}$;

**Output:** ensemble label estimation $H(\mathbf{x})=\text{sign}\left[\sum_{t=1}^{T}q^t(\mathbf{x})\right]$

Fig. 3. Original Real AdaBoost algorithm [10].

**Input:** feature pool $\mathcal{F}=\{f_k, k=1,\ldots,d\}$, training set $\mathcal{R}$, $\mathcal{IR}$

1) Initialization: set $w^0(\mathbf{x})=\frac{1}{2|\mathcal{R}|}$ or $\frac{1}{2|\mathcal{IR}|}$ for $\mathbf{x}\in\mathcal{R}$ or $\mathcal{IR}$, respectively;
2) Repeat for $t=1,\ldots,T$
   a) For $k=1,\ldots,d$
      - Calculate $\mathcal{S}(\mathcal{R}_k,\mathcal{IR}_k)$ by Equations (3, 6);

   b) Get $f^t$ : $f^*=\arg\min_k \mathcal{S}(\mathcal{R}_k,\mathcal{IR}_k)$;
   c) Construct an FKNN classifier along $f^t$, and get $p(h^t(\mathbf{x})=1|\mathbf{x})$ by Equation (8);
   d) Set $q^t(\mathbf{x})=\frac{1}{2}\log\frac{p(h^t(\mathbf{x})=1|\mathbf{x})}{1-p(h^t(\mathbf{x})=1|\mathbf{x})}$;
   e) Update $w^t(\mathbf{x}) = w^{t-1}(\mathbf{x})\exp(-y_{\mathbf{x}}q^t(\mathbf{x}))$, and renormalize $w^t(\mathbf{x})$ over $\mathbf{x}$;

**Output:** the ensemble value $V^T(\mathbf{x})$ by Equation (9)

Fig. 4. Our feature selection algorithm.

on the similarity between $\mathcal{R}$ and $\mathcal{IR}$ along each feature axis, instead of the training error. Furthermore, the target of the CBIR relevance feedback learning is to output a "relevant" degree for every image in the database, according to which the images can be ranked to give out the retrieval results. Thus, instead of the ensemble class label estimation, a set of *soft* ensemble class probabilities about the "relevant" degree of the images are preferred.

Our feature selection algorithm is given in Fig. 4. We project the training samples to each feature axis $f_k$ as $\mathcal{R}_k$ and $\mathcal{IR}_k$, and then calculate the similarity $\mathcal{S}(\mathcal{R}_k,\mathcal{IR}_k)$ with the current weights $\mathcal{W}^{t-1}$ according to (3) and (6). The optimal feature along which the two image sets have the smallest similarity is selected as $f^t$ by (7). A weak classifier is constructed along $f^t$, and a set of class label estimations $h^t(\mathbf{x})$ and class probability estimations $p(h^t(\mathbf{x})=1|\mathbf{x})$ are given. Then samples are re-weighted by the same method as Real AdaBoost. Specifically, the *fuzzy K-nearest neighbor* (FKNN) classifier [18] is adopted as the weak classifier for each selected feature, which does not have the one-cluster assumption for the "relevant" or "irrelevant" set. For a FKNN classifier over feature axis $f^t$, it estimates the memberships $\phi^{\mathcal{R}}(\mathbf{x})$ and $\phi^{\mathcal{IR}}(\mathbf{x})$ of each image $\mathbf{x}$ to the "relevant" set $\mathcal{R}$ and "irrelevant" set $\mathcal{IR}$, respectively, as

$$\phi^{\mathcal{R}}(\mathbf{x}) = \frac{\sum_{j=1}^{K}\left[w^{t-1}(\tilde{\mathbf{x}}_j)(f^t(\mathbf{x})-f^t(\tilde{\mathbf{x}}_j))^{-2}\cdot\delta(\tilde{\mathbf{x}}_j\in\mathcal{R})\right]}{\sum_{j=1}^{K}(f^t(\mathbf{x})-f^t(\tilde{\mathbf{x}}_j))^{-2}}$$

$$\phi^{\mathcal{IR}}(\mathbf{x}) = \frac{\sum_{j=1}^{K}\left[w^{t-1}(\tilde{\mathbf{x}}_j)(f^t(\mathbf{x})-f^t(\tilde{\mathbf{x}}_j))^{-2}\cdot\delta(\tilde{\mathbf{x}}_j\in\mathcal{IR})\right]}{\sum_{j=1}^{K}(f^t(\mathbf{x})-f^t(\tilde{\mathbf{x}}_j))^{-2}}$$

where $\tilde{\mathbf{x}}_j, j=1,\ldots,K$ are the $K$ nearest neighbors of $\mathbf{x}$ in the training set along $f^t$, and $\delta(\cdot)$ is the indicator function. When $\phi^{\mathcal{R}}>\phi^{\mathcal{IR}}$, image $\mathbf{x}$ is predicted to be "relevant." The class probability that image $\mathbf{x}$ belongs to $\mathcal{R}$ can be given by

$$p(h^t(\mathbf{x})=1|\mathbf{x})=\phi^{\mathcal{R}}/(\phi^{\mathcal{R}}+\phi^{\mathcal{IR}}). \tag{8}$$

$h^t(\mathbf{x}) = 1$ if $p(h^t(\mathbf{x}) = 1|\mathbf{x}) > 0.5$; otherwise $h^t(\mathbf{x}) = -1$. Similar to Real AdaBoost, we set $q^t(\mathbf{x}) = (1/2)\log(p(h^t(\mathbf{x})=1|\mathbf{x})/1-p(h^t(\mathbf{x}))=1|\mathbf{x})$. After selecting $t$ feature axes, to rank the images, an ensemble value can be obtained for $\mathbf{x}$ as

$$V^t(\mathbf{x})=\sum_{i=1}^{t}q^i(\mathbf{x}). \tag{9}$$

The larger $V^t(\mathbf{x})$ is, the better image $\mathbf{x}$ accords with the query concept determined by $\mathcal{R}$ and $\mathcal{IR}$. Assume that there are totally $T$ features selected, the images can be ranked according to $V^T(\mathbf{x})$ in descending order. The top images are returned to the user as the retrieval result.

The optimal dimensionality $T^*$ of the selected feature space is determined by minimizing the training error $\epsilon^T$. That is

$$T^*=\arg\min_T \epsilon^T. \tag{10}$$

$\epsilon^T$ is given based on the ensemble label estimations of the training samples calculated by the *leave one out* method [11] as $H(\mathbf{x}) = \text{sign}[V^T(\mathbf{x})]$. There are $d$ possible choices for $T^*$ in total, and in practice we quantize number $d$ into $\gamma$ levels to find $T^*$ (that is, the possible values for $T^*$ are $(d/\gamma),(2d/\gamma),\ldots,d$). Considering both the computational cost and the retrieval accuracy, we set $\gamma=3$ in our system empirically.

### B. Computational Complexity Analysis

The time complexity of our feature selection algorithm to select one optimal feature axis is

$$O(m\times n\times d+M\times(m+n)) \tag{11}$$

where $M$ is the total number of images in the database; $m$ and $n$ are the sizes of $\mathcal{R}$ and $\mathcal{IR}$, respectively; $d$ is the dimensionality of the original feature space $\mathcal{F}$. This time cost is comprised by mainly two parts: $O(m\times n\times d)$ to calculate the similarity between $\mathcal{R}$ and $\mathcal{IR}$ along $d$ feature axes; and $O(M\times(m+n))$ to construct the FKNN classifier and get class probability estimations for all images.

As for MDA, BiasMap, and SMMS, they all contain matrix computation [at least $O((m+n)\times d^2)$ for calculating $S_w$ and $S_b$, at least $O(d^3)$ for SVD [2], etc.]. Usually, $m$ and $n$ are much smaller than $d$, and our feature selection method costs much less time than the DA approaches in most cases. In the conventional AdaBoost feature selection approach, to select one feature axis, $d$ classifiers are constructed to calculate and compare the training errors, whose computational complexity increases dramatically when the complexity of weak classifiers increases. For example, when FKNN classifier is used in conventional AdaBoost, the time complexity is $O(d\times M\times(m+n))$. As for our feature selection algorithm, since only one classifier is constructed to select one feature axis, the influence of the complexity of weak classifiers is much smaller. Thus, our method is usually faster, especially when we use complex weak classifiers. The experimental results in Figs. 11(b) and 12(b) verify our analysis.

## IV. Online Feature Selection in RBIR Systems

As mentioned in Section I, not much work has been done on feature selection in the RBIR context. This is partly due to the difficulty to represent images in a uniform feature space in RBIR systems. Recently, a binary region-based feature representation has been proposed [15], [36], which extracts a codebook from the original region-based features with grouping method, and uses the indexes of the codebook to represent images. The codebook are more concise than the original features, and the curse of dimensionality is alleviated. However, much valuable low-level information may be lost in the binary codebook representation, and the effectiveness of this method is limited.

Here, we propose a novel region-based image representation. Based on a generative model, a *fuzzy codebook* is learned to represent images in a uniform feature space with real-valued fuzzy features. With the new feature representation, the boosting feature selection algorithm described in Sections II and III is applied in the RBIR system. In the following sections, we introduce the method to get the original region-based features, followed by the details of extracting the fuzzy codebook and representing images by the fuzzy codebook.

### A. Feature Extraction and Image Segmentation

An image is partitioned into small blocks without overlapping, and feature vectors (color features and texture features) are extracted based on each block. In our system each block has $16\times 16$ pixels, which is a tradeoff between the computational complexity and the effectiveness of the extracted features. The JSEG algorithm [8] is adopted to segment the image, which can adaptively determine the number of regions, and the segmentation performance is fairly good in most situations. After segmentation, an image $\mathbf{x}$ is divided into a set of regions $\nabla(\mathbf{x})=\{r_1(\mathbf{x}),\ldots,r_m(\mathbf{x})\}$ without overlapping, with each region represented by the mean feature vector of its member blocks. Each region $r_i(\mathbf{x})$ corresponds to a saliency membership $v(r_i(\mathbf{x}))$, which describes how well this region represents image $\mathbf{x}$. Fig. 5 gives the segmentation result of an example image. The region saliency $v(r_i(\mathbf{x}))$ is given by

$$v(r_i(\mathbf{x}))=\frac{1}{Z}\cdot\frac{\mathrm{Area}(r_i(\mathbf{x}))}{\mathrm{Avg}_{b_j\in r_i(\mathbf{x})}\|b_j-c(\mathbf{x})\|}$$

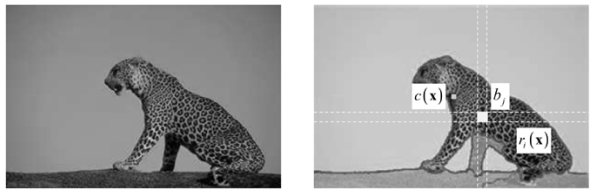


Fig. 5. Example for image segmentation. (a) Original image and (b) segmentation result.

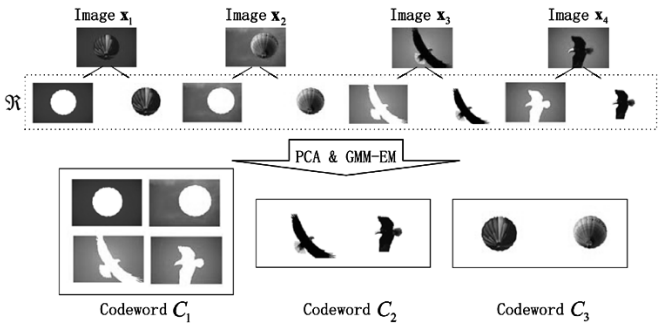


Fig. 6. Example for generating the fuzzy codebook from the regions of four images $\mathbf{x}_1,\ldots,\mathbf{x}_4$, where three codewords $C_1, C_2, C_3$ are learned.

where $b_j\in r_i(\mathbf{x})$ is a member block in $r_i(\mathbf{x})$; $c(\mathbf{x})$ is the center of image $\mathbf{x}$; $Z$ is the normalization factor to make $\sum_i^m v(r_i(\mathbf{x}))=1$.

### B. Learning the Fuzzy Codebook

Let $\mathcal{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_M\}$ denote a database with $M$ images, and $\Re=\bigcup_{i=1}^{M}\nabla(\mathbf{x}_i)$ are all the segmented regions from all the images in $\mathcal{X}$. The *principle component analysis* (PCA) is exploited to simplify the low-level feature representation based on $\Re$, which can reduce the redundancy in low-level features, and can alleviate the singularity problem in matrix computation when generating the fuzzy codebook. After PCA, the original $\Re$ is projected into a new feature space $\tilde{\mathcal{F}}$ with a much lower dimensionality $\tilde{d}$, and we denote it by $\tilde{\Re}$. Rename the elements in $\tilde{\Re}$ as $\tilde{\Re}=\{r_1,\ldots,r_n\}$. We assume that the region $r_j$ is generatively formed by $N$ codewords $C_1,\ldots,C_N$ in the form of a *Gaussian mixture model* (GMM) as

$$p(r_j)=\sum_{k=1}^{N}p(C_k)p(r_j|C_k) \tag{12}$$

where $p(C_k)$ is the prior probability of codeword $C_k$. $p(r_j|C_k)$ follows Gaussian distribution

$$p(r_j|C_k)=(2\pi)^{-\frac{\tilde{d}}{2}}|\Sigma_k|^{-\frac{1}{2}}\cdot\exp\left\{-\frac{1}{2}(r_j-\vec{\boldsymbol{\mu}}_k)^T(\Sigma_k)^{-1}(r_j-\vec{\boldsymbol{\mu}}_k)\right\}$$

where $\vec{\boldsymbol{\mu}}_k$ and $\Sigma_k$ are the mean vector and covariance matrix of $C_k$ respectively. Then the GMM-EM algorithm proposed in [1] is adopted to group the regions in $\tilde{\Re}$ into $N$ clusters and calculate the following parameters iteratively: the codeword priors $p(C_k)$, the mean vectors $\vec{\boldsymbol{\mu}}_k$ and the co-variance matrixes $\Sigma_k, k=1,\ldots,N$. Fig. 6 illustrates how to generate the fuzzy codebook

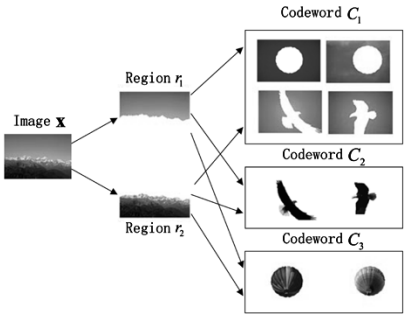


Fig. 7. Relationship among codewords, regions, and images, represented by a directed graph model.

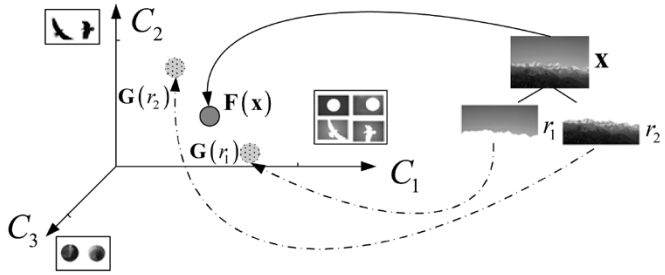


Fig. 8. Example for representing images by the fuzzy codebook. $C_1, C_2, C_3$ span a 3-D feature space, in which image $\mathbf{x}$ is represented as a fuzzy feature vector $\mathbf{F}(\mathbf{x}) = \sum_{i=1,2} p(r_i|\mathbf{x})\mathbf{G}(r_i)$.

from the images. In our experiment, we set $N = 600$ empirically (see Section V-A for details).

Note that although there are model selection methods which can adaptively determine $N$, such as the mixture-based clustering method that selects $N$ by Minimum Message Length criterion [19], from experiments we find that these methods tend to select a small $N$ (no more than 200), and it is difficult to choose appropriate parameters in practice. By experiments, their retrieval results are usually worse than that of fixing $N$.

### C. Representing Images by the Fuzzy Codebook

We assume that the probability of an image belonging to a codeword is determined by its region partitions. The process can be illustrated by a directed graph model [17] shown in Fig. 7. Given an image $\mathbf{x} \in \mathcal{X}$ and its region partitions $\{r_1, \ldots, r_m\}$, we have the conditional independence relationship $p(C_k|\mathbf{x}, r_i) = p(C_k|r_i), i = 1, \ldots, m$. Thus

$$p(C_k|\mathbf{x}) = \frac{1}{p(\mathbf{x})} \sum_{i=1}^{m} p(\mathbf{x}|r_i) p(C_k, r_i) = \sum_{i=1}^{m} p(r_i|\mathbf{x}) p(C_k|r_i) \tag{13}$$

where

$$p(C_k|r_i) = \frac{p(C_k)p(r_i|C_k)}{\sum_{j=1}^{N} p(C_j)p(r_i|C_j)} \tag{14}$$

and $p(r_i|\mathbf{x})$ can be understood to be equal to the region saliency $v(r_i)$. $p(C_j)$ and $p(r_i|C_j)$ are obtained in the previous clustering process in Section IV-B. $p(C_k|\mathbf{x})$ can be interpreted as the appropriate degree of using the codeword $C_k$ to represent the image $\mathbf{x}$. Thus, vector $\mathbf{F}(\mathbf{x}) = [p(C_1|\mathbf{x}), \ldots, p(C_N|\mathbf{x})]^T$ can be treated as a set of real-valued features for $\mathbf{x}$ represented in the feature space spanned by codewords $C_1, \ldots, C_N$. $\mathbf{F}(\mathbf{x})$ is a fuzzy feature vector, with each entry denoting how well each codeword interprets the image. The feature representation process can be intuitively explained as follows: the region of image $\mathbf{x}, r_i$, is represented in the feature space spanned by $C_1, \ldots, C_N$ as $\mathbf{G}(r_i)$, where $\mathbf{G}(r_i) = [p(C_1|r_i), \ldots, p(C_N|r_i)]^T$. Then $\mathbf{x}$ is represented as $\mathbf{F}(\mathbf{x}) = \sum_{i=1}^{m} p(r_i|\mathbf{x})\mathbf{G}(r_i)$. Fig. 8 illustrates the process to represent images by the fuzzy codebook.

With the new region-based representation, images are put into one uniform real-valued feature space. Since the new features are fuzzy features, the boosting online feature selection algorithm proposed in Sections II and III can be applied in the RBIR system to select codewords one by one from the fuzzy codebook, and to construct the ensemble classifier for retrieval during the relevance feedback rounds.

## V. Experimental Results

The experiments are carried out on a database with 10 000 real-world images from the Corel gallery [7]. These images are pre-assigned to 100 categories by high-level semantics (defined by a large group of human observers as standard ground truth), such as autumn, balloon, bird, dog, eagle, sunset, tiger, etc., each containing 100 images. Two types of color features and two types of texture features are used in the experiments, which are: the nine-dimensional color moments in LUV color space (the first three-order moments) and the 64-dimensional color histogram in HSV color space; the ten-dimensional coarseness vector and the eight-dimensional directionality (the Tamura's texture features [26]). Assume that during each query session the user looks for images in one semantic category, and conducts five feedback rounds. At each feedback round, the top ten images among the returned images, which have not been labeled in previous feedback rounds, are labeled as feedback information. The statistical average top-$k$ precision is used as the performance measurement

$$P_k = \frac{\text{the "relevant" image number in } k \text{ returned images}}{k}. \tag{15}$$

Since each semantic category has 100 images, the top-$k$ recall ($\mathrm{Re}_k$) has the following relationship with $P_k$

$$\mathrm{Re}_k = \frac{\text{the "relevant" image number in } k \text{ returned images}}{100} = \frac{k \times P_k}{100}.$$

When $k$ is fixed, $\mathrm{Re}_k \propto P_k$, and we do not show $\mathrm{Re}_k$ in the experiments for the sake of simplicity. We randomly select ten images from each semantic category as query images, and run totally 1000 query sessions to calculate the average precision.

### A. Determing the Codebook Size

Considering both the computational complexity and the effectiveness to represent images, the size of the fuzzy codebook is selected through the following experiment. The average $P_{20}, P_{50}$,

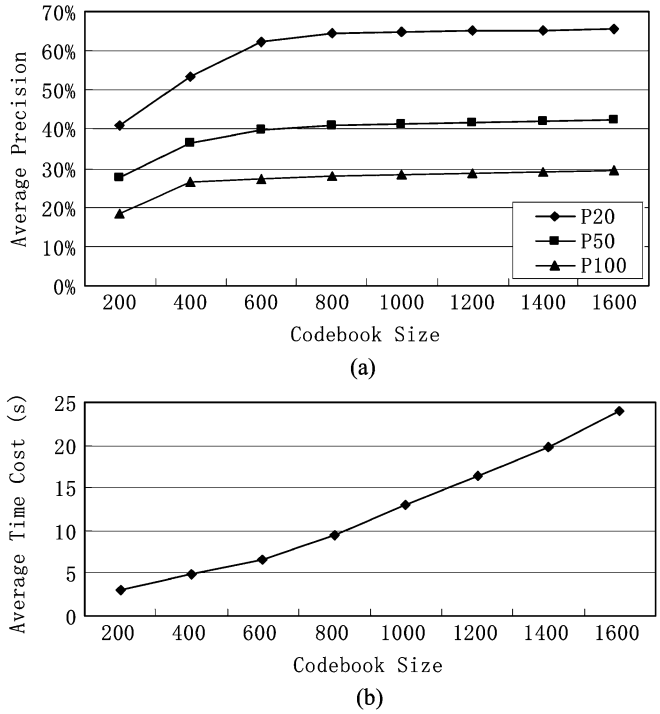


Fig. 9. Precision and time cost for different codebook sizes.

and $P_{100}$ for the fifth feedback round are calculated with different codebook sizes. The results are shown in Fig. 9(a). At the same time, the corresponding average retrieval time costs for different codebook sizes are given in Fig. 9(b). In this experiment, we set $\alpha = 0.7, \beta = 0.4$. The figures show that for all $P_{20}, P_{50}$, and $P_{100}$, the larger the codebook size, the better the retrieval performance, and the longer the retrieval time. When the number of codewords exceeds 600, the improvement of the precision becomes not very significant, but the time cost increases greatly. Thus, in practical, the system uses 600 codewords as a tradeoff.

We have also evaluated the mixture-based clustering method with model selection [19], which can adaptively determine $N$ by minimum message length criterion. Through experiments, we find that this method tends to have a small $N$ (no more than 200), and it is difficult to choose appropriate parameters. Thus, practically we prefer to fix the codebook size.

### B. Comparison With the State-of-the-Arts

We evaluate the performance of our algorithm by making three sets of comparisons with the state-of-the-art algorithms. 1) We compare our method with the classical RBIR methods in [6] (which straightforwardly calculates the image-to-image similarity for retrieval) and [36] (which is a typical binary feature representation method with online relevance feedback learning without feature selection). 2) We compare our algorithm with other feature selection algorithms, including the DA approaches (MDA [33], BDA [37] and SMMS [35]) and the conventional training-error-based AdaBoost approach [27]. 3) We compare our method with the SVM-classifier-based method, which yields almost the most promising retrieval performance in current CBIR context.

In the following experiments of this section, we set $\alpha = 0.7, \beta = 0.4$ for our method. For a fair comparison, all the methods use the same query images. Also, all the algorithms are initiated in the same way. That is, the 1st round of retrieval is carried out with the same method by calculating the image-to-image similarity between each image and the query image according to the straightforward algorithm in [6], and then ranking the images by the similarity.

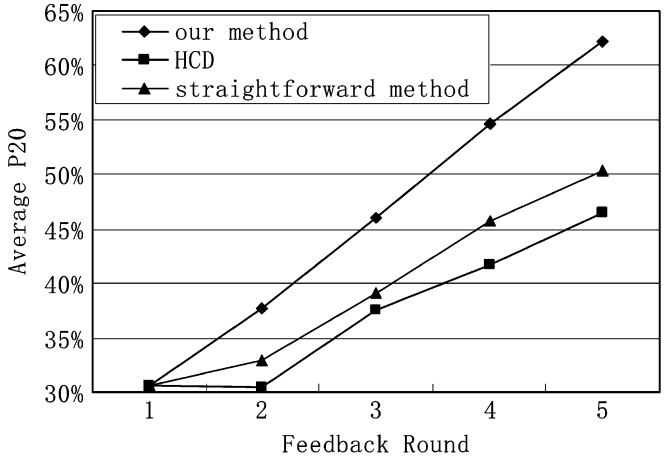


Fig. 10. Average precision of our method, HCD, and the straightforward similarity-based method.

*1) Comparison With Classical RBIR Approach:* In this experiment, we compare our algorithm with two kinds of classical RBIR approaches: the straightforward similarity-based approach [6], and the unsupervised hidden concept discovery (HCD) method [36]. Note that the method proposed in [6] does not use relevance feedback. To make a full and fair comparison, we introduce the relevance feedback mechanism to the straightforward method as follows: for each image in the database, assume that its average similarity to the "relevant" and "irrelevant" images are $\psi^+$ and $\psi^-$, respectively. $\psi^+ - \psi^-$ can be used to denote how this image accords with the current query concept. Images are ranked by this value and the top ones are returned. As for HCD, the codebook size is 600.

Fig. 10 gives the average $P_{20}$ of our method, HCD, and the straightforward similarity-based method. From the figure, we can see that both HCD and our method are better than the straightforward approach. This indicates that online learning mechanism can improve the retrieval performance of the RBIR system. Further more, our method outperforms HCD significantly and consistently from the second feedback round. For example, the precision improvement in round 5 attains 24%. This shows that the proposed image representation by fuzzy codebook and the online feature selection mechanism can improve the retrieval accuracy.

*2) Comparison With Other Feature Selection Algorithms:* In this experiment, we compare our boosting online feature selection algorithm with MDA, BDA, SMMS and the conventional AdaBoost approach.

First, the average $P_{20}$ of our algorithm and the DA approaches (MDA, BDA, SMMS) after each feedback round are given in Fig. 11(a). The result clearly shows that our method consistently outperforms the others from the 2nd feedback round. For example, the average precision improvements of our method after the fifth feedback round are 22%, 20%, 32%, respectively, compared with MDA, BDA, and SMMS. Fig. 11(b) gives the cor-

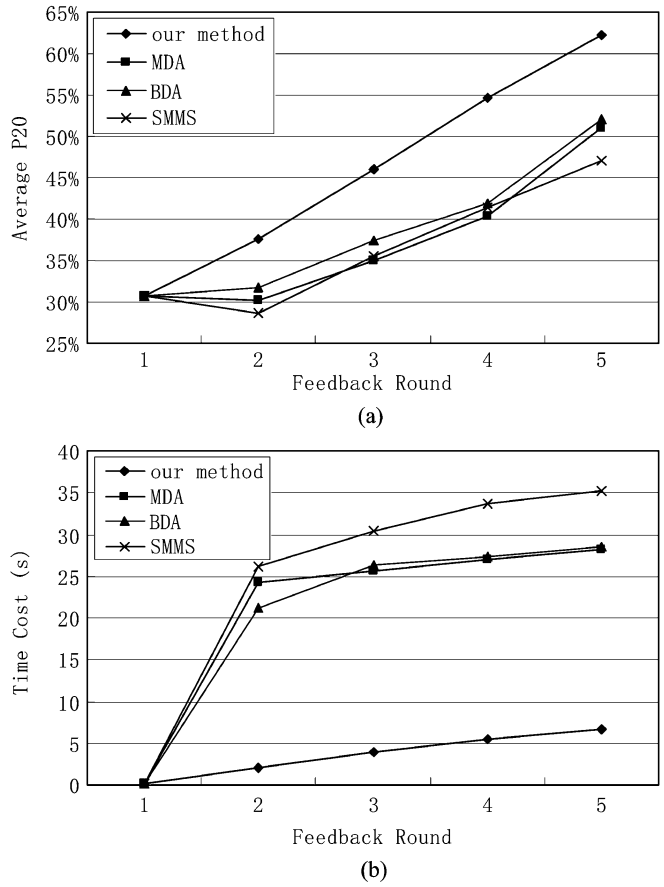


Fig. 11. Comparison of our algorithm, MDA, BDA, and SMMS.

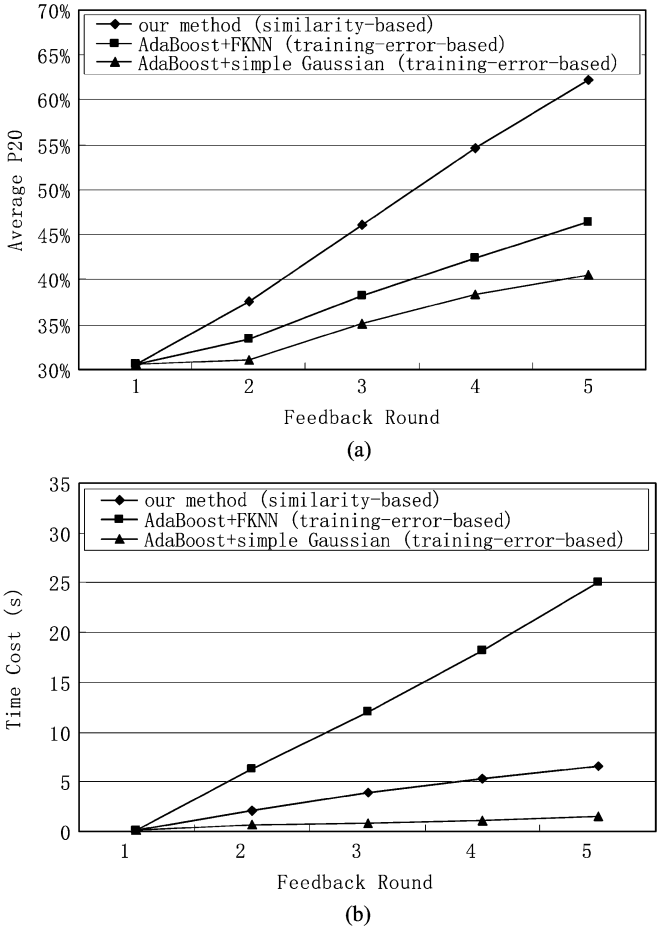


Fig. 12. Comparison of our algorithm (which uses similarity-based feature selection criterion) and the AdaBoost approaches (which use training-error-based feature selection criterion).

responding time costs for our algorithm and the other three DA methods during each feedback round, which shows that our algorithm is much faster than the compared DA approaches.

Second, in order to verify the effectiveness of our proposed similarity-based feature selection criterion, we compare our algorithm with two kinds of methods using conventional AdaBoost mechanism with training-error-based feature selection criterion: 1) The AdaBoost image retrieval approach described in [27], which uses the simple Gaussian classifier as the weak learner; and 2) the AdaBoost algorithm using FKNN classifier as the weak learner. The average $P_{20}$ of these algorithms are given in Fig. 12(a), and Fig. 12(b) gives their corresponding time costs. The figures show that our online feature selection algorithm has much better retrieval precision than both two kinds of AdaBoost approaches. The AdaBoost algorithm using simple Gaussian classifier has the smallest time cost, but the retrieval performance is far from satisfactory. When more complex classifiers are used, such as the FKNN classifier, the time cost of AdaBoost grows dramatically. Since the complexity of classifiers has small effect on the computational complexity of our algorithm, our method is much faster than the AdaBoost algorithm using FKNN classifier. This phenomenon is consistent with the analysis in Section III-B.

*3) Comparison With SVM-Based Algorithm:* In this experiment, we compare our method with the SVM-classifier-based method. As discussed before, SVM method yields nearly the most promising performance in the CBIR context. In this experiment, the SVM classifier uses the RBF kernel $K(\vec{\mathbf{x}}, \vec{\mathbf{x}}_i) = \exp\{-\|\vec{\mathbf{x}} - \vec{\mathbf{x}}_i\|^2/\gamma T\}$, where $T$ is the dimensionality of $\vec{\mathbf{x}}$, and the polynomial kernel $K(\vec{\mathbf{x}}, \vec{\mathbf{x}}_i) = (1+\vec{\mathbf{x}}\cdot\vec{\mathbf{x}}_i)^{\lambda}$. We use the software $\mathrm{SVM}^{\mathrm{light}}$ [16] to construct the SVM classifier, and tune $\gamma$ and $\lambda$ within a large range to find good performances for the SVM classifier.

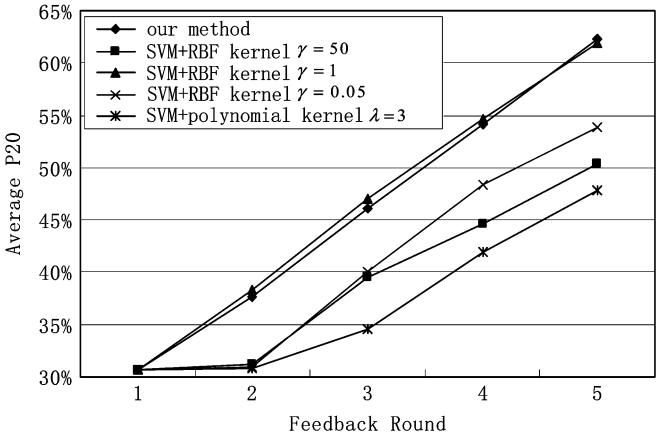


Fig. 13. Comparison of our method and the SVM-classifier-based algorithm.

The average $P_{20}$ of our algorithm and the SVM method with different parameter settings are given in Fig. 13. From the figure, we can see that when using RBF kernel with $\gamma = 1$, the SVM

TABLE I
INFLUENCE OF $\alpha$ WITH $\beta = 0.4$

| $\alpha$ | $t=1$ | $t=2$ | $t=3$ | $t=4$ | $t=5$ |
|---|---|---|---|---|---|
| 0.1 | 30.69% | 34.4% | 42.16% | 48.46% | 54.52% |
| 0.4 | 30.69% | 35.6% | 43.4% | 49.64% | 55.66% |
| 0.7 | 30.69% | 37.64% | 46.06% | 54.64% | 62.27% |
| 0.9 | 30.69% | 36.58% | 45.36% | 51.53% | 58.7% |

TABLE II
INFLUENCE OF $\beta$ WITH $\alpha = 0.7$

| $\beta$ | $t=1$ | $t=2$ | $t=3$ | $t=4$ | $t=5$ |
|---|---|---|---|---|---|
| 0.1 | 30.69% | 35.22% | 43.14% | 50.46% | 57.32% |
| 0.4 | 30.69% | 37.64% | 46.06% | 54.64% | 62.27% |
| 0.7 | 30.69% | 34.52% | 42.22% | 49.54% | 55.65% |
| 0.9 | 30.69% | 33.47% | 41.36% | 48.58% | 54.58% |

classifier has the best performance. This is almost the best performance we can get in practice after carefully tuning the parameters. Through experiments, we find that it is difficult to get an appropriate parameter setting for polynomial kernel to make the retrieval performance comparable with that of using RBF kernel. Thus, we only show the result of using polynomial kernel with $\lambda = 3$. The result indicates that our algorithm yields comparable retrieval performance with SVM-based approach. Unlike the SVM classifier which depends on choosing a proper kernel function and appropriate parameters, our method is fairly robust to the parameter setting, which will be shown in the next section.

From all the comparisons in this section, we can see that the algorithm we proposed in this paper is suitable for CBIR online relevance feedback learning, which has fairly good retrieval accuracy and tolerable processing time.

### C. Influence of Parameters

There are two parameters in our feature selection algorithm: $\alpha$ and $\beta$ in (6). First, we fix $\beta = 0.4$ and let $\alpha$ change from 0 to 1 to test the influence of $\alpha$ on the retrieval performance. Table I shows the average $P_{20}$ of our method in this parameter setting. Then we fix $\alpha = 0.7$, and let $\beta$ change from 0 to 1 to test the influence of $\beta$. Table II shows the average $P_{20}$ in such cases. From the tables, we can see that the performance attains a local optimum when $\alpha = 0.7, \beta = 0.4$. And the precision for $\alpha > \beta$ is obviously better than that for $\alpha \leq \beta$ in average. This phenomenon is consistent with the asymmetry requirement of CBIR problem. Generally speaking, when $\alpha > \beta$, the influences of these parameters are not very significant, and our algorithm is fairly robust.

## VI. CONCLUSIONS AND DISCUSSIONS

In this paper, we explore the issue of online feature selection in the CBIR online learning context. A novel feature selection criterion based on calculating the similarity between the "relevant" and "irrelevant" image sets is proposed, and an effective online feature selection algorithm is implemented in a boosting manner to select the most representative feature axes for the current query concept, and to combine the incrementally constructed classifiers into an ensemble classifier to retrieve images. This method accounts for the asymmetry requirement to treat the "relevant" and "irrelevant" images differently, and makes no assumption for image distribution. To apply the proposed online feature selection algorithm to RBIR systems, a novel region-based image representation is proposed, which represents images in a uniform real-valued feature space. Experimental results, including comparisons with many state-of-the-arts, show that the proposed method can improve the retrieval performance and save the processing time.

As mentioned before, the definitions of operators $\cap, \cup$ and $-$ in FFCM and UFM measurements may take many forms in the fuzzy set theory. Different choices result in different feature selection criterions, and yield different retrieval performances. More evaluations will be done in the future. In addition, Tavesky also suggested an alternative form for the feature matching function, the ratio model [30], $\check{S}(a,b) = (\mathbf{A} \cap \mathbf{B}/|\mathbf{A} \cap \mathbf{B}| + \alpha|\mathbf{A} - \mathbf{B}| + \beta|\mathbf{B} - \mathbf{A}|)$, to measure the similarity between two stimuli $a$ and $b$. More work can be done to generalize the ratio model to fuzzy features and use it in our online learning framework. Moreover, similar to the AdaBoost approach [27], our algorithm in this paper can only select feature axes parallel to the original ones, another aspect of our future work is to adopt some optimization method (such as the 1-D sequential method proposed in [20]) to alleviate this problem. Furthermore, the training-error-based dimensionality decision method used in this paper is not optimal, because the small sample learning difficulty makes the training error unrepresentative for the generalization error. How to determine the appropriate dimensionality is still an open problem. We will also investigate this problem in our future work.

## VII. ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

## REFERENCES

[1] J. Bilmes, "A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," Int. Comput. Sci. Inst. (ICSI), Berkeley, CA, Tech. Rep. ICSI-TR-97-021, 1997.
[2] M. Brand, "Incremental singular value decomposition of uncertain data with missing values," in *Proc. Eur. Conf. Computer Vision*, vol. 2350, Copenhagen, Denmark, 2002, pp. 707–720.
[3] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik, "Blobworld: A system for region-based image indexing and retrieval," in *Proc. Int. Conf. Visual Information Systems*, Amsterdam, The Netherlands, 1999, pp. 509–516.
[4] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 8, pp. 1026–1038, Aug. 2002.
[5] Y. Q. Chen, X. Z. Zhou, and T. S. Huang, "One-class SVM for learning in image retrieval," in *Proc. IEEE Int. Conf. Image Processing*, vol. 1, Thessaloniki, Greece, 2001, pp. 34–37.
[6] Y. X. Chen and J. Z. Wang, "A region-based fuzzy feature matching approach for content-based image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1252–1267, Sep. 2002.
[7] *Corel stock photo library*, Ontario, Canada: Corel.
[8] Y. Deng, B. S. Manjunath, and H. Shin, "Color image segmentation," in *Proc. Int. Conf. Computer Vision Pattern Recognition*, vol. 2, Ft. Collins, CO, 1999, pp. 446–451.

[9] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in *Proc. Int. Conf. Machine Learning*, Bari, Italy, 1996, pp. 148–156.
[10] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *Ann. Stat.*, vol. 38, no. 2, pp. 337–374, 2000.
[11] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.
[12] G. D. Guo, A. K. Jain, W. Y. Ma, and H. J. Zhang, "Learning similarity measures for natural image retrieval with relevance feedback," *IEEE Trans. Neural Networks*, vol. 13, no. 4, pp. 811–820, Apr. 2002.
[13] P. Hong, Q. Tian, and T. S. Huang, "Incorporate support vector machines to content-based image retrieval with relevant feedback," in *Proc. Int. Conf. Image Processing*, vol. 3, Vancouver, BC, Canada, 2000, pp. 750–753.
[14] T. S. Huang, X. S. Zhou, M. Nakazato, Y. Wu, and I. Cohen, "Learning in content-based image retrieval," in *Proc. IEEE Int. Conf. Development and Learning*, Cambridge, MA, 2002, pp. 155–162.
[15] F. Jing, M. J. Li, H. J. Zhang, and B. Zhang, "An effective and efficient region-based image retrieval framework," *IEEE Trans. Image Process.*, vol. 13, no. 5, pp. 699–709, May 2004.
[16] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods—Support Vector Learning*, B. Scholkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1999.
[17] M. Jordan, Ed., *Learning in Graphical Models*. Cambridge, MA: MIT-Press, 1999.
[18] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE Trans. Syst., Man Cybern.*, vol. SMC-15, no. 4, pp. 580–585, Apr. 1985.
[19] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1154–1166, Sep. 2004.
[20] C. Liu and H. Y. Shum, "Kullback–Leibler boosting," in *Proc. Int. Conf. Computer Vision Pattern Recognition*, vol. 1, 2003, pp. 587–594.
[21] T. P. Minka and R. W. Picard, "Interactive learning using a society of models," *Pattern Recognit.*, vol. 30, no. 4, pp. 565–581, 1997.
[22] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A powerful tool in interactive content-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 644–655, May 1998.
[23] S. Santini and R. Jain, "Similarity measures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 9, pp. 871–883, Sep. 1999.
[24] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the ealy years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2002.
[25] J. B. Tenenbaum and T. L. Griffiths, "Generalization, similarity, and Bayesian inference," *Behav. Brain Sci.*, vol. 24, pp. 629–640, 2001.
[26] H. Tamura, S. Mori, and T. Yamawaki, "Texture features corresponding to visual perception," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-8, no. 6, pp. 460–473, 1978.
[27] K. Tieu and P. Viola, "Boost image retrieval," in *Proc. Int. Conf. Computer Vision Pattern Recognition*, vol. 1, Hilton Head Island, SC, 2000, pp. 228–235.
[28] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proc. ACM Multimedia*, Ottawa, ON, Canada, 2001.
[29] K. Torkkola, "Feature extraction by nonparametric mutual information maximization," *J. Mach. Learn. Res.*, vol. 3, pp. 1415–1438, 2003.
[30] A. Tvesky, "Feature of similarity," *Psychol. Rev.*, vol. 84, no. 4, pp. 327–352, 1977.
[31] N. Vasconcelos and A. Lippman, "Learning from user feedback in image retrieval system," in *Proc. NIPS*, Denver, CO, 1999, pp. 977–983.
[32] L. Wang, K. L. Chan, and Z. H. Zhang, "Bootstrapping SVM active learning by incorporating unlabeled images for image retrieval," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, vol. 1, 2003, pp. 629–634.
[33] Y. Wu, Q. Tian, and T. S. Huang, "Discriminant EM algorithm with application to image retrieval," in *Proc. Int. Conf. Computer Vision Pattern Recognition*, vol. 1, 2000, pp. 222–227.
[34] L. Zhang, F. Z. Lin, and B. Zhang, "Support vector machine learning for image retrieval," in *Proc. Int. Conf. Image Processing*, vol. 2, Thessaloniki, Greece, 2001, pp. 721–724.
[35] W. D. Zhang and T. H. Chen, "Classification based on symmetric maximized minimal distance in subspace (SMMS)," in *Proc. Int. Conf. Computer Vision Pattern Recognition*, vol. 2, 2003, pp. 100–105.
[36] R. F. Zhang and Z. F. Zhang, "Hidden semantic concept discovery in region based image retrieval," in *Proc. Int. Conf. Computer Vision Pattern Recognition*, vol. 2, Washington, DC, 2004, pp. 996–1001.
[37] X. S. Zhou and T. S. Huang, "Small sample learning during multimedia retrieval using BiasMap," in *Proc. Int. Conf. Computer Vision Pattern Recognition*, vol. 1, 2001, pp. 11–17.
[38] H. J. Zimmermann, *Fuzzy Set Theory and Its Applications*, 2nd ed. Boston, MA: Kluwer, 1991.

**Wei Jiang** (S'04) received the B.S. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2002, where she is currently pursuing the M.S. degree.

Her research interests include content-based image retrieval/management, pattern recognition, and image processing.

**Guihua Er** (M'00) received the B.S. degree from the Automation Department, Tianjin University, China, in 1984, and the M.S. degree from the Automation Department, Beijing Institute of Technology, Beijing, China, in 1989.

She is now an Associate Professor and the Vice Director of Broad-band Networks and Digital Media Lab in the Automation Department, Tsinghua University, Beijing. Her research interests include multimedia database and multimedia information coding.

**Qionghai Dai** (SM'00) received the B.S. degree in mathematics from Shanxi Normal University, China, in 1987, and the M.E. and Ph.D. degrees in computer science and automation from Northeastern University, China, in 1994 and 1996, respectively.

Since being a Postdoctoral Researcher in the Automation Department, he has been with the Media Lab, Tsinghua University, China, where he is currently an Associate Professor and Head of the Lab. His research interests are in signal processing, broad-band networks, video processing, and communication.

**Jinwei Gu** (S'04) received the B.S. degree from the Automation Department, Tsinghua University, China, in 2002, where he is currently puruing the M.S. degree.

His research interests are in pattern recognition, computer vision, and intelligent information processing.