In Proc. of IEEE Conf. on CVPR'2000, Vol.I, pp.222-227, Hilton Head Island, SC, 2000

Discriminant-EM Algorithm with Application to Image Retrieval

Ying Wu, Qi Tian, Thomas S. Huang
Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL 61801
{yingwu,qitian,huang}@ifp.uiuc.edu

Abstract

In many vision applications, the practice of supervised learning faces several difficulties, one of which is that insufficient labeled training data result in poor generalization. In image retrieval, we have very few labeled images from query and relevance feedback so that it is hard to automatically weight image features and select similarity metrics for image classification. This paper investigates the possibility of including an unlabeled data set to make up the insufficiency of labeled data. Different from most current research in image retrieval, the proposed approach tries to cast image retrieval as a transductive learning *problem, in which the generalization of an image classifier is only defined on a set of images such as the given image database. Formulating this transductive problem in a probabilistic framework, the proposed algorithm, Discriminant-EM (D-EM), not only estimates the parameters of a generative model, but also finds a linear transformation to relax the assumption of probabilistic structure of data distributions as well as select good features automatically. Our experiments show that D-EM has a satisfactory performance in image retrieval applications. D-EM algorithm has the potential to many other applications.*

1 Introduction

Recent years have witnessed a rapid increase of the volume of digital image collections, which motivates the research of image retrieval [2, 9, 11]. Early research of image retrieval is searching by manually annotating every image in a database. However, these text-based techniques are impractical for two reasons: large size of image databases and subjective meanings of images. To avoid manual annotating, an alternative approach is content-based image retrieval (CBIR), by which images would be indexed by their visual contents such as color, texture, shape, etc. Many research efforts have been made to extract these low-level image features [7, 12], evaluate distance metrics [10, 13], and look for efficient searching schemes [14, 16].

However, one of the difficulties of CBIR is the gap between high-level concepts and low-level image features, due to the rich content but subjective concepts of an image. The mapping between them would be highly nonlinear such that it is impractical to represent it explicitly. A promising approach to this problem is machine learning, by which the mapping could be learned through a set of examples. In our proposed approach, image retrieval is cast as a statistical learning problem.

Although learning techniques offer a flexible and tractable means to many vision applications, the generalization of learning results has to depend not only on training algorithms but also on training data sets. If the training data set is large and representative enough, good generalization may be obtained. On the other hand, if the training data set is small or not informative, it is hard to guarantee a good generalization. If a classifier is over-trained on the training data set, *overfitting* will probably occur.

In some cases, good and large training data sets can be easily obtained. However, in many situations in vision or image understanding, collecting a large and informative training data set is not a trivial task. First, data collecting may not be straightforward. Second, collecting representative data may be difficult due to the large variation in visual inputs. Third, manually labeling a large date set is always time-consuming.

In fact, it seems that it might not be necessary to have every sample labeled in supervised learning. A very interesting result given by the theory of the support vector machine (SVM) [15] is that the classification boundary is related only to some support vectors, rather than the whole data set. Although the identification of these support vectors is not trivial, it motivates us to think about the roles of non-support vectors. Fortunately, it is easier to collect unlabeled data. The issue of combining unlabeled data in su-

pervised learning begins to receive more and more research efforts recently and the research of this problem is still in its infancy. Without assuming parametric probabilistic models, several methods are based on the SVM [4, 1, 5]. However, when the size of unlabeled data becomes very large, these methods need formidable computational resources for mathematical programming. Another difficulty of these SVM-based methods is that the way of selecting the kernel function is heuristic. Some other alternative methods try to fit this problem into the EM framework and employ parametric models [8, 6], and have some applications in text classification. Although EM offers a systematic approach to this problem, these methods largely depend on the *a priori* knowledge about the probabilistic structure of data distribution.

This paper looks into the image retrieval problem in the view of transductive learning, and presents a probabilistic approach to employ unlabeled data.

2 Problem Formulation

The task of image retrieval is to find as many as possible "similar" images to the query images in a given database. The retrieval system acts as a classifier to divide the images in the database into two classes, either relevant or irrelevant.

In image retrieval, an image can be represented by a feature vector $\mathbf{x}$ and its label y. Since the image space is huge, it is practical to represent images in lower dimensional feature space instead of raw image space. Physical features and mathematical features are two typical representations. Many research efforts have been made to extract physical features such as color features, texture features, edge features, structure features, or combination of these features [2, 7, 12]. However, images are too rich to represent by these physical features. An alternative representation is mathematical features, which only performs dimension reduction in mathematical senses. Principal component analysis (PCA) is a typical technique to obtain such mathematical features [14]. Generally, it is up to us to determine how many principal components we would use.

Both representations are facing the same problem: automatic feature weighting, which is partly the reason of the gap between high-level concepts and low-level image features. For example, if images are represented as a set of physical features, sometimes color features such as color histogram or color moments are more suitable for retrieval, but sometimes a combination of color and texture features will have better performance. A possible approach is to specify a set of rules to select better features. However, it is impractical to construct such rules for every possible image class. If mathematical features are used, it would be difficult to understand the meaning of important features. In this situation, learning approaches can be taken into account to obtain the rules implicitly and dynamically. However, the difficulty facing many learning approaches is that the labeled training sample set may be very small. In the application of image retrieval, there are a limited number of labeled training samples given by the query and relevance feedback, so that it is difficult to learn some concepts. Pure supervised learning from such a small training data set will have poor generalization performance.

However, there are a large number of unlabeled images in the given database, which can be used to help supervised learning. Unlabeled data contain information about the joint distribution over features. If the probabilistic structure of data distribution is known, parameters of probabilistic models can be estimated by unsupervised learning alone, but it is still impossible to assign class labels without labeled data [3]. This fact suggests that labeled data (if enough) can be used to label the class and unlabeled data can be used to estimate the parameters of generative models.

In such circumstance, the hybrid training data set $\mathcal{D}$ consists of a labeled data set $\mathcal{L} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, N\}$, where $\mathbf{x}_i$ is its feature vector, y_i is its label and N is the size of the set, and an unlabeled data set $\mathcal{U} = \{\mathbf{x}_i, i = 1, \ldots, M\}$, where M is the size of the set. In image retrieval, the query images act as the labeled data, and the whole database or a subset can be treated as the unlabeled set.

In this sense, image retrieval is formulated as a *Transductive Problem*, which is to generalize the mapping function learned from the labeled training data set $\mathcal{L}$ to a specific unlabeled data set $\mathcal{U}$. We make an assumption here that $\mathcal{L}$ and $\mathcal{U}$ are from the same distribution. This assumption is reasonable, because the query images are drawn from the same image database. Essentially, image retrieval is to classify the images in the database by:

$$y_i = arg \max_{j=1,\ldots,C} p(y_j|\mathbf{x}_i, \mathcal{L}, \mathcal{U} : \forall \mathbf{x}_i \in \mathcal{U}) \qquad (1)$$

where C is the number of classes, and $C = 2$ for image retrieval. In this sense, we do not care the performance of the classifier over images outside the given database.

3 Using Unlabeled Data

The Expectation-Maximization (EM) approach can be applied to this transductive learning problem, since

the labels of unlabeled data can be treated as missing values. We assume that the hybrid data set is drawn from a mixture density distribution of C components $\{c_j, j = 1, \ldots, C\}$, which are parameterized by $\mathbf{\Theta} = \{\theta_j, j = 1, \ldots, C\}$. The mixture model can be represented as:

$$p(\mathbf{x}|\mathbf{\Theta}) = \sum_{j=1}^{C} p(\mathbf{x}|c_j; \theta_j) p(c_j|\theta_j) \tag{2}$$

where $\mathbf{x}$ is a sample drawn from the hybrid data set $\mathcal{D} = \mathcal{L} \bigcup \mathcal{U}$. We make another assumption that each component in the mixture density corresponds to one class, i.e. $\{y_j = c_j, j = 1, \ldots, C\}$.

Since the training data set $\mathcal{D}$ is a union of a set of labeled data set $\mathcal{L}$ and a set of unlabeled set $\mathcal{U}$, the joint probability density of the hybrid data set can be written as:

$$p(\mathcal{D}|\mathbf{\Theta}) = \prod_{\mathbf{x}_i \in \mathcal{U}} \sum_{j=1}^{C} p(c_j|\mathbf{\Theta}) p(\mathbf{x}_i|c_j; \mathbf{\Theta}) \bullet$$
$$\prod_{\mathbf{x}_i \in \mathcal{L}} p(y_i = c_i|\mathbf{\Theta}) p(\mathbf{x}_i|y_i = c_i; \mathbf{\Theta}) \tag{3}$$

This equation holds when we assume that each sample is independent to others. The first part of Equation 3 is for the unlabeled data set, and the second part is for the labeled data.

The parameters $\mathbf{\Theta}$ can be estimated by maximizing *a posteriori* probability $p(\mathbf{\Theta}|\mathcal{D})$. Equivalently, this can be done by maximizing $\lg(p(\mathbf{\Theta}|\mathcal{D}))$. Let $l(\mathbf{\Theta}|\mathcal{D}) = \lg(p(\mathbf{\Theta})p(\mathcal{D}|\mathbf{\Theta}))$, and we have

$$l(\mathbf{\Theta}|\mathcal{D}) = \lg(p(\mathbf{\Theta})) + \sum_{\mathbf{x}_i \in \mathcal{U}} \lg(\sum_{j=1}^{C} p(c_j|\mathbf{\Theta}) p(\mathbf{x}_i|c_j; \mathbf{\Theta}))$$
$$+ \sum_{\mathbf{x}_i \in \mathcal{L}} \lg(p(y_i = c_i|\mathbf{\Theta}) p(\mathbf{x}_i|y_i = c_i; \mathbf{\Theta})) \tag{4}$$

Since the log of sum is hard to deal with, a binary indicator $\mathbf{z}_i$ is introduced, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iC})$. And $z_{ij} = 1$ iff $y_i = c_j$, and $z_{ij} = 0$ otherwise, so that

$$l(\mathbf{\Theta}|\mathcal{D}, \mathcal{Z}) = \lg(p(\mathbf{\Theta}))$$
$$+ \sum_{\mathbf{x}_i \in \mathcal{D}} \sum_{j=1}^{C} z_{ij} \lg(p(O_j|\mathbf{\Theta}) p(\mathbf{x}_i|O_j; \mathbf{\Theta}))$$

The EM algorithm can be used to estimate the probability parameters $\mathbf{\Theta}$ by an iterative hill climbing procedure, which alternatively calculates $E(\mathcal{Z})$, the expected values of all unlabeled data, and estimates the parameters $\mathbf{\Theta}$ given $E(\mathcal{Z})$. The EM algorithm generally reaches a local maximum of $l(\mathbf{\Theta}|\mathcal{D})$. It consists of two iterative steps:

- E-step: set $\hat{\mathcal{Z}}^{(k+1)} = E[\mathcal{Z}|\mathcal{D}; \hat{\Theta}^{(k)}]$
- M-step: set $\hat{\Theta}^{(k+1)} = arg\ max_\theta\, p(\Theta|\mathcal{D}; \hat{\mathcal{Z}}^{(k+1)})$

where $\hat{\mathcal{Z}}^{(k)}$ and $\hat{\Theta}^{(k)}$ denote the estimation for $\mathcal{Z}$ and Θ at the k-th iteration respectively.

When the size of the labeled set is small, EM basically performs an unsupervised learning, except that labeled data are used to identify the components. If the probabilistic structure, such as the number of components in mixture models, is known, EM could estimate true parameters of the probabilistic model. Otherwise, the performance can be very bad.

Generally, when we do not have such *a prior* knowledge about the data distribution, a Gaussian distribution is always assumed to represent a class. However, this assumption is often invalid in practice, which is partly the reason that unlabeled data hurt the classifier.

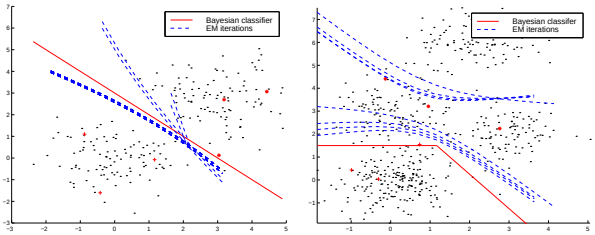
Figure 1 shows a simple example. In Figure 1.a, there are two classes of data drawn from two Gaussian distributions respectively, and only six samples are labeled. EM assumes Gaussian for both classes. The iteration begins with a weak classifier learned from these labeled samples. This weak classifier is used to estimate the labels of all the other unlabeled samples. Then, all these data are employed to learn a new classifier, which labels the unlabeled samples again in next iteration. In this special case, EM converges to the Bayesian classifier. On the other hand, if the guess of probabilistic structure is not correct, EM may not give a good estimation. In Figure 1.b, one class of data are drawn from a 3-component Gaussian mixtures, but the model still assumes Gaussian distribution. EM fails to give a good classifier.

4 D-EM Algorithm

EM often fails when structure assumption does not hold. One approach to this problem is to try every possible structure and select the best one. However, it needs more computational resources. An alternative is to find a mapping such that the data are clustered in the mapped data space, in which the probabilistic structure could be simplified and captured by simpler Gaussian mixtures.

4.1 Multiple Discriminant Analysis

Multiple Discriminant Analysis (MDA) [3] is a natural generalization of Fisher's linear discrimination (LDA) in the case of multiple classes. MDA offers many advantages and has been successfully applied to many tasks such as face recognition. The basic idea behind MDA is to find a linear transformation $\mathbf{W}$ to



(a) (b)

Figure 1: "." represents unlabeled sample. "+" and "*" denotes labeled sample. Six samples are labeled. Solid lines are Bayesian classifier, and dash lines are the iteration results of EM. (a) Data are drawn from two Gaussian distributions. EM converges to the Bayesian classifier. (b) One class of data is drawn from a 3-component Gaussian mixture, but EM still assumes Gaussian. One component is mislabeled. EM fails and unlabeled data do not help.

map the original d_1 dimensional data space to a new d_2 space such that the ratio between the between-class scatter and within-class scatter is maximized in the new space.

$$\mathbf{W} = arg\,\max_{\mathbf{W}} \frac{|\mathbf{W}^T S_b \mathbf{W}|}{|\mathbf{W}^T S_w \mathbf{W}|} \tag{5}$$

Suppose $\mathbf{x}$ is an m-dimensional random vector drawn from C classes in the original data space. The ith class has a probability P_i, a mean vector $\mathbf{m}_i$. The within-class scatter matrix S_w is defined by

$$S_w = \sum_{i=1}^{C} P_i E[(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T | c_i] \tag{6}$$

where c_i denotes the i-th class. The between-class scatter matrix S_b defined by

$$S_b = \sum_{i=1}^{C} P_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \tag{7}$$

where the grand mean $\mathbf{m}$ is defined as $\mathbf{m} = E[\mathbf{x}] = \sum_{i=1}^{C} P_i \mathbf{m}_i$. Details can be found in [3].

MDA offers a means to catch major differences between classes and discount factors that are not related to classification. Some features most relevant to classification are automatically selected or combined by the linear mapping $\mathbf{W}$ in MDA, although these features may not have substantial physical meanings any more. Another advantage of MDA is that the data are clustered to some extent in the projected space, which makes it easier to select the structure of Gaussian mixture models.

4.2 D-EM Algorithm

It is apparent that MDA is a supervised statistical method, which requires enough labeled samples to estimate some statistics such as mean and covariance. However, when the available labeled data are not enough, it is difficult to expect MDA to output good results.

By combining MDA with the EM framework, our proposed method, Discriminant-EM algorithm (D-EM), supplies MDA enough labeled data by combining supervised and unsupervised paradigms. The basic idea of D-EM is to identify some "similar" samples in the unlabeled data set to enlarge the labeled data set so that supervised techniques are made possible in such an enlarged labeled set.

D-EM begins with a weak classifier learned from the labeled set. Certainly, we do not expect much from this weak classifier. However, for each unlabeled sample $\mathbf{x}_j$, the classification confidence $\mathbf{w}_j = \{w_{jk}, k = 1,\ldots,C\}$ can be given based on the probabilistic label $\mathbf{l}_j = \{l_{jk}, k = 1,\ldots,C\}$ assigned by this weak classifier.

$$l_{jk} = \frac{p(\mathbf{W}^T\mathbf{x}_j|c_k)p(c_k)}{\sum_{k=1}^{C} p(\mathbf{W}^T\mathbf{x}_j|c_k)p(c_k)} \tag{8}$$

$$w_{jk} = \lg(p(\mathbf{W}^T\mathbf{x}_j|c_k))\ \ k = 1,\ldots,C \tag{9}$$

Euqation(9) is just a heuristic to weight unlabeled data $\mathbf{x}_j \in \mathcal{U}$, although there may be many other choices.

After that, MDA is performed on the new weighted data set $\mathcal{D}' = \mathcal{L} \bigcup \{\mathbf{x}_j, \mathbf{l}_j, \mathbf{w}_j : \forall \mathbf{x}_j \in \mathcal{U}\}$, by which the data set $\mathcal{D}'$ is linearly projected to a new space of dimension $C-1$ but unchanging the labels and weights, $\hat{\mathcal{D}} = \{\mathbf{W}^T\mathbf{x}_j, y_j : \forall \mathbf{x}_j \in \mathcal{L}\} \bigcup \{\mathbf{W}^T\mathbf{x}_j, \mathbf{l}_j, \mathbf{w}_j : \forall \mathbf{x}_j \in \mathcal{U}\}$. Then parameters $\mathbf{\Theta}$ of the probabilistic models are estimated by maximizing a posteriori probability on $\hat{\mathcal{D}}$, so that the probabilistic labels are given by the Bayesian classifier according to Equation(8). The D-EM algorithm iterates over these three steps, "Expectation-Discrimination-Maximization". The algorithm can be terminated by several methods such as presetting the iteration times, comparing a threshold and the difference of the parameters between two consecutive iterations, and using cross-validation. The following is the description of the D-EM algorithm.

Discriminant-EM algorithm (D-EM)
inputs: labeled set $\mathcal{L}$, unlabeled set $\mathcal{U}$
output: classifier with parameters $\mathbf{\Theta}$
begin Initialize: number of components C
  $\mathbf{W} \leftarrow MDA(\mathcal{L})$
  $lset \leftarrow Projection(\mathbf{W}, \mathcal{L})$

$uset \leftarrow Projection(\mathbf{W}, \mathcal{U})$
$\mathbf{\Theta} \leftarrow MAP(lset)$
D-E-M iteration
E-step:
$plabel \leftarrow Labeling(\mathbf{\Theta}, uset)$
$weight \leftarrow Weighting(plabel)$
$\mathcal{D}' \leftarrow \mathcal{L} \bigcup \{\mathcal{U}, plabel, weight\}$
D-step:
$\mathbf{W} \leftarrow MDA(\mathcal{D}')$
$lset \leftarrow Projection(\mathbf{W}, \mathcal{L})$
$uset \leftarrow Projection(\mathbf{W}, \mathcal{U})$
$\hat{\mathcal{D}} \leftarrow lset \bigcup \{uset, plabel, weight\}$
M-step:
$\mathbf{\Theta} \leftarrow MAP(\hat{\mathcal{D}})$
return $\mathbf{\Theta}$
end

It should be noted that the simplification of probabilistic structures is not guaranteed in MDA. If the components of data distribution are mixed up, it is very unlikely to find such a linear mapping.

4.3 Image Retrieval by D-EM

By the approach of relevance feedback in image retrieval, several relevant and irrelevant examples are labeled by human. Generally, it is under a large risk to weight image features by such a small labeled data set, since the similarity among these images would be vague. Using a random subset of the database or even the whole database as an unlabeled data set, the D-EM algorithm identifies some "similar" images to the labeled images to enlarge the labeled data set. Therefore, good discriminating features could be automatically selected through this enlarged training data set to better represent the implicit concepts.

The application of D-EM to image retrieval is straightforward. In our current implementation, in the transformed space, both classes are represented by a Gaussian distribution with three parameters, the mean μ_i, the covariance $\mathbf{\Sigma}_i$ and *a priori* probability of each class P_i. The D-EM iteration tries to boost an initial weak classifier.

5 Experiments

In order to give some analysis and compare several different methods, we manually label an image database of 134 images, which is a subset of the COREL database. Our dataset has 7 classes such as airplane, bird, car, church painting, flower, mountain view and tiger. All images in the database have been labeled as one of these classes. In all the experiments, these labels for unlabeled data are only used to calculate classification error.

To investigate the effect of the unlabeled data used in D-EM, we feed the algorithm a different number of labeled and unlabeled samples. The labeled images are obtained by relevance feedback. When using more than 100 unlabeled samples, the error rates drop to less than 10%. From Figure 2, we find that D-EM brings about 20% to 30% more accuracy. In general, combining some unlabeled data can largely reduce the classification error when labeled data are very few.

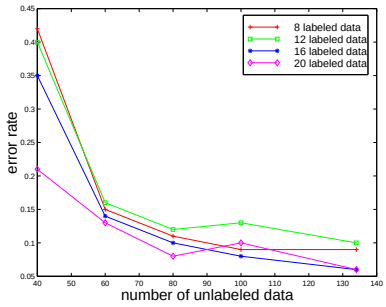


Figure 2: The effect of labeled and unlabeled data in D-EM. Error rate decreases when adding more unlabeled data. Combining some unlabeled data can largely reduce the classification error.

We test and compare four methods. The first one is to weight each features by relevance feedback (WRF) [12], in which 37 physical features which are precalculated and pre-stored. The top 20 most similar images are obtained through ranking each image by comparing the Mahalanobis distances to the means of query images. The second method is a simple probabilistic method (SP), in which both classes (relevant and irrelevant) are assumed Gaussian distributions, and the model parameters are estimated by feedback images. The third method is the basic EM (EM) algorithm, which assumes Gaussian distributions for both classes. The fourth is the D-EM algorithm. In the last three probabilistic methods, the label of each image is given by maximizing *a posteriori* probability, $l_j = arg\,max_k\, p(c_k|\mathbf{x}_j)$.

We also compare a set of physical features (P-Features) and mathematical features (M-Features). We use the same physical features as that in WRF [12], in which 9 color features include the mean, std and skew of the HSV space, 10 texture features are extracted by wavelets, and 18 structure features are represented by the statistics of the edge map. The mathematical features are extracted by PCA, in which the number of principal components is 30, and the resolution of image is reduced to 20×20. Except for WRF, both P-Features and M-Features are tested.

These four methods are compared on this fully la-

beled database. Classification error for each method is calculated for evaluation, although these errors are not available for the training. Suppose the database has N samples, C classes, and the k-th class has N_k samples, and $N = \sum_{k=1}^{C} N_k$. The method to calculate error in WRF is different from the other three methods. In WRF, if the query images belong to the j-th class, and m_j samples in the top N_j belongs to the j-th class, the error for this query is defined as $e_j = 2(N_j - m_j)/N$. In the other three methods, if there are m samples in total that are not correctly labeled, the error is defined as $e_j = m/N$. The average error is obtained by averaging over M experiments, i.e. $e = \sum_{j=1}^{M} e_j/M$.

Algorithm	P-Features	M-Features
WRF	6.3%	N/A
SP	21.2%	15.7%
EM	23.4%	25.8%
D-EM	3.9%	5.3%

Table 1: Error rate comparison among different algorithms. All comparisons are based on the first time relevance feedback with 6 relevant and 6 irrelevant images. D-EM outperforms the other three methods.

Our algorithm is also tested by several large databases. The COREL database contains more than 70, 000 images over a wide range of more than 500 categories with 120×80 resolution. The VISTEX database is a collection of 832 texture images. Satisfactory results are obtained.

6 Conclusion

The gap between high-level concepts and low-level visual features is one of the difficulties of CBIR. Different from other methods in image retrieval, our approach formulates it as a transductive learning problem, in which the unlabeled samples in the given database combined with labeled data are both used in training. The proposed method, Discriminant-EM algorithm (D-EM), approaches this problem in the EM framework. In D-EM, the assumption of probabilistic structure in EM is relaxed and the most relevant features to classification can be automatically selected. Our experiments show that the D-EM algorithm could be an effective way to CBIR. This algorithm can be easily expanded to retrieve other media types.

One of the future research directions of this approach is to explore the non-linear case of MDA. The proposed approach needs to be tested on more databases. To accelerate the algorithm, the size of the unlabeled data set could decrease through the iteration.

7 Acknowledgment

This work was supported in part by National Science Foundation Grants CDA-96-24396, IRI-96-34618 and EIA-99-75019. The authors would like to appreciate the anonymous reviewers for their comments.

References

[1] K.Bennett, A.Demiriz, "Semi-Supervised Support Vector Machines", *Proc. NIPS*, Denver, 1998

[2] S.Chang, J.Smith, M.Beigi and A.Benitez, "Visual Information Retrieval from Large Distributed Online Repositories", *Communications of ACM*, Dec. pp.12-20, 1997

[3] R.Duda and P.Hart, "Pattern Classification and Scene Analysis", New York:Wiley, 1973 (The 2nd Version with D.Stork unpublished)

[4] A.Gammerman, V.Vapnik, V.Vowk, "Learning by Transduction", *Conf. Uncertainty in Artificial Intelligence*, pp.148-156, 1998

[5] T.Joachims, "Transductive Inference for Text Classification using Support Vector Machines", *Int'l Conf. on Machine Learning (ICML)* 1999

[6] K.Nigam, A,Mccallum, S.Thrun, T.Mitchell, "Text Classification from Labeled and Unlabeled Documents Using EM", *Machine Learning*, 1999

[7] B.Manjunath and W. Ma, "Texture Features for Browsing and Retrieval of Image Data", *IEEE T-PAMI*, Nov. 1996

[8] T.Mitchell, "The Role of Unlabeled Data in Supervised Learning", *Proc. Sixth Int'l Colloquium on Cognitive Science*, Spain, 1999

[9] A.Pentland, R.Picard and S.Sclaroff, "Photobook: Content-based Manipulation of Image Database", *Int'l Journal of Computer Vision*, 1996

[10] M.Popescu and P.Gader, "Image Content Retrieval From Image Database Using Feature Integration by Choquet Integral", *SPIE Storage and Retrieval for Image and Video Database*, VII, 1998

[11] Y.Rui, T.Huang and S.Chang, "Image Retrieval: Current Techniques, Promising Directions and Open Issues", *Journal of Visual Communication and Image Representation*, Vol.10, pp.1-23, 1999

[12] Y.Rui, T.Huang, M.Ortega, S.Mehrotra, "Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval", *IEEE Circuits and Systems for Video Technology*, Vol 8, No.5, pp644-655, 1998

[13] S.Santini and R.Jain, "Similarity Measures", *IEEE T-PAMI*, Vol.21, No.9, 1999

[14] D.Swets, J.Weng, "Hierarchical Discriminant Analysis for Image Retrieval", *IEEE T-PAMI*, Vol.21, No.5, pp.386-400, 1999

[15] V.Vapnik, "The Nature of Statistical Learning Theory", Springer-Verlag, 1995

[16] H.Zhang, D.Zhong, "A Scheme for Visual Feature Based Image Retrieval", *Proc. SPIE Storage and Retrieval for Image and Video Database*, 1995