Incorporate Discriminant Analysis with EM Algorithm in Image Retrieval

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

Abstract

One of the difficulties of Content-Based Image Retrieval (CBIR) is the gap between high-level concepts and low-level image features, e.g., color and texture. Relevance feedback was proposed [1] to take into account of the above characteristics in CBIR. Although relevance feedback incrementally supplies more information for fine retrieval, two challenges exist: (1) the labeled images from the relevance feedback are still very limited compared to the large unlabeled images in the image database. (2) relevance feedback does not offer a specific technique to automatically weight the low-level feature. In this paper, image retrieval is formulated as a transductive learning problem by combining unlabeled images in supervised learning to achieve better classification. Experimental results show that the proposed approach has a satisfactory performance for image retrieval applications.

1. Introduction

With the advances in technology to generate, transmit, and store large amounts of digital images and video, research in content-based image retrieval (CBIR) has gained more attention recently. In CBIR, images are indexed by their visual contents such as color, texture, etc. Many research efforts have been made to extract these low-level features [2, 3], evaluate distance metrics [4, 5] and look for efficient searching schemes [6, 7].

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. Relevance feedback was proposed [1] to take into account the above characteristics. Although relevance feedback incrementally supplies more information for fine retrieval, two challenges exist: (1) the labeled images from relevance feedback are still very limited compared to the large unlabeled images in the image database. (2) Relevance feedback does not offer a specific technique to automatically weight the low-level feature in CBIR.

A possible approach to this problem is machine learning, by which the mapping could be learned through a set of examples. In this paper, to obtain a possible better high-level concept from several given images, image retrieval problem is formulated as a *transductive* learning problem. The Expectation-Maximization (EM) algorithm can be applied to this transductive learning problem [8]. Based on the EM framework and discriminant analysis, the proposed approach employs both labeled images (from relevance feedback) and unlabeled images (from image database). It not only estimates the parameters of a generative model, but also estimates a linear transformation that maps the original feature space to a new feature space. The role of the linear transformation is to relax the assumption of the probabilistic structure of data distribution as well as to construct a new set of features that is "best" for the classification.

The rest of the paper is organized as follows. Our approach is described in Section 2. Image retrieval with the proposed algorithm and experimental results are discussed in Section 3 and 4, respectively. Conclusions and future work are given in Section 5.

2. Our Approach

In the application of image retrieval, there are a limited number of labeled training image samples given by the query and relevance feedback so that it is difficult to learn the similarities. Therefore pure supervised learning 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. In such circumstance, the hybrid training dataset D consists of a labeled data set $L=\{(\mathbf{x}_i, y_i), i=1,\ldots,N\}$, where N is the size of the set, $\mathbf{x}_i$ is the feature vector and y_i is its label, and an unlabeled data set $U=\{\mathbf{x}_i, i=1,\ldots,M\}$, 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 dataset L to a specific unlabeled data set U. We assume that the hybrid dataset is drawn from a mixture density distribution of C components $\{c_j, j=1,\ldots,C\}$, which are parameterized by $\Theta=\{\theta_j, j=1,\dots,C\}$. The mixture model can be represented as:

$$p(\mathbf{x}\mid\Theta)=\sum_{j=1}^{C}p(\mathbf{x}\mid c_j;\theta_j)p(c_j\mid\theta_j) \qquad (1)$$

where $\mathbf{x}$ is a sample drawn from the hybrid data set $D=L\cup U$. We make another assumption that each component in the mixture density corresponds to one class, i.e., $\{y_j=c_j, j=1,\dots,C\}$. Essentially, image retrieval is to classify the images in the database by:

$$y_i=\arg\max_{j=1,\dots,C} p(y_j\mid\mathbf{x}_i,L,U:\forall\mathbf{x}_i\in U\} \qquad (2)$$

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. Image Retrieval by D-EM Algorithm

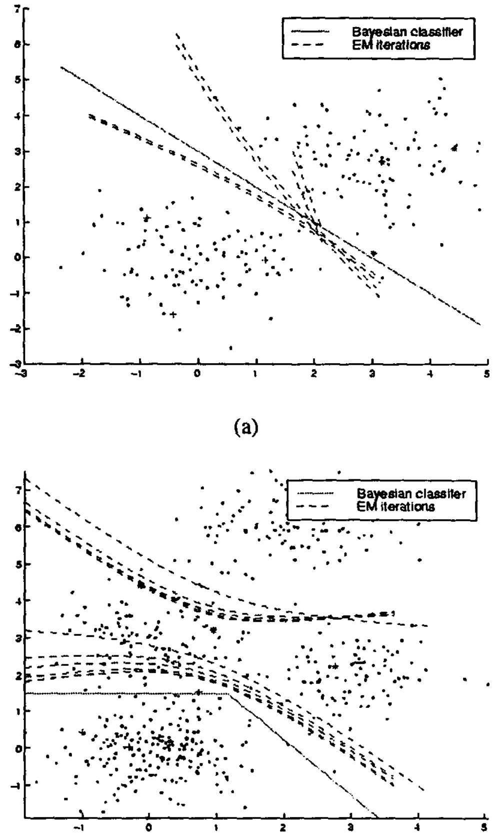
The Expectation-Maximization (EM) algorithm can be applied to this transductive learning problem, since the labels of unlabeled data can be treated as missing values. The parameters Θ can be estimated by an iterative hill climbing procedure [8].

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 probabilistic models. Otherwise, the performance can be very bad.

Generally, when we do not have such *a priori* knowledge about the data distribution, a Gaussian distribution is always assumed to represent a class, which is a special case of the assumption of one-to-one component-class correspondence in the generative model. However, this assumption is often invalid in practice, which is partly the reason that unlabeled set hurts the classifier.

Figure 1 shows a simple example. In Fig. 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 sample 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 Fig. 1(b), one class of data is drawn from 3-component Gaussian mixtures, but the model still assumes Gaussian distribution. EM fails to give a good classifier.

One possible approach is multiple discriminant analysis (MDA). Multiple discriminant analysis [9] is a natural generalization of Fisher's linear discrimination in the cases of multiple classes. The basic idea behind MDA is to find a linear transformation $\boldsymbol{W}$ to 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.



(a)

(b)

Figure 1 "." Denotes unlabeled samples. "+" 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.

MDA offers a means to catch major differences between classes and discount factors that are not related to classification. Another advantage of MDA is that the data are clustered in the projected space to some extent, which makes it easier to select the structure of Gaussian mixture models.

It is apparent that MDA is a supervised statistical method, which requires enough labeled samples to estimate mean and covariance. MDA offers many advantages and has been successfully applied to many tasks such as face recognition. However, when the labeled data from the query and the relevance feedback are not enough, it is difficult to expect MDA to achieve good results.

By combining MDA with the EM framework, our proposed method is such a way to combine supervised and unsupervised paradigms. The basic idea is to identify some "similar" samples in the unlabeled data set to enlarge the labeled data set so that supervised techniques can be applied in such an enlarged labeled set.

Our method begins with a weak classifier learned from the labeled data 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.

$$l_{jk} = \frac{p(\mathbf{x}_j|c_k)p(c_k)}{\sum_{k=1}^{C} p(\mathbf{x}_j|c_k)p(c_k)} \qquad (3)$$

$$w_{jk} = \log(p(\mathbf{x}_j \mid c_k)) \qquad k = 1,\ldots,C \qquad (4)$$

Eq. (4) is just a heuristic to weight unlabeled data $\mathbf{x}_j \in U$, although there may be many other choices. Then, MDA is performed on the new weighted data set $D' = L \cup \{\mathbf{x}_j, \mathbf{l}_j, \mathbf{w}_j : \forall \mathbf{x}_j \in U\}$, by which the data set D' is linearly projected to a new space $\hat{D}$ of dimension $C-1$ but unchanging the labels and weights.

$$\hat{D} = \{\mathbf{W}^T\mathbf{x}_j, y_j : \forall \mathbf{x}_j \in L\} \cup \{\mathbf{W}^T\mathbf{x}_j, \mathbf{l}_j, \mathbf{w}_j : \forall \mathbf{x}_j \in U\} \ (5)$$

Then parameters Θ of the probabilistic models are estimated by maximizing *a posteriori* probability on $\hat{D}$, so that the probabilistic labels are given by the Bayesian classifier according to Eq. (3).

This approach consists of a three-step loop, Expectation-Discrimination-Maximization. The algorithm can be terminated by several methods such as presetting the iteration times, thresholding the difference of the parameters between consecutive two iterations, and using cross-validation.

4. Results

We manually labeled an image database of 134 images, which is a subset of COREL database. The manually labeled dataset has 7 classes such as car, flower, mountain, airplane, church painting, tiger and bird. 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 test the algorithm performance, different numbers of labeled images and unlabeled images are fed into the D-EM algorithm. Figure 2 shows the error rate for bird and non-bird classification. The following conclusions can be obtained: (1) For a fixed number of labeled data, incorporating more unlabeled data in the training will greatly reduce the classification error. The error rate drops below 10% when using more than 100 unlabeled samples. The error rate is gradually becoming flat when the number of unlabeled samples exceeds a certain number, e.g., 100 in this example. This fact can be very useful, especially for large image database. It means that a subset of the database could be a good representation for the whole database if the number of samples in the subset exceeds some threshold. The parameters of the generative model can be estimated from a subset of the large database instead of the whole database. (2) For a fixed number of unlabeled data, increasing the number of labeled data generally results in the decreased error rate. This is easy to understand. With more *a priori* known information, the smaller error rate will be. In general, combining some unlabeled data can greatly reduce the classification error when labeled data are very few.

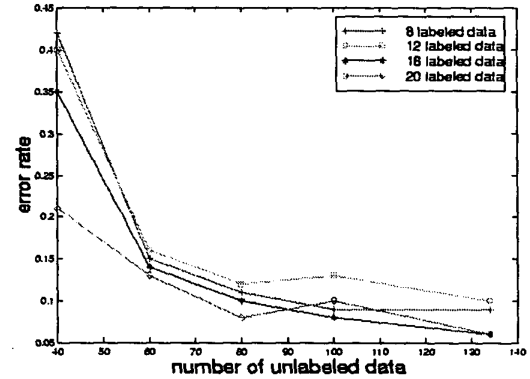


Figure 2: Error rate decreases when more unlabeled data are available. Combining more unlabeled data can greatly reduce the classification error

Four algorithms are compared (1) Weight each feature by relevance feedback (WRF) [1]. 37 color (9 color moments), texture (10 wavelet moments) and structure features (18 water-filling feature) are extracted from each image. The top 20 most similar images are obtained through ranking each image by comparing the *Mahalanobis* distances to the mean of the query images. (2) A simple probabilistic method (SP), in which both classes (relevant and irrelevant) are assumed Gaussian distributions, and the model parameters, *W* and Θ are estimated by feedback images, i.e., labeled data only. The unlabeled data is not used. (3) The basic EM algorithm, which assumes Gaussian distributions for both classes (4) Our proposed discriminant analysis with EM algorithm (D-EM). In the last three probabilistic methods, the label of each image is given by maximizing *a posteriori* probability, $\mathbf{l}_j = \arg\max_k p(c_k \mid \mathbf{x}_j)$.

We also compare a set of physical features and mathematical features. The physical features are color, texture and structure that are same as in WRF [1]. The mathematical features are extracted by PCA, in which the number of principle components is 30, and the resolution of image is reduced to 20×20.

These four algorithms are compared on this fully labeled database. Classification errors for each method are calculated for evaluation. Suppose the database has N images, I classes, and the k-th class has N_k images. Note that $N = \sum_{k=1}^{I} N_k$. The error rate for the last three methods is calculated as:

$$e_j = \frac{m}{N} \tag{6}$$

where m is the total number of samples that are not correctly labeled in the all N images.

The error rate for 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 rate for this query is defined as

$$e_j = \frac{2(N_j - m_j)}{N} \tag{7}$$

The average error is obtained by averaging M experiments, i.e.

$$e = \frac{1}{M} \sum_{k=1}^{M} e_j \tag{8}$$

Table 1 shows the error rate comparisons for the four algorithms.

Table 1. Error Rate Comparisons for different algorithms. All comparisons are based on the first time relevance feedback with 6 relevant and 6 irrelevant images. Clearly, D-EM performs best.

Algorithm	Physical-Features	Mathematical-Features
WRF	6.3%	N/A
SP	21%	16%
EM	23%	26%
D-EM	3.9%	5.3%

5. Conclusions

In this paper, image retrieval is formulated as a transductive learning problem, in which the unlabeled images in the given database combined with labeled images are used in training. The Discriminant-EM algorithm (D-EM) approaches this problem in the EM framework by taking advantage of a generative model. A linear transformation is used in the D-EM algorithm. This transformation is obtained by discriminant analysis such that the probability assumption of the data distribution is relaxed. Our preliminary experiments show that the D-EM algorithm could be an effective way for Content-Based Image Retrieval. Combining the database with queries can greatly enhance the accuracy of relevant/irrelevant classification, and therefore, the quality of image retrieval.

A small image database is used for testing the algorithm performance in this paper. In our future work, we will apply D-EM algorithm to large databases. This algorithm will also be applied to retrieve other media types in our future work.

6. Acknowledgements

This work was supported in part by National Science Foundation Grants CDA-96-24396, EIA-99-75019 and IRI-96-34618.

References

[1] Y. Rui, T. S. Huang, et al., "Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval", *IEEE Circuits and systems for Video technology*, vol. 8, no.5, 1999.

[2] Michael Swain and Dana Ballard, "Color Indexing", International Journal of Computer Vision, 7(1), 11-32,1991.

[3] Calvin C. Gotlieb and Herbert E. Kreyszig, "Texture Descriptors based on co-occurrence matrices", *Computer Vision, Graphics, and Image Processing*, 51, 1990.

[4] S. Santini and R. Jain, "Similarity Measures", *IEEE PAMI*, vol. 21, no.9, 1999.

[5] M. Popescu and P. Gader, "Image Content Retrieval From Image Databases Using Feature Integration by Choquet Integral", in *SPIE Conference Storage and Retrieval for Image and Video Databases VII*, San Jose, CA, 1998.

[6] D. M. Squire, H. Müller, and W. Müller, "Improving Response Time by Search Pruning in a Content-Based Image Retrieval System, Using Inverted File Techniques", *Proc. of IEEE workshop on CBVIAL*, Fort Collins, June 1999.

[7] D. Swets, J. Weng, "Hierarchical Discriminant Analysis for Image Retrieval", *IEEE PAMI*, vol. 21, no.5, 1999

[8] Y. Wu, T. S. Huang, "Using Unlabeled Data in Supervised Learning by Discriminant-EM Algorithm", *NIPS'99 workshop on using unlabeled data for supervised learning*, CO, 1999.

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