Integrating Unlabeled Images for Image Retrieval Based on Relevance Feedback

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

Abstract

Retrieval techniques based on pure similarity metrics are often suffered from the scales of image features. An alternative approach is to learn a mapping based on queries and relevance feedback by supervised learning. However, the learning is plagued by the insufficiency of labeled training images. Different from most current research in image retrieval, this paper investigates the possibility of taking advantage of unlabeled images in the given image database to make feasible a hybrid statistical learning. Assuming a generative model of the database, the proposed approach casts image retrieval as a transductive learning problem in a probabilistic framework. Our experiments show that the proposed approach has a satisfactory performance in image retrieval applications.

1. Introduction

To avoid manual annotating large image databases, an alternative approach of retrieving images 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 [1, 4, 9], evaluate distance metrics [7, 10], and look for efficient searching schemes [11, 14].

However, images are too rich to represent by these low-level 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 [11]. Both representations confront 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.

The mapping between them would be highly nonlinear such that it is impractical to represent it explicitly. In this situation, learning approaches can be taken into account to learn the possible mapping implicitly and dynamically. However, in the application of image retrieval, there are a limited number of labeled training images given by the queries and relevance feedback, so that it is difficult to learn the image similarity measurement correctly. Pure supervised learning from such a small training data set will have poor generalization performance.

To obtain a possible better similarity measurement from several given images, this paper looks into the image retrieval problem in the perspective of transductive learning, and presents a probabilistic approach to employ both labeled images and unlabeled images. Based on the EM framework and discriminant analysis, the proposed algorithm, Discriminant-EM (D-EM), learns a generative model in a lower-dimensional subspace obtained by discriminant analysis, which relaxes the assumption of the probabilistic structure of the data distribution. A new formulation of the image retrieval problem is given in section 2. The proposed algorithm is presented in section 3. Experimental results and conclusion are given in 4 and 5, respectively.

2. Problem Formulation and EM

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. By the approach of relevance feedback in image retrieval, several relevant and irrelevant examples are labeled by the user.

Generally, it is under a large risk to perform supervised learning techniques on such a small labeled data set, since the similarity among these images would be vague such that the generalization would be very poor. However, when we weaken the requirement of generalization to a known sub-

0-7695-0750-6/00 $10.00 © 2000 IEEE

set of the whole data space, and provide more unsupervised data to describe this subset, a good generalization would be obtained on such subset, since here we do not care the generalization of the data outside this subset. For image retrieval, we try to learn an good image classifier in the sense of the given database.

The basic idea of our approach is to identify 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.

In such circumstance, we employ a hybrid training data set $\mathcal{D}$ which 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. Here, 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.

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 $\Theta = \{\theta_j, j = 1, \ldots, C\}$. The mixture model can be represented as:

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

where $\mathbf{x}$ is a sample drawn from the hybrid data set $\mathcal{D} = \mathcal{L} \bigcup \mathcal{U}$.

Let $\mathcal{Z} = \{\mathbf{z}_j : \forall \mathbf{x}_j \in \mathcal{U}\}$, in which $\mathbf{z}_j$ is the probabilistic label for the sample $\mathbf{x}_j$ in the unlabeled set $\mathcal{U}$. The EM algorithm can be used to estimate the probability parameters Θ by an iterative hill climbing procedure, which alternatively calculates $E(\mathcal{Z})$, the expected values of all unlabeled data, and estimates the parameters Θ given $E(\mathcal{Z})$. 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 probabilistic model parameters. 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 this direct EM method performs poor in many cases.

3. Discriminant-EM Algorithm

Since we generally do not know the probabilistic structure of data distribution, EM often fails when structure assumption does not hold. Instead of trying every possible structure in EM, 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.

Multiple Discriminant Analysis (MDA) [2] is a natural generalization of Fisher's linear discrimination (LDA) in the case of multiple classes. The basic idea behind MDA is to find a linear transformation $\mathbf{W}$ 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.

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 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.

It is apparent that MDA is a supervised statistical method, which requires a large number of labeled samples to estimate some statistics such as mean and covariance. 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.

D-EM algorithm 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 $\mathbf{z}_j = \{z_{jk}, k = 1, \ldots, C\}$ assigned by this weak classifier.

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

$$w_{jk} = \lg(p(\mathbf{x}_j|c_k)) \; k = 1, \ldots, C \tag{4}$$

Euqation(4) 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 $\mathcal{D}' = \mathcal{L} \bigcup \{\mathbf{x}_j, \mathbf{z}_j, \mathbf{w}_j : \forall \mathbf{x}_j \in \mathcal{U}\}$, by which the data set $\mathcal{D}'$ is linearly projected to a new space of dimension $C-1$ but unchanging the labels and weights, $\hat{\mathcal{D}} = \{\mathbf{W}^T\mathbf{x}_j, y_j : \forall \mathbf{x}_j \in \mathcal{L}\} \bigcup \{\mathbf{W}^T\mathbf{x}_j, \mathbf{z}_j, \mathbf{w}_j : \forall \mathbf{x}_j \in \mathcal{U}\}$. Then parameters Θ 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(3). The D-EM algorithm iterates over these three steps, "Expectation-Discrimination-Maximization".

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. 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 images are only used to calculate classification error.

To investigate the effect of the unlabeled training data used in D-EM, we feed the algorithm a different number of labeled and unlabeled images. 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 1, we find that D-EM brings about 20% to 30% more accuracy than without using any unlabeled images. In general, combining a number of unlabeled images can largely reduce the classification error when labeled data are very few.

We experimented with physical features (P-Features), which consist of 9 color features including the mean, std and skew of the HSV color space, 10 texture features extracted by wavelets, and 18 structure features represented by the statistics of the edge map[9]. The mathematical features (M-Features) are extracted by PCA, in which the number of principle components is 30, and the resolution of image is reduced to 20×20.

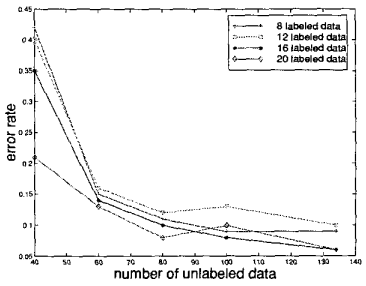

Figure 1. The effect of labeled and unlabeled data in D-EM. Error rate decreases when adding more unlabeled data. Combining some unlabeled data can largely reduce the classification error.

We test and compare four methods. The first method is to incrementally find a similarity measurement by weighting each feature from relevance feedback (WRF) [9], in which 37 physical features are pre-calculated and pre-stored. The top 20 most similar images are obtained through ranking each image by comparing the Mahalanobis distances to the mean of query images. The second method is a simple probabilistic method (SP) which only employs the labeled images. In this method, both classes (relevant and irrelevant) are assumed Gaussian distributions, and the model parameters are estimated by labeled images alone. The third method is the basic EM (EM) algorithm, which assumes Gaussian distributions for both classes. This method employs both labeled and unlabeled images, but it does not perform discriminant analysis and has to estimate the parameters of a high dimensional generative model. The fourth is the D-EM algorithm, which has been described in section 3. In the last three probabilistic methods, the label of each image is given by maximizing *a posteriori* probability (MAP), $l_j = arg\,max_k\, p(c_k|\mathbf{x}_j)$. Except for WRF, both P-Features and M-Features are tested.

These four methods are compared on this fully labeled 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

$$e_j = \frac{2(N_j - m_j)}{N} \tag{5}$$

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 on several large databases. The COREL database contains more than 17, 000 images. The VISTEX database is a collection of 832 texture images.

5. Conclusion

Different from many other methods in content-based image retrieval, our approach formulates it as a transductive learning problem, in which both the image queries and unlabeled images in the given database are employed the training of an image classifier. The proposed method, Discriminant-EM algorithm (D-EM), approaches this problem in the EM framework. Since the simple EM algorithm confronts several difficulties, such as learning in high dimensionality and probabilistic structure assumption, the D-EM algorithm introduces a Discrimination-step in the EM iteration to relax the assumption of the probabilistic structure of data distribution and automatically select the most relevant features to classification. Our experiments show that the D-EM algorithm could be an effective way to multimedia databases.

Future work should includes the study of the convergence and stability of the algorithm. Currently, D-EM uses a linear transformation, but non-linear transform may have better performance. Another future research direction of this approach is to explore the non-linear case of MDA. To accelerate the algorithm, the size of the unlabeled data set could decrease through the iteration. More large image databases should be tested by this approach.

6. Acknowledgment

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

References

[1] 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

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

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

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

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

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

[7] 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

[8] 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

[9] 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

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

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

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

[13] Y.Wu, T.S.Huang, "Using Unlabeled Data in Supervised Learning by Discriminant-EM Algorithm", *NIPS'99 Workshop on Using Unlabeled Data in Supervised Learning*, Colorado, 1999

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