Discriminant Features for Model-Based Image Databases

Anlei Dong and Bir Bhanu
Center for Research in Intelligent Systems
University of California, Riverside, California 92521, USA
{adong, bhanu}@cris.ucr.edu

Abstract

A challenging topic in content-based image retrieval is to determine the discriminant features that improve classification performance. An approach to learn concepts is by estimating mixture model for image databases using EM algorithm; however, this approach is impractical to be implemented for large databases due to the high dimensionality of the feature space. Based on the over-splitting nature of our EM algorithm and the Bayesian analysis of the multiple users' labelling information derived from their relevance feedbacks, we propose a probabilistic MDA to find the discriminating features, and integrate it with the EM framework. The experimental results on Corel images show the effectiveness of concept learning with the probabilistic MDA, and the improvement of the retrieval performance.

1 Introduction

Recently, several researchers working on content-based image retrieval (CBIR) adopt Gaussian mixture model (GMM) to model the database image distribution [1] [2] [3]. The knowledge of the mixture model for an image database may provide good classifiers and thus, improve retrieval performance.

One major approach to estimate mixture model is the *Expectation-Maximization* (EM) algorithm [4]. However, due to the high dimensionality of the feature space of image databases, the direct implementation of EM on the feature vectors is impractical for two reasons: (1) it is difficult in finding the most discriminating features; (2) the computation of EM is formidable. Wu and IIuang [5] propose to integrate *multiple discriminant analysis* (MDA) [6] with the EM framework for the hybrid (labelled and unlabelled) data, so that the weak classifiers are boosted by exploring discriminating features in a self-supervised fashion. This is called D-EM, in which the number of classes (clusters) is assumed to be known by the system, and each class provides some labelled images.

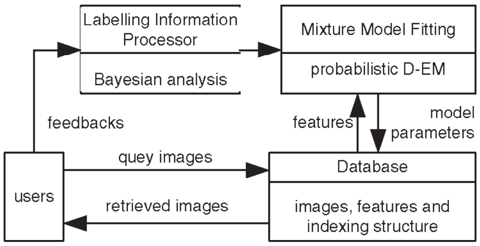


Figure 1: System diagram for concept learning .

However, to estimate the mixture model for a real image database, the number of classes is unknown in advance by the system. We propose an over-splitting EM algorithm to avoid estimating the number of classes. Our EM algorithm will be fed with the positive and negative labelling information derived from the relevance feedbacks by multiple users. As the labelling information provided by different users may be inconsistent, we propose to process and exploit it for learning based on Bayesian analysis.

For our over-splitting EM algorithm, the direct D-EM may mislead the system to find the most discriminating features. Fortunately, our Bayesian analysis on the users' labelling information may provide the probabilities between some mixture components in EM. Such probabilities may help to avoid the misleading effect on finding the most discriminating features.

Figure 1 illustrates the diagram for our system with concept learning by exploiting users' labelling information. The main contributions of this paper are (1) a probabilistic MDA to find the most discriminating features and (2) a framework for the semi-supervised EM algorithm integrating with the probabilistic MDA to achieve concept learning.

2 Technical approach

2.1 Labelling information analysis

We define the labelling information derived from the relevance feedback of a user as a *retrieval experience* $\mathcal{E} = \{\mathcal{X}^+, \mathcal{X}^-\}$, where $\mathcal{X}^+ = \{x_1^+, x_2^+, \ldots, x_{N+}^+\}$ are labeled as belonging to (positive for) a certain but unknown class while another portion of samples

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

IEEE COMPUTER SOCIETY

$\mathcal{X}^- = \{x_1^-, x_2^-, \ldots, x_{N^-}^-\}$ are labeled as NOT belonging to (negative for) that unknown class. Note that x_i^+ $(i = 1, 2, \ldots, N^+)$ and x_j^- $(j = 1, 2, \ldots, N^-)$ are image visual feature vectors.

It is necessary to find an efficient way to store the information contained in the retrieval experiences from multiple users. We assume that each database image belongs to one and only class (decided by the opinion of the majority of people), and we regard this assignment as the true labelling information, and all the other contradicting labelling information on this image as *labelling noise.* We define the *labelling noise rates* as

$\alpha = prob(I \notin \mathcal{E} | I \in C; \mathcal{E} \subset C)$, and

$\beta = prob(I \in \mathcal{E} | I \notin C; \mathcal{E} \subset C)$,

where "$\mathcal{E} \subset C$" means that the user who has provided retrieval experience $\mathcal{E}$ is seeking the images of Class C, "$I \in (\notin)\ C$" represents that Image I is (not) belonging to Class C, and "$I \in (\notin)\ \mathcal{E}$" denotes that Image I is labelled as positive (negative) in retrieval experience $\mathcal{E}$. Basically, α denotes the probability that a user labels an image as negative when this user is seeking a class and this image belongs to this class, and β denotes the probability that a user labels an image as positive when this user is seeking a class and this image does NOT belong to this class.

If α and β are known, some retrieval experiences are merged into a single concept experience using Bayesian analysis [7]. We denote the collection of such concept experiences as Φ. We can compute $prob(C_i, C_j)$, the probability that the ith and jth concept experiences are for the same concept, and such probability will be used for the following probabilistic MDA.

2.2 Probabilistic D-EM

2.2.1 Over-splitting: To estimate the mixture model using EM, the number of components for the algorithm has to be given, although the system usually does not know the real number of classes c_{real}. With the knowledge of highest possible number of classes c_{max} and the experience concept collection Φ with $|\Phi|$ being the number of elements in Φ, we can determine the initial number of components c for the EM algorithm as $c = max(c_{max}, |\Phi|)$. We initialize $|\Phi|$ of the c component centers to be the $|\Phi|$ experience concept centers, and randomly select $c - |\Phi|$ data as the remaining $c - |\Phi|$ component centers. We also initialize the c component covariances matrices to be identity matrices.

To evaluate the over-splitting clustering result $\mathcal{R} = \{\mathcal{R}_1, \ldots, \mathcal{R}_c\}$, we compare it with the groundtruth mixture model $\mathcal{C} = \{C_1, \ldots, C_{c_{real}}\}$ $(c > c_{real})$ by using a statistical index. A pair of vectors $\{x_i, x_j\}$ are referred to as (I) BS if both vectors belong to the same component in $\mathcal{C}$ and to the same cluster in $\mathcal{R}$, (II) BB if both vectors belong to the same component in $\mathcal{C}$. Let ξ_1 and ξ_2 be the number of BS and BB respectively, we use $SI = \frac{\xi_1}{\xi_2}$ to evaluate an over-splitting clustering result.

2.2.2 Probabilistic MDA: For the case of multiple classes (c classes), multiple discriminant analysis (MDA) [6] finds a linear transformation to map the original d-dimensional feature space to a new d'-dimensional feature space $(d' < d)$, by maximizing the objective function

$J(W) = \frac{|W^t S_B W|}{|W^t S_W W|}$,

where W is the transformation matrix, S_B is the between-class scatter matrix, and S_W is the within-class scatter matrix. Note that the within-class scatter matrix is

$S_W = \sum_{i=1}^{c} \sum_{x \in \mathcal{X}_i} (x - m_i)(x - m_i)^t$,

where m_i is the mean for the ith class $(i = 1, 2, \ldots, c)$, and the between-class scatter matrix is

$S_B = \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^t$,

where m is the mean for all of the data and n_i is the number of data in the ith class $(i = 1, 2, \ldots, c)$.

For our EM algorithm, the clustering may result in over-splitting due to the unavailability of the number of classes. A naive approach to find the discriminating features is to implement D-EM algorithm [5] directly on the data. However, since some different components during EM iteration may correspond to the same class due to over-splitting, the direct MDA on the data may not be optimal. For example, for the two components C_1 and C_2 that actually belong to the same class, the transformation W of the direct MDA still attempts to separate C_1 and C_2 instead of combining them. Furthermore, since MDA is a global optimization problem, the attempt to separate C_1 and C_2 weakens the discriminating ability of the resulting features for separating the other "real different" classes. Thus, the direct D-EM in [5] may mislead the exploration of the discriminating features in our system.

Some components in our semi-supervised EM algorithm correspond to the experience concepts in the collection Φ, and such correspondence is determined at the initialization stage of the EM since the center of each experience concept in Φ is set to be the initial center of one component in EM as we have introduced in Section 2.2.1. For each pair of components C_i and C_j $(i, j = 1, 2, \ldots, c)$ in EM, since we can compute the probability that their corresponding experience concepts in Φ are belonging to the same class [7], this probability can be regarded as the probability that C_i and C_j are belonging to the same class, and we denote it as $prob(C_i, C_j)$ $(i, j = 1, 2, \ldots, c)$. Note that such probability is also a kind of labelling information since it is derived from people, although it is not in the form of $\{0, 1\}$.

Now we reconsider the MDA with the probabilities $prob(C_i, C_j)$ $(i, j = 1, 2, \ldots, c)$. Intuitively, if the value of $prob(C_i, C_j)$ is low, the transformation of the MDA should tend to separate C_i and C_j and vice versa. To achieve this idea, we first rewrite the between-class scatter matrix in the pairwise form as

$S'_B = \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} n_i n_j (m_i - m_j)(m_i - m_j)^t$.

As each term on the right of the above equation represents the separation between a pair of components C_i and C_j $(i, j = 1, 2, \ldots, c)$, we assign the weight to this separation by directly multiplying the probability $(1 - prob(C_i, C_j))$ to the term corresponding to C_i and C_j, i.e., our criterion for the separation of different components is given by modifying the between-class scatter matrix S'_B as

$$S''_B = \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} (1 - prob(C_i, C_j)) n_i n_j (m_i - m_j)(m_i - m_j)^t.$$

Correspondingly, the objective function for MDA is modified as

$J'(W) = \frac{|W^t S''_B W|}{|W^t S_W W|}$.

We call this probabilistic MDA. The probabilistic MDA is more suitable to deal with the case where mixture components are related to each other with probabilities, instead of the binary relationship which MDA usually can handle.

2.2.3 EM framework: Figure 2 presents our EM algorithm for concept learning. Our probabilistic MDA is integrated with the E-step and M-step.

Since over-split clustering may lead to very small components, which cause the singularity problem, we remove those small components right after the component proportion estimation at each iteration. The removal criterion is: if $\alpha_j < \delta(\frac{1}{c})$, $j = 1, 2, \ldots, c$, the jth component is removed. This is called *component annihilation* [8]. Although component annihilation is usually based on the relationship between sample sizes and dimensionality, we simply use a constant parameter δ ($\delta = 0.1$ in this work) since the purpose of our component removal is only to avoid singularity.

2.2.4 Indexing and search: We directly use the clustering result of our EM algorithm for indexing, by which the images belonging to the same cluster are stored consecutively on hard drive. When a query images comes, the system chooses the cluster with the highest probability that the query belongs to this cluster based on the mixture model parameters. Thus, the search is limited to the images belonging to this cluster so that the search time is saved compared with the global search.

```
• Given the data $\mathcal{X}$, the collection of the experience concept $\Phi$, and $c_{max}$.
• Initialization (see Section 2.2.1).
• Estimate component proportions $\{\alpha_1, \alpha_2, \ldots, \alpha_c\}$.
• Repeat
    1. Remove the $j$th component if $\alpha_j < \delta\frac{1}{c}$, $j = 1, 2, \ldots, c$. Normalize the proportions for the remaining $c'$ components, $c \leftarrow c'$.
    2. E-step: Estimate component-indicators $\mathcal{Z}$ .
    3. Modify $\mathcal{Z}$ by concept collection $\Phi$.
    4. $\mathcal{X} \leftarrow$ probabilistic MDA($\mathcal{X}, \mathcal{Z}, \Phi$).
    5. M-step: Compute component proportions, means and covariances respectively ([4]).
    6. Stop if termination criterion is met.
```

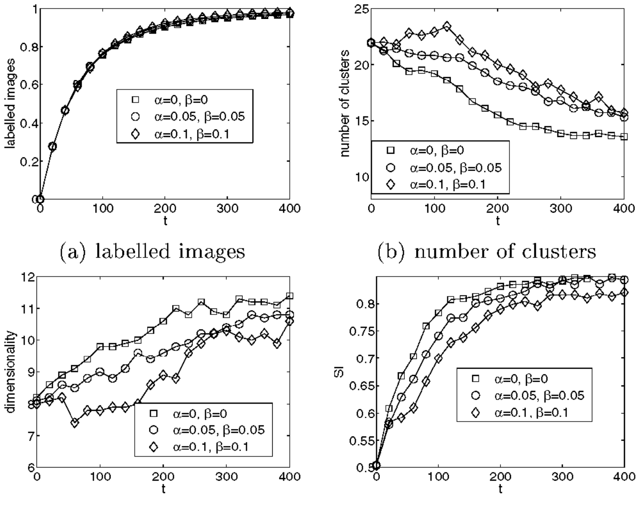
Figure 2: Probabilistic D-EM algorithm for concept learning.

3 Experiments

We collect 1,200 images from Corel stock photo library and divide them into 12 classes [3]. Images are represented by texture and color features only. The texture features are derived from 16 Gabor filters. We also extract means and standard deviations from the three channels in HSV color space. Thus, each image is represented by 22 features.

We set the system running time as $t = 0, 1, 2, \ldots$; at each t, a user makes his/her queries and provides a retrieval experience by executing relevance feedback. At each relevance feedback iteration, the system presents 20 images for users to label. We assume that the system only knows that $c_{max} = 2c_{real} = 24$.

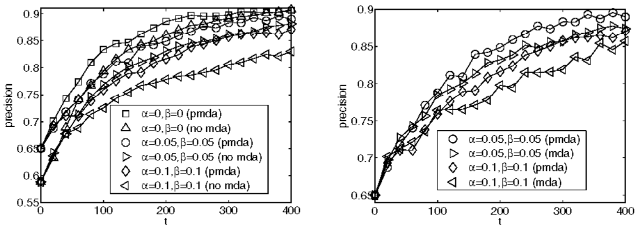
Figure 3 provides various performances for concept learning with different labelling noise rates. Figure 3(a) shows that the percentage of the images ever being labelled (no matter positive or negative) increases with retrieval experiences increased. From Figure 3(b), we observe the number of clusters (after EM algorithm) over time. For lower labelling noise rate, this estimation converges to c_{real} (=12) faster. Figure 3(c) shows that integrating probabilistic MDA with EM effectively reduce the dimensionality of feature space (original dimensionality is 22). Since the main computational load of EM is the computation of the inverse of each component's covariance, whose complexity is $o(cd^3)$ (c is the number of components and d is feature dimensionality). In this experiment, the dimensionality is reduced from 22 to 7 $\sim$ 12. Thus, the computational load for the EM is alleviated significantly. Figure 3(d) validates that the clustering (evaluated by the statistical index



(a) labelled images (b) number of clusters

(c) dimensionality (d) statistical index

Figure 3: Performance for concept learning with various labelling noise rates.



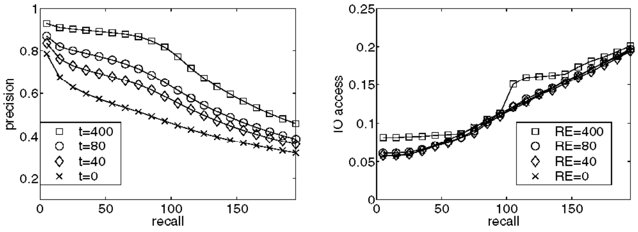
(a) pmda vs. no mda (b) pmda vs. mda

Figure 4: Retrieval precisions of different approaches.

SI) is improved with more labelling information provided by retrieval experiences.

Figure 4 demonstrates that the retrieval precision improves faster with lower labelling noise rates α and β. Furthermore, Figure 4(a) compares our probabilistic MDA (pmda) approach with non-MDA one, and we observe that pmda is better than non-MDA approach in all of the cases with various labelling noise rates. Figure 4(b) shows that the integration of pmda with EM is better than the direct D-EM in [5].

Figure 5 presents two important results without labelling noise. Since the results for the cases with labelling noise have the similar trends (although their learning improvement is slower), we do not show them due to space limitation. Figure 5(a) shows the precision-recall curves from which we observe that the retrieval performance is improved with more retrieval experiences. Using the clustering result, the indexing structure can be derived. Figure 5(b) helps us to conclude that the indexing structure requires much less search (measured by IO accesses since the database images are stored on hard drive) compared with the global search.



(a) recall (b) IO access

Figure 5: Noise rate = 0.0.

4 Conclusions

The paper proposes a concept learning approach for image databases, which is achieved by an EM algorithm integrating the probabilistic MDA, so that the difficulties brought by the high dimensionality of feature space are alleviated and the discriminating features are determined. The retrieval performance is improved using the concept learning results in two aspects: 1) the retrieval precision is improved, and (2) the search time (represented by IO access) is saved.

Acknowledgements

This research was supported in part by grants IIS-0114036 and F49620-02-1-0315. The contents of the information do not necessarily reflect the position or the policy of the US Government.

References

[1] N. Vasconcelos, *Bayesian Models for Visual Information Retrieval*, Ph.D thesis, MIT, 2000.

[2] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," *Proc. IEEE Int. Conference on Computer Vision*, vol. 2, pp. 408–415, 2001.

[3] A. Dong and B. Bhanu, "Active concept learning for image retrieval in dynamic databases," *Proc. ICCV*, pp. 90–95, 2003.

[4] G. McLachlan and D. Peel, *Finite Mixture Models*, John Wiley & Sons, 2000.

[5] Y. Wu and T. S. Huang, "Towards self-exploring discriminating features for visual learning," *Engineering Applications of Artifical Intelligence*, vol. 15, pp. 139–150, April 2002.

[6] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, John Wiley & Sons, second edition, 2001.

[7] A. Dong and B. Bhanu, "On labeling noise and outliers for robust concept learning for image databases," *IEEE Workshop on Learning in Computer Vision and Pattern Recognition*, 2004.

[8] M. Figueiredo and A.K. Jain, "Unsupervised learning of finite mixture models," *IEEE PAMI*, vol. 24, no. 3, pp. 381–396, March 2002.