Proceedings of International Conference on Pattern Recognition (ICPR), Hong Kong, August 20-24, 2006

Adaptive Discriminant Projection for Content-based Image Retrieval

Jie Yu
Department of Computer Science
University of Texas at San Antonio
jyu@cs.utsa.edu

Qi Tian
Department of Computer Science
University of Texas at San Antonio
qitian@cs.utsa.edu

Abstract

Content-based Image Retrieval (CBIR) is a computer vision application that aims at automatically retrieving images based on their visual content. Linear Discriminat Analysis and its variants have been widely used in CBIR applications because of their effectiveness in finding a projection that maps the original high-dimensional space to a low-dimensional one and preserves the most discriminant features. Those techniques assume images from certain class(es) are all visually similar and try to cluster them in the projected space. In this paper we show that the human high-level concept of semantic similarity between images may not arise only from the low-level visual similarity and consequently that assumption is inappropriate in many cases. We propose an Adaptive Discrimant Projection framework which could model different data distributions based on the clustering of different classes. To learn the best model fitting the real scenario, Boosted Adaptive Discriminant Projection is further proposed. Extensive experiments are designed to evaluate our methods and compare them to the state-of-the-art techniques on benchmark data set and real image retrieval applications. The results show the superior performance of our proposed methods.

1. Introduction

Content-based image retrieval is a computer vision application that automatically retrieves images of user interest from large image databases based on the visual content. The mapping between high-level semantic concept and low-level image features is obtained by a learning process. The images could be pre-processed to extract statistical features, such as color, texture and shape. An image feature vector is often used to represent an image as data point in a high-dimensional space. Although Content-based image retrieval has been successfully applied in many fields, it still faces two major challenges.

Small Sample Set: In CBIR, a set of samples with categorical information are used to train a classifier. Because labeling the training samples requires human interference and could be computational expensive, the size of the training set is often very small. In that case the learning process tends to bias to the training set and overfitting could occur.

High Dimensionality: In many data analysis application, the observed data have very high dimensionality. Specifically the images in CBIR are represented by image feature vector whose dimensionality ranges from tens to hundreds in most cases. Traditional statistical approaches have difficulties in modeling data directly in such a high dimensional space.

Some techniques have been proposed to alleviate the two problems. For the small sample set problem, researchers have been using unlabeled data along with the labeled data to avoid overfitting. However our research in [1] shows that the unlabeled data and labeled data must be from the same statistical source to improve the performance. Otherwise using unlabelled data may deteriorate the classification. For the high dimensionality problem, it is almost a common practice to conduct dimension reduction to find a compact representation of data in a low dimensional space. Traditional techniques, such as Principal Component Analysis (PCA) [2] and Linear Discriminat Analysis (LDA) [3], are widely used. For a classification task such as CBIR, LDA is often preferred because it incorporates class information and discovers the most discriminant features in the projected space.

2. LDA and BDA

2.1. Linear discriminat analysis

LDA tries to find a mapping from originally high-dimensional space to a low-dimensional space in which the most discriminant features are preserved. Intuitively it makes samples from same class cluster to each other and samples from different classes separate from each other. Mathematically it could be modeled as finding an optimal projection that maximizes the following ratio:

$$W_{LDA} = \arg\max_{w} \frac{|WS_B W^T|}{|WS_w W^T|} \quad (1)$$

, where $S_B = \sum_{i=1}^{C} P_i (m_i - m_G)(m_i - m_G)^T$ is between-class scatter matrix and $S_W = \sum_{i=1}^{C} \sum_{x_j \in \text{class } i} (x_j - m_i)(x_j - m_i)^T$ is within-class scatter matrix. C is the number of classes. m_i is the class mean of i_{th} class and m_G is the grand mean of all samples. P_i is the priori probability of i_{th} class. The maximization problem has a closed-form solution: W is the eigen vectors of $S_B S_w^{-1}$ corresponding the largest non-zero eigen values [3]. It can be denoted as follows:

$$W_{LDA} = \underset{\max}{eig}(S_B S_w^{-1}) \quad (2)$$

If the number of classes C is greater than 2, we obtain a Multiple Discriminant Analysis (MDA). Otherwise the two-class discriminant analysis is obtained and often known as Fisher Discriminat Analysis (FDA). In most CBIR applications, the users only label the images as relevant to their interest (positive) or irrelevant to their interest (negative). The reason lies in two folds: 1) the precise subjects of images within positive and negative classes are still too complicated and ambiguous to define and 2) the user may group images with dramatically different visual contents into the same semantic class. For instance one who is interested in images of fruit may result in labeling strawberry and banana as positive in spite of the fact that they are visually quite dissimilar.

Although LDA is one of the most widely used techniques in CBIR, it still faces some major problems:

Effective Dimension: The number of non-zero eigen values in $S_B S_W^{-1}$ determines the maximal dimensionality of the projected space

for LDA, which is known as *effective dimension*. Because the rank of S_B is less than or equals to *C-1*, the effective dimension could be at most *C-1*. As we discussed before, for most CBIR applications *C=2*. Thus the projected space can only have dimensionality of 1, which prevents accurate modeling of the data in higher dimensional projected space.

Regularization: Before conducting eigen operation on $S_B S_W^{-1}$, we have to make sure the S_W is a full rank matrix to calculate its inverse. When the small sample set and/or high dimensionality problem occurs, S_W is often singular and has no inverse. A common practice to handle this problem is to use regularization by adding small quantities to the diagonal elements of S_w and force it to be full rank [5]. However the soundness of the regularization technique hasn't been theoretically justified.

2.2. Biased discriminant analysis

In two-class LDA, the equivalent effort has been taken to cluster negative and positive samples. Intuition suggests that clustering the negative samples may be difficult and unnecessary because they may be from visually different classes. Zhou and Huang propose a Biased Discriminant Analysis (BDA) which clusters only positive samples and makes the negative samples far away from the positive ones [3]. There is no effort to cluster negative samples. The assumption behind the BDA is that the samples from the positive class are visually similar and should be clustered in the projected space. On the other hand the negative samples might be from different classes and it is difficult to find a mapping to make them close to each other. For classification tasks we just need to make the negative samples far away from the center of positive ones.

Although the idea of BDA is easy to accept, we found that its assumption is inappropriate in some scenarios which will be explained in Section 3. The complex nature of human concept requires a classification method that can adaptively fit the distribution of images from different classes.

3. Boosted adaptive discriminant projection

As we discussed in Section 2, LDA assumes images from positive and negative classes are from the same sources respectively and they could be clustered in the projected space. BDA assumes that positive samples must be visually similar and negative samples may be from different sources. However one can easily find many CBIR applications that don't fit into either assumption.

3.1. Adaptive discriminant projection

To provide a more accurate model of the complex distribution for positive and negative images, we propose an Adaptive Discriminant Projection (ADP) framework:

$$W_{ADP} = \arg\max_W \frac{|W[\lambda S_{P->N} + (1-\lambda) S_{N->P}]W^T|}{|W[\eta S_P + (1-\eta) S_N]W^T|} \quad (3)$$

in which

$$S_{N->P} = \sum_{i \in Negative} (x_i - m_P)(x_i - m_P)^T \quad (4)$$

$$S_{P->N} = \sum_{j \in Positive} (x_j - m_N)(x_j - m_N)^T \quad (5)$$

$$S_P = \sum_{j \in Positive} (x_j - m_P)(x_j - m_P)^T \quad (6)$$

$$S_N = \sum_{j \in Ne} (x_i - m_N)(x_i - m_N)^T \quad (7)$$

The m_P and m_N are the means of positive and negative samples, respectively. The two parameters λ and η controls the bias between positive and negative samples. Proper setting of parameters may fit the real distribution of data. The following table gives examples of the special cases for ADP:

Table 1. Special cases of ADP

(λ,η)	Optimal Projection	Note
(0,0)	$W_{ADP} = \arg\max_W \frac{\|WS_{N->P}W^T\|}{\|WS_N W^T\|}$	Case 1
(0,1)	$W_{ADP} = \arg\max_W \frac{\|WS_{N->P}W^T\|}{\|WS_P W^T\|}$	Case 2 (BDA)
(1,0)	$W_{ADP} = \arg\max_W \frac{\|WS_{P->N}W^T\|}{\|WS_N W^T\|}$	Case 3 (Counter -BDA)
(1,1)	$W_{ADP} = \arg\max_W \frac{\|WS_{P->N}W^T\|}{\|WS_P W^T\|}$	Case 4
(0.5,0.5)	$W_{ADP} = \arg\max_W \frac{\|W(S_{P->N} + S_{N->P})W^T\|}{\|W(S_P + S_N)W^T\|}$	Case 5 (LDA-like)

From the above table we can find that the ADP recovers BDA when λ and η are set to 0 and 1, respectively in Case 2. Case 5 corresponds to a LDA-like projection which assumes the positive and negative sample are both from single sources, respectively. Case 3 finds a projection that is on the contrary side of BDA while Case 1 and 4 show some distribution scenarios that haven't been discussed in literature before. We find that all 5 cases fit certain distributions and have correspondence with some CBIR query scenarios as illustrated in Figure 1.

Case 1 may handle the distribution that the size of positive sample set is much larger than that of negative samples and the negative samples may be from different smaller classes (*Fig. 1 (a)*). Normal LDA or BDA would severely bias to cluster the positive samples (S_P) since they dominate the training set. A real CBIR query scenario that fits this case would be quality control based on the scanned image of internal structure of material. In that application most images are scanned from qualified products (positive) and only few of them reflects defectiveness in unqualified products (negative).

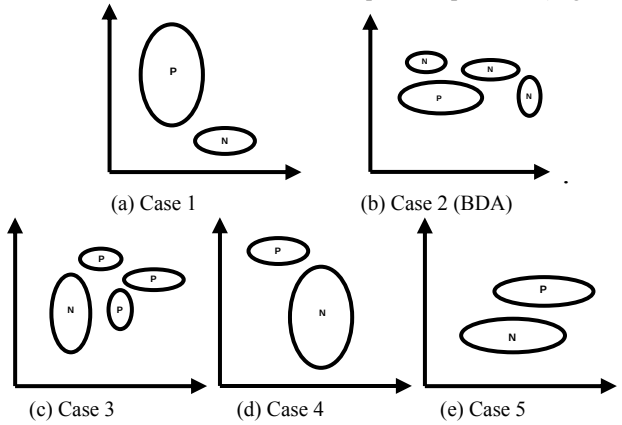

Figure 1. Distributions and labels corresponding to special cases (*P*: positive samples, *N*: negative samples)

Similarly Case 3 may correspond to applications where the size of negative sample set is much larger than that of positive samples and the positive samples may from different sub-class clusters. One possible application with this distribution is medical imaging, radiation or MRI, for diagnosis. In that case most images are from normal person (negative) and one may only interest in the images that are from patients (positive).

Case 2 may best fit distribution illustrated in *Figure 1 (b)* where positive images all look alike while negative ones may be irrelevant to each other and from different distributions. BDA finds the optimal projection that fits that situation. In practice one may find that case fits the face identification application where one person's facial images need to be classified from among facial images of 4 persons.

Case 4 is on the opposite side of Case 2, in which negative samples share strong correlation while positive samples may be quite different. An example of this kind of application will be telling images of fruit, e.g. apple, orange and banana, from those of green vegetables.

The above discussion shows that our ADP framework could model more distributions than LDA and BDA. The case 1 and case 4 could handle the imbalanced size of positive and negative training set while case 2 and 3 may correspond to the scenario that one semantic class contain multiple subclasses. Although we give example applications that fit the extreme cases of ADP, more accurate fitting could be achieved by parameter tuning.

3.2. Boosting adaptive discriminant projection

To find the best parameter setting for a particular application, one may compare the performance of the ADP projection corresponding to different settings.

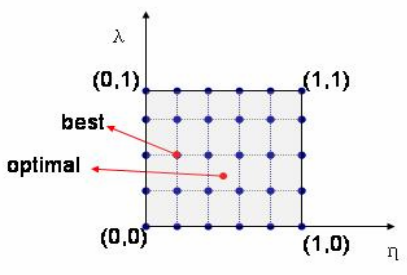

Figure 2. Best ADP found in the parameter space by subsampling may not be the one with optimal setting.

However to find an optimal setting one has to do exhaustive searching in the square region in Figure 2, which is computational expensive. Besides, the parameter setting one find to perform the best on training set can't guarantee best performance on real data set as illustrated in Figure 2. To solve that problem, we adopt the idea of AdaBoosting [6] to generalize a set of classifiers that are trained on different data sets projected by ADP with different parameter settings into a more accurate one.

The basic idea of AdaBoosting lies in the following two folds: 1) the boosting process trains each classifier iteratively with weighted training samples. The misclassified training samples receive more weights in the next run. Thus the classifier is forced to pay more attention to those difficult to learn samples; 2) the final classifier is the weighted combination of a set of weak classifiers. The weight for each classifier is set according to its performance, that is, the better the performance the larger the weight.

4. Experiments and analysis

4.1 ADP vs LDA

Our first experiment is designed to evaluate the effectiveness of the proposed ADP and its boosted version. The methods are tested on benchmark data sets from UCI repository. For comparison purpose, LDA and BDA is also implemented and tested. Due to limited space Figure 3 only shows the results on Heart data set. Similar results are obtained on other data sets. In all the experiments we conducted, our boosted ADP is trained on 36 ADP classifiers with (λ, η) evenly sampled from 0 to 1 with step size of 0.2. In all the experiments, Bayesian classifier is used on the projected data.

In Figure 3 as iteration goes on, the error rate decreases for best single ADP classifier and the boosted ADP. The performance of LDA and BDA are shown for reference as straight line. Although Boosted ADP starts with a set of weak classifiers (compared to the best ADP classifier), but after one iteration, the boosted ADP outperforms the single best ADP classifier. Both the ADP and its boosted version outperform the LDA and BDA in this experiment. Although the improvement seems not significant, it could be because the benchmark data sets usually contain clean and sufficient data for training.

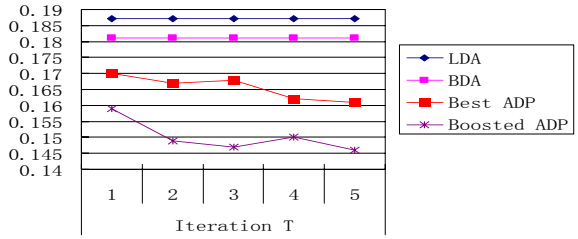

Figure 3. AdaBoost on Heart benchmark data set

4.2. Comparison to state-of-the-art

In the second experiment, we test the performance of our proposed methods when handling the small sample set problem. The state-of-the-art linear and nonlinear variants of discriminant analysis including DEM, kernel DEM (KDEM) [7], BDA, kernel BDA (KBDA) [4] are also tested as comparison to our methods. The boosted ADP used in this experiment are trained on ADP classifiers with parameters evenly sampled from 0 to 1 with step size of 0.05.

The data sets used in the experiments are the MIT facial image dataset [8] and non-face images from Corel database. All the face and non-face images are scaled down to 16×16 gray images and normalized feature vector of dimension 256 is used to represent each image. The size of the training set is 100, 200, 400, and 800, respectively. Compared with the feature vector dimension of 256, the training sample size is set from relatively small to relatively large. Table 2 gives the experiment results with smallest error rate in bold.

From the results in Table 2 we find that our proposed methods perform well when the training set size is small compared to the feature dimensionality. When compared with linear techniques of DEM and BDA with simple regularization, the ADP performs much better than them and doesn't require regularization. Even when compared with the KDEM and KBDA, the boosted ADP performs better in 3 out of 4 tests. It should be noted only linear transformation is used in our ADP, but it is more efficient than nonlinear algorithms

such as KDEM and KBDA. All these show the robust performance of the ADP.

Table 2. Comparison to DEM, BDA, KDEM and KBDA

Error Rate (%)	Size of Training Set			
	100	200	400	800
DEM w/ reg.	10.5	19.3	15.0	9.0
BDA w/ reg.	34.7	25.4	18.5	19.3
KDEM	6.93	1.93	1.7	**0.5**
KBDA	3.04	2.89	2.58	1.44
ADP (λ^*,η^*)	2.5 (0.35,0.1)	1.9 (0.55,0.2)	1.7 (0.1,0.15)	1.6 (0.1,0.1)
Boosted ADP	**1.85**	**1.63**	**1.41**	0.84

We also compare the performance of our proposed methods with the state-of-the-arts on Harvard facial image database with different illumination conditions. The experiment settings are the same as in [9]. The following are examples of the images:

Figure 4. Facial images with different illumination condition

From the experiment result in the Table 3, we can find the proposed projection could find the most discriminant features given the change in illumination condition and perform better than the other techniques.

Table 3. Comparison of PCA, LDA and its variants with ADP and Boosted ADP

Method	Error Rate (%)		
	Subset 1	Subset 2	Subset 3
PCA w/o 1st 3 features	1.2	5.4	25.3
LDA	0.7	1.4	3.7
BDA	1.5	3.2	7.5
KDEM	0.6	1.1	2.9
KBDA	1.3	1.9	3.4
ADP (λ^*,η^*)	0.42 (0.35,0.1)	0.67 (0.25,0.15)	2.01 (0.3,0.2)
Boosted ADP	0.33	0.52	1.84

4.3. Effective dimension

In the last experiment, we test the ADP versus the projection dimension using the same MIT face databases and COREL databases. Since it is a two-class classification problem (face vs. non-face) in traditional LDA, the effective projection dimension is one. The feature dimension and experimental setting are same as before except that the size of the training dataset is fixed at 100.

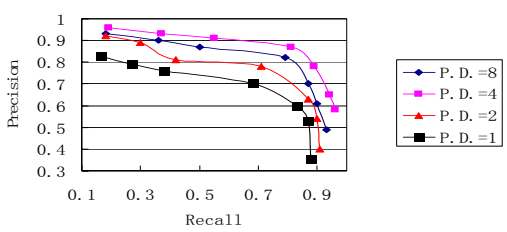

Figure 5. Precision-Recall graph for different Project Dimension (P.D.)

Figure 5 shows the Precision-Recall graph for different projection dimensions. Clearly, the projection dimension of 4 gives the best performance and all the higher projection dimensions yield smaller error rate than that of *C-1 (C=2)*. This shows that the ADP increases the effective dimension and as a result the classification performance improves since the data structure can be more accurately modeled in a space with dimensionality higher than *C-1* .

5. Conclusions and future work

In this paper, we propose a novel Adaptive Discriminat Projection to better model the distribution of image data from different classes. It provides a much richer set of alternatives to LDA and BDA. As a result, it not only compensates for regularization that is afflicted by all sample-based estimation methods, but also increases the effective dimension of the projected subspace. In order to avoid parameter searching, the boosted ADP is also proposed. We found it can provide the two desirable properties. First, the boosted ADP can provide a unified and stable solution to finding optimal projection. Second, the weighted training schemes in boosting add indirect non-linearity and adaptivity to the linear methods and thus enhance it by iterations. The experimental tests on benchmark databases and image retrieval applications have shown the superior performance of ADP and boosted ADP.

In the future, we are interested in continuing this research work in the following two directions: 1) using fusion methods to combine ADP classifiers and 2) exploring learning based approaches to find optimal parameter settings for ADP.

Acknowledgement: This work was supported in part by the Army Research Office (ARO) grant under W911NF-05-1-0404, and by the Center of Infrastructure Assurance and Security (CIAS), the University of Texas at San Antonio.

6. References

[1] Q. Tian, J. Yu, Q. Xue, N. Sebe, "A New Analysis of the Value of Unlabeled Data in Semi-Supervised Learning for Image Retrieval," *International Conference on Multimedia and Expo (ICME 2004)*, Taipei, Taiwan, ROC, June 27-30, 2004.

[2] I. T. Jolliffe, *Principal Component Analysis*. 2nd edition, New-York: Springer-Verlag2, 002.

[3] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd edition, John Wiley & Sons, Inc., 2001.

[4] X. Zhou, and T.S. Huang, "Small sample learning during multimedia retrieval using biasMap," *Proc. of IEEE Conf. CVPR*, Hawaii, December 2001.

[5] J. Friedman, "Regularized discriminant analysis," *Journal of American Statistical Association*, vol. 84, 1989.

[6] Y. Freund, and R. E. Schapire, "A short introduction to boosting," *Journal of Japanese Society for Artificial Intelligence,* , Sep., 1999.

[7] Q. Tian, Y. Wu, J. Yu, and T.S. Huang, "Self-Supervised Learning Based on Discriminative Nonlinear Features for Image Classification," *Pattern Recognition, Special Issue on Image Understanding for Digital Photographs,* Vol. 38, No. 6, pp. 903-917, 2005.

[8] H. A. Rowley and S. Baluja and T. Kanade, "Neural Network-Based Face Detection," *IEEE Trans. PAMI,* Vol. 20, 1998.

[9] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE TPAMI*, vol. 19, pp. 711-720, 1997.