Multi-class object recognition using boosted linear discriminant analysis combined with masking covariance matrix method

Masashi Tanigawa and Takio Kurita
National Institute of Advanced Industrial Science and Technology,
Umezono 1-1-1, Tsukuba JAPAN

Abstract

We propose a new algorithm, boosted linear discriminant analysis (bLDA), for classification of a non-linear pattern distribution, and masking covariance matrix method (MCM) for robust and fast computation of object recognition. bLDA integrates classifiers on multiple linear discriminant spaces. Each linear discriminant space is spanned by eigenvectors so as to maximize ratio of within-class variance and between-class variance of training data. The weights of samples are updated for each boosting step; increasing weights for misclassification while decreasing weights for correct classification. bLDA performs well to classify a complicated data distribution, such as face images. In addition, we propose MCM to find optimal local features, instead of traditional exhaustive search in huge number of candidates feature, for robust and real-time object recognition. In MCM, the covariance vectors of training data set are restricted to be locally correlated, by multiplication of covariance mask. bLDA combined with MCM automatically and effectively extract spatially-local features. Especially, the covariance mask on Haar space induces anisotropic collinearity of object's contours. bLDA-MCM algorithm performed 98.70% correct for face/non-face classification task, after 100 rounds of boosting.

1 Introduction

We propose a new method for improving a traditional linear discriminant analysis (LDA), which combines multiple classifiers built on a linear discriminant space (LDS).

LDS are spanned by eigenvectors which are derived from maximizing variance ratio of within-class and between-class of a given data set. LDA has been successfully applied for many applications [1]. But actual distributions in real world, such as face images, are complicated so that LDA fails to reach to enough performance of classification. To improve classification performance for such complicated distributions, several approaches have been proposed in decades; neural networks, kernel methods[4], clustering techniques and so on. These nonlinear classifiers can successfully adapt to a complicated structure of data distribution appeared in real-world applications. We use boosting method to integrate classifiers to fit a given distribution. The boosting combines weak classifiers which do not have sufficient ability to segregate the target from the background by itself. The boosting methods are known that total classifiers has a good generalization for a set of samples. In this paper, we propose boosted linear discriminant analysis (bLDA), which combines weak-classifiers underlying on linear discriminant space. The LDS is basically constructed as traditional manner, except for weighting samples. The weights of samples are assigned by boosting procedure, weighting misclassified samples in previous rounds of boosting. In the AdaBoost algorithm, the weights of samples decrease when the classifier responds correctly; on the other hand, those of samples increase when the classifier responds faultily. Sequential weight updating process determines which samples should be learned intensively in the next round of boosting. Firstly, we will verify well-classification performance of bLDA for simple statistical data Secondly, we propose masking covariance matrix method (MCM) for selecting optimal local features. The MCM is just multiplication of covariance matrix of training data set. The image object are comprised of a spatially local parts , for example, face images can decomposed into some primitives, nose, mouth, and so on. To design these combination of local features, we exhaust a number of times to find the best. In this paper, we assume the local features correlates. The assumption is implemented in covariance matrix constraints of LDA calculation.

0-7695-2506-7/06 $20.00 © 2006 IEEE

2 bLDA: composing classifiers on linear discriminant spaces.

2.1 Traditional linear discriminant space

The linear discriminant space (LDS) is constructed on eigenvectors which are column-vectors of transformation matrix such that the ratio between within-variance Σ_W and between-variance Σ_B is maximized to distinguish class distributions. The transformation matrix U is linear transformation from feature space x to linear discriminant space y: $y = U^T x$. The matrix U are derived from the discriminant criterion (1):

$$\hat{U} = \underset{U}{\operatorname{argmax}} \frac{U \Sigma_B U^T}{U \Sigma_W U^T} \tag{1}$$

The discriminant criterion shown above are equivalent to a generalized eigenvalue problem:

$$\Sigma_B U = \Sigma_W U \Lambda \tag{2}$$

where Λ is generalized eigenvalues. In the image analysis in section 3, we reduce data's dimension using PCA before LDA, because of computation cost. PCA was applied in section 3, 90% contribution subspace are used.

2.2 Design of classifier on LDS

We now introduce a new type of classifier similar to naive-bayes method. In this method, eigenvectors on LDS are used as feature vectors u_i^T, and the marginal probability on training data are referred as a output function $f(y)$: feature's output is $z = f(u_i^T x)$. The similar output function, using marginal distribution, are utilized in face-recognition system[14].

Here, let $p(y_1, \cdots, y_n|k)$ to be data distribution of class k, where y_i is a position on the linear discriminant coordinate system, and n is dimension number of LDS, usually, $n = \min(\text{rank}(\Sigma_B), \text{\# of data sample-1})$. The aim of designing classifier is to estimate a posterior distribution $p(k|y_1, \cdots, y_n)$ of a given class k. Using Bayes' law, the posterior distribution is expressed as follows:

$$P(k|y_1, \cdots, y_n) = \frac{1}{Z} p(y_1, \cdots, y_n|k) P(k) \tag{3}$$

where $p(k)$ is a prior distribution, which is estimated by weight sum in class k in the training data set. Because $p(y_1, \cdots, y_n|k)$ is high dimensional function, we assume that it can decompose into marginal distribution of each LDS axis:

$$p(y_1, \cdots, y_n|k) = \prod_{i=1}^{m} p(y_i|k) \tag{4}$$

This assumption is a bit rough, and it would cause large deviation from true probability. But, as described later, the boosting method compensates and reduces the deviation by another classifiers. Substituting (4) into (3), the posterior probability forms:

$$P(k|y_1, \cdots, y_n) = \frac{P(k)}{Z} \prod_{i=1}^{n} p(y_i|k) \tag{5}$$

Simply, in this paper, $p(y_i|k)$ is parametrized with Gaussian distribution $p(y_i|k) \approx \mathcal{N}(\mu_i, \sigma_i)$. Alternatively $p(y_i|k)$ can be represented non-parametrically, e.g. by histogram.

2.3 Boosting

Figure (1) shows the procedure of AdaBoost algorithm. The weight of each sample is scalar, and error function for updating weight procedure is as follows.

$$\begin{aligned} \text{err} &= \sum_{k \neq i} P(i|y) - P(k|y) && (6) \\ &= 1 - 2P(k|y) && (7) \end{aligned}$$

where k is the label number of a current data y. In two class discrimination, these criterion are reduced to be the error function for the original Adaboost.

2.4 Experiment: IRIS

We demonstrate three-class discrimination task using Fisher's iris data which are consisted of 150 samples of four-dimensional features.

Figure 2(A) shows data distribution for each iris, projected on two dimensional subspace spanned by dominant eigenvectors of LDS. One of three iris is easily discriminated from two others. The marginal distributions for each class are shown in Figure 2(B). The AdaBoost reduces the importance weights of samples that are correctly responded by current classifier. Figure 2(B) shows the weight distribution, The color saturation of dots are depicted to the strength of the weight; stronger saturation means larger weight. The overlap region of the two class tend to increase their weights since it is difficult to separate. After four rounds of boosting, the weight strength of the data are plotted in right side of Figure 2. Successive LDS are constructed by underlying these weights of samples.

We measure bLDA using the ten-fold cross-validation method; iris data are randomly divided into ten subsets. One of them is used for measuring correct ratio, and the other subsets are used for training. The ten combinations of the subsets are examined. The ten-fold cross-validation experiments showed 95.63% correct of performance.

```
1. Let $N$ be the number of samples, $M$ be the number of boosting steps,
2. Initialize with weights $w_i = 1/N$,$i = 1, 2, \dots, N$.
3. For $m = 1$, to $M$:
   (a) Spanning subspace so as to maximizing the ratio of variances $\Sigma_B$ and $\Sigma_W$.
   (b) Select axises of LDA as feature vectors. eq.(2)
   (c) Project data distribution to the axises and approximate marginal probability density,
       for example Gaussian distribution ($\mu_i$, $\sigma_i$), and histogram and so on.
   (d) Estimate confidence of classifiers eq.(5).
   (e) Estimate error; err is equation eq.(7).
   (f) Update data weight $w_i \leftarrow w_i \exp[err]$, $i = 1, 2, \dots, N.$, and normalize the
       weights such that $\sum_i w_i = 1$.
4. Output the total classifier's confidence value: $p(k|y)$ by eq.(5). and total classifier's re-
   sults: $\underset{k}{\operatorname{argmax}}(p(x_k))$.
```

Figure 1. bLDA algorithm: The classifiers are built on linear discriminant space, errors are estimated by one-other comparison method; It defines the correct only if the largest output is true label.

3 Local feature selection

3.1 The constraint of local correlation

In natural scene, it often encounters that target objects are occluded, shaded and deformed partially. To develop object detector to be robust against various environment, number of methods have been proposed. One of the strategies is to select local features which cover a given small restricted area, and combines them globally. The important points are how to select local feature efficiently among a lot of candidates. For example, Viola and Jones developed object detectors using rectangle features which are simple edge detectors similar to Haar wavelet [13]. In their method, they select optimal features, maximizing separability between target and background, among a huge number of features pool, filled with possible variation; position, size and shape, over ten thousands of features. Finding optimal features are usually exhaustive search [8]

In addition, fast rise of boosting strongly depends on which feature we should select at early epoch of detection. Consider that we have sufficient amount of learners that are *too-weak*-learners, the combination of them somewhat improves total performance, but the rise of boosting is not steep. Original Viola's rectangle features are too weak to classify for complicated classification in natural environment, such as multiple detection of faces, cars, building, etc. because these simplest features are likely to be member of every objects. Therefore, we should select bit strong features that is sufficient strong alone.

To find appropriate features, we used bLDA as described above. We modify bLDA method to be easy to select optimal local feature, using masking covariance matrix method (MCM). The MCM multiply covariance matrix with covariance mask which constraints covariance matrix with a priori spatial correlation. We can estimate optimal covariance matrix, if we have a complete set of partially occluded image ideally. But we can not always prepare complete variations of partially occluded images. To obtain the semi-optimal covariance matrix, we adjust the empirical covariance matrix with a prior information that features are only locally correlated. The constraint covariance matrix is

$$\tilde{\Sigma}_{ij} = G_{ij}\Sigma_{ij} \tag{8}$$

where G_{ij} indicates local correlation between i-th pixel and j-th pixel. The local correlation is calculated by the following Gaussian spread function:

$$G_{ij} = \exp(||x_i - x_j||^2/2\sigma^2) \tag{9}$$

where x_i is a position of i-th pixel and σ_{MCM} is spatial extension of local correlation. Replacing Σ to $\tilde{\Sigma}$ in bLDA procedure provides automatic and efficient extraction of local features.

Figure 3 shows various local features that are extracted in each round of boosting. Figure 4 shows the results for varying σ_{MCM} for the face database was CMU-MIT database

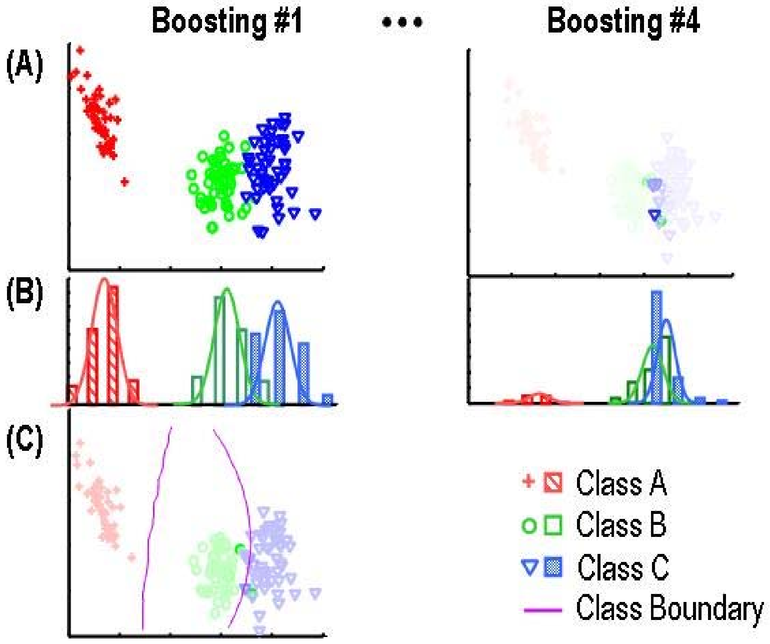

Figure 2. Example of bLDA process; The database *iris* is composed of four-dimensional feature vector (Petal width, petal length sepal width sepal length) of three types of iris plant. (A) iris distribution on two dimensional planes which are spanned by dominant eigenvectors of LDA for each round of boosting. Points are colored for each iris type; and the darkness of point indicates the strength of data weight. (B) The marginal distribution are projected to a primal axis. The marginal distribution are approximated by Gaussian distribution. The simple histogram are drawn by bars, and approximated Gaussian distribution are drawn by lines. (C) The updated weights are plotted. The boundary of classes are drawn with curves.

consisted of 2430 faces and 9137 non-face images for test samples, and 472 faces and 23573 non-face images for training samples. The smaller σ_{MCM} improves the correct ratio. The CRR dropped at 4 pixel. The unpluged MCM corresponds to σ_∞ The best performance was found at 1.0 pixel.

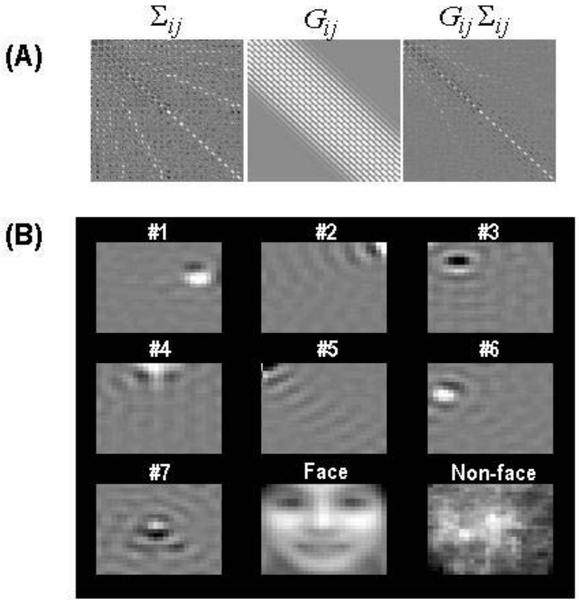

Figure 3. Local feature selection: (A) Original correlation matrix (left), Mask of local correlation G (center), Correlation matrix for LDA (right). (B) Extracted features (#1-7), averaged face and non-face. Various features are extracted, #1:left-eye, #2:left-hair, #3:right-eye, #4:forehead, #5:right-hair, #6:right cheek, #7:nose and mouth.

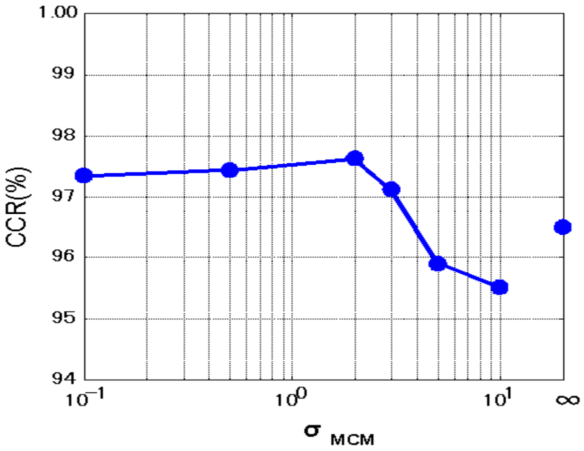

Figure 4. spatial width of covariance mask σ_{MCM}.

3.2 Multi-resolution local selection

As described above, effective local features for detection are automatically selected with respect to the pattern of the covariance mask. We extend covariance matrix constraints in multi-resolution domain. In this report, we use two-dimensional Haar wavelet. The LDA space is independent to selection of coordinate system. Changing coordinate system $y = Qx$ alters the transformation matrix U to QUQ^T. When we use different coordinate system. the covariance mask differently affects local feature extraction. The transformation of covariance mask is HGH^T, where H denotes matrix representation of Haar transformation. The covariance matrix $\tilde{\Sigma}$ for bLDA is obtained by

$$
\begin{aligned}
\tilde{\Sigma}_{ij}^{Haar} &= G_{ij}^{Haar}\Sigma_{ij}^{Haar} &(10)\\
&= (HGH^T)_{ij}(H\Sigma H^T)_{ij} &(11)
\end{aligned}
$$

Note that the operation between G and Σ is not matrix multiplication. The two Hs between G and Σ can not be diminished.

Interestingly, the representation of covariance matrix on Haar space induces anisotropic collinearity of object's contours (Figure.5). It means that the covariance mask constraints that a continuous contour tends to be in same local object.

There is an another advantage of use of Haar wavelet. Viola and Johns utilized Haar-like rectangle features as a feature basis, and they combined integral image method[13]. Once integral image is made before object search, the average of luminance in a given region is calculated by referring just few points of integral image. Since the same technique is applicable for the rectangle feature like Haar wavelet, the objects search cost would be considerably reduced.

For multi-resolution MCM, we measured the performance of bLDA using our original database. The image database are composed of two classes of face and non-face images. 300 face images and 700 non-face images are used in training phase, 425 face images and 1500 non-face images are testified. The best performance is 98.70% correct.

For multi-class classification we measures the performance to classify facial emotion; anger, joy, yawn. These images are sampled from AR face database[10]. The data are composed of 354 images for training, and 45 images for test. The correct performance reaches to 82.2%

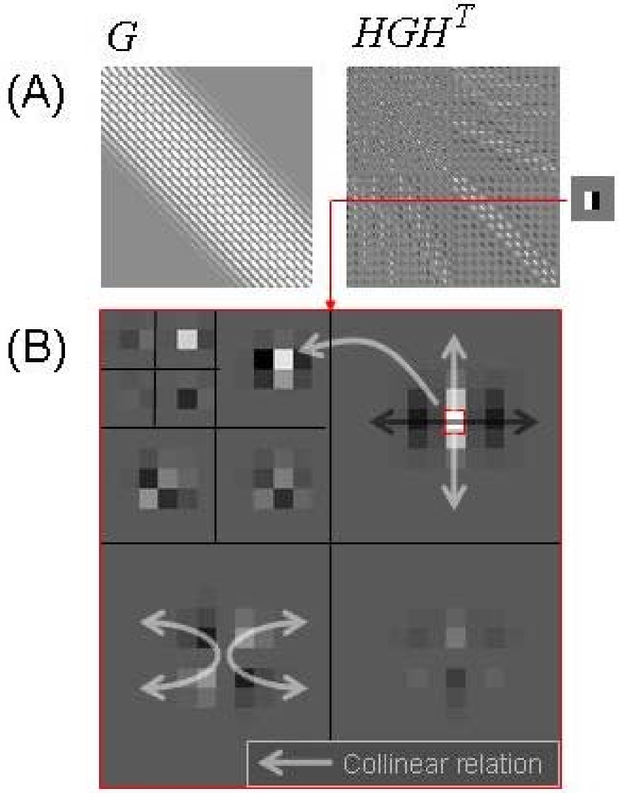

Figure 5. Covariance mask on Haar space: (A) Left matrix is the covariance matrix G on feature space. The covariance mask is designed so as to be correlated within neighborhood, σ = 5.0 pixel. Right matrix is the covariance matrix on Haar space. The transformation of the matrix is HGH^T, where H is the Haar transformation matrix. (B) The slice of matrix HGH^T at a given point (red lines), which indicates vertical-bar feature located at the center, are expanded as two-dimensional image (24x24). This image represents relation of correlation between Haar components for a given feature. The white arrows indicates the collinear relation of contour continuity. The black arrows indicates the lateral inhibition of the vertical-bar feature. There is the relation between different level of resolution.

4 Conclusion

We proposed boosted linear discriminant analysis (bLDA) and masking covariance matrix method (MCM). bLDA combined with MCM extracts effective local features and performs high discriminant ability, 98.70%, for face/non-face discrimination task.

References

[1] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recongition using class specific linear projection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[2] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting, 1998.

[3] T.-K. Kim and J. Kittler. Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(3):318–327, 2005.

[4] K.Nishida and T.Kurita. Pedestrian detection by boosting soft-margin svm with local feature selection. In *IAPR on MVA, 2005*, 2005.

[5] S. kopf, B. Mika, S. Burges, C. Knirsch, P. Miiller, K.itsch, and G. Smola. Input space vs. feature space in kernel-based methods, 1999.

[6] T. Kurita and T. Taguchi. A kernel-based fisher discriminant analysis for face detection. *IEICE Trans. on Information and Systems*, 2005.

[7] S. Li, Z. Zhang, H. Shum, and H. Zhang. Floatboost learning for classification, 2002.

[8] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *ICIP (1)*, pages 900–903, 2002.

[9] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. Boosting linear discriminant analysis for face recognition. In *ICIP (1)*, pages 657–660, 2003.

[10] A. Martinez and R.Benavente. The ar face database, 1998.

[11] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. In *Proc. 14th International Conference on Machine Learning*, pages 322–330. Morgan Kaufmann, 1997.

[12] R. E. Schapire and Y. Singer. Improved boosting using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

[13] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features, 2001.

[14] B. Wu, H. Ai, C. Huang, and S. Lao. Fast rotation invariant multi-view face detection based on real adaboost. In *FGR*, pages 79–84, 2004.

[15] J. Yang, A. F. Frangi, J. yu Yang, and D. Zhang. Kpca plus lda: A complete kernel fisher discriminant framework for feature extraction and recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(2):230–244, 2005.