Normalized LDA for Semi-supervised Learning

Bin Fan Zhen Lei Stan Z. Li
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun Donglu, Beijing 100190, China

Abstract

Linear Discriminant Analysis (LDA) has been a popular method for feature extracting and face recognition. As a supervised method, it requires manually labeled samples for training, while making labeled samples is a time consuming and exhausting work. A semi-supervised LDA (SDA [3]) has been proposed recently to enable training of LDA with partially labeled samples.

In this paper, we first reformulate supervised LDA based on the normalized perspective of LDA. Then we show that such a reformulation is powerful for semi-supervised learning of LDA. We call this approach Normalized LDA, which uses total diversity to normalize intra-class diversity and aims to find projection directions that minimize normalized intra-class diversity. Although the Normalized LDA is identical to LDA in the supervised situation, a semi-supervised approach can be easily incorporated into its framework to make use of unlabeled samples to improve the performance in the learned subspace. Moreover, different with SDA which uses unlabeled samples to preserve neighboring relations, unlabeled samples in the Normalized LDA are used for a more accurate estimation of data space. Experiments of face recognition on the FRGC version 2 database and CMU PIE database demonstrate that the Normalized LDA outperforms SDA.

1. Introduction

In the past two decades, face recognition has been widely researched in the area of pattern recognition and computer vision [12] [4] [15]. This is not only for its potential applications such as biometrics, human-computer interaction and security, but also for its theoretical challenges since it is a hard pattern classification problem with large amount of classes and large diversity in each class in its framework. One of the most successful and well-studied techniques to face recognition is the appearance-based method. PCA [7] and LDA [8] are two popular ones of this kind, which learn a subspace of face from a set of training samples. In PCA, the subspace is constructed to retain most information about original samples by Karhunen-Loeve transform. However, it does not utilize any class information and thus it may drop some important clues for classification. LDA is then proposed and it seeks subspace of features that best separating different face classes by maximizing the between class covariance and simultaneously minimizing the within class covariance. Because of the high dimensions of feature space (*e.g.* the total number of pixels in a facial image) and small sample size, the within-class scatter matrix S_w is often singular, so the optimal solution of LDA cannot be found directly. Efforts have been made to attenuate the singularity problem in the estimation of S_w. These methods including PCA+LDA (Fisher LDA) [8], Direct LDA (D-LDA) [16], Null space LDA (N-LDA) [5], regularized LDA [6] and so on. However, all these LDAs suffer the same problem, which is the weak capability of generalization when there do not exist enough training samples.

In reality, it is difficult to obtain a large number of labeled samples, but unlabeled ones are abundant, *e.g.* we can find thousands of millions facial images on the Internet. Can we use these unlabeled samples to help us improve the classifier's accuracy? If the answer is definite, then how can we use them? These are the problems to be solved in semi-supervised learning [18]. It uses both labeled and unlabeled data to find a hypothesis that accurately labels unseen examples. The research of semi-supervised learning has gained much attention in recent years [13] [17] [19] [2]. Rosenberg et al. [11] have presented a semi-supervised approach to train object detectors based on self-training, and their experimental results show that the trained model for detection is comparable to the model trained with a large number of labeled samples. Balcan et al. [1] have applied semi-supervised learning to the problem of person identification in webcam images. Roli and Marcialis [10] have used self-training to develop a semi-supervised face recognition algorithm on the basis of the standard PCA-based algorithm.

In this paper, we proposed a novel approach for semi-supervised LDA learning, called Normalized LDA, which

978-1-4244-2154-1/08/$25.00 ©2008 IEEE

is motivated by reformulating LDA in a normalized way. The Normalized LDA aims to find a group of projection directions on which the normalized intra-class diversity is minimized, equivalent to maximizing the diversity of total samples and minimizing the intra-class diversity simultaneously. With comparison to supervised LDA, a semi-supervised approach can be easily incorporated into the Normalized LDA to learn a more discriminative subspace. Specifically, the unlabeled and labeled samples are used together to get a more accurate estimation of the diversity of total sample space to improve the performance in the learned subspace. In the aspect of semi-supervised learning of LDA, a related work is Semi-supervised Discriminant Analysis (SDA) [3] proposed by Cai et al, which is essentially a regularized approach based on the assumption that nearby points will have similar embeddings. It uses unlabeled samples to preserve neighboring relations. However, different from SDA, the unlabeled samples in the Normalized LDA is used to estimate the diversity of total sample space. Statistically, more samples leads to a more accurate estimation, and hence improves the generalization of the proposed approach. Moreover, we generalize our algorithm by introducing weights for labeled samples to improve the proposed algorithm's robustness further. The experimental results show that our algorithm outperforms SDA.

The rest of this paper is organized as follows. Section 2 describes how to enable semi-supervised learning of LDA by reformulating LDA in a normalized way. In Section 3, we introduce weights for labeled samples to further improve the performance of our approach. The algorithmic procedure of our approach is presented in Section 4. Then the experimental results are shown in Section 5 and finally we conclude this paper in Section 6.

2. Semi-supervised learning of LDA

In this section, we will reformulate supervised LDA as an approach which aims to find projection vectors minimizing the normalized intra-class diversity. The intra-class diversity is normalized by the diversity of total samples. Then we will show that such a reformulation is powerful for semi-supervised learning of LDA.

2.1. Normalized Perspective of LDA

Let $\mathcal{X}^L = \{x_1, x_2, \cdots, x_N\} \subset \mathbb{R}^d$ be a training set with N labeled samples belonging to K classes $\{C_1, C_2, \cdots, C_K\}$, and $X = [x_1, x_2, \cdots, x_N]$ is the data matrix of these training samples. We denote the number of samples in the l-th class as N_l.

In LDA, intra-class scatter matrix S_w and between-class scatter matrix S_b are defined as follows:

$$S_w = \frac{1}{N}\sum_{l=1}^{K}\sum_{x_i \in C_l}(x_i - m_l)(x_i - m_l)^T \tag{1}$$

$$S_b = \frac{1}{N}\sum\nolimits_{l=1}^{K} N_l(m_l - m)(m_l - m)^T \tag{2}$$

in which $m_l = \frac{1}{N_l}\sum_{x_i \in C_l} x_i$ is the mean vector of class C_l and $m = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the total mean vector. The total scatter matrix S_t is defined as:

$$S_t = \frac{1}{N}\sum\nolimits_{i=1}^{N}(x_i - m)(x_i - m)^T \tag{3}$$

It follows that $S_t = S_w + S_b$, and then using $tr(S_t)$ to normalize $tr(S_w)$ and $tr(S_b)$, we get

$$\frac{tr(S_w)}{tr(S_t)} + \frac{tr(S_b)}{tr(S_t)} = 1 \tag{4}$$

in which $tr(A)$ denotes the trace of matrix A.

The objective of LDA is to maximize the between-class scatter matrix while minimize the intra-class scatter matrix. In Equ. (4), we can find that these two objectives are identical after normalization and can be achieved simultaneously. In other words, LDA aims to find projection vectors that either minimize $\frac{tr(S_w)}{tr(S_t)}$ or maximize $\frac{tr(S_b)}{tr(S_t)}$, which can be expressed as follows:

$$minimize \quad \frac{tr(P^T S_w P)}{tr(P^T S_t P)} \tag{5}$$

$$maximize \quad \frac{tr(P^T S_b P)}{tr(P^T S_t P)} \tag{6}$$

in which P is the transform matrix constituted by projection vectors as columns. Since

$$tr(P^T S_w P) = \frac{1}{N}\sum_{l=1}^{K}\sum_{x_i \in C_l}\left\|P^T x_i - P^T m_l\right\|^2 \tag{7}$$

$$tr(P^T S_b P) = \frac{1}{N}\sum\nolimits_{l=1}^{K} N_l\left\|P^T m_l - P^T m\right\|^2 \tag{8}$$

$$tr(P^T S_t P) = \frac{1}{N}\sum\nolimits_{i=1}^{N}\left\|P^T x_i - P^T m\right\|^2 \tag{9}$$

the $tr(P^T S_w P)$ represents the intra-class diversity, the $tr(P^T S_b P)$ represents the between-class diversity and the $tr(P^T S_t P)$ represents the total diversity. Thus the $\frac{tr(P^T S_w P)}{tr(P^T S_t P)}$ is the normalized intra-class diversity and the $\frac{tr(P^T S_b P)}{tr(P^T S_t P)}$ is the normalized between-class diversity. Therefore, the objective of LDA is reformulated either to maximize the normalized between-class diversity or to minimize the normalized intra-class diversity.

Based on the above reformulation, we propose an approach named *Normalized LDA* for semi-supervised learning of LDA. In real applications, *e.g.* face recognition, there are thousands of millions classes that makes estimating relationship between classes difficult with only a number of training samples. However, we can assume that samples in each class have a similar distribution. Taking facial images for example, the variety of different image of one person is similar to that of another person. Therefore, we take Equ. (5) as the objective of our method, which is to minimize the normalized intra-class diversity. The small value of normalized intra-class diversity means that the samples in each class scatter compactly while all the samples in the data space scatter dispersively, which indicates different classes are separated well. In the supervised situation, it is identical to LDA. However, as we shall see later, a semi-supervised approach can be easily incorporated to the Normalized LDA to enable training with partially labeled samples for the semi-supervised learning of LDA. The advantage of Normalized LDA is training with unlabeled samples.

2.2. Training with unlabeled samples

In LDA, the discriminating subspace is learned by just considering the labeled samples. When there are not enough labeled samples, its performance on testing samples can not be guaranteed. In real applications especially in face recognition, obtaining samples is an easy work but labeling these samples is a time consuming and exhausting work. Therefore, developing an algorithm that uses unlabeled samples to improve performance in the learned subspace will largely reduce human labor. In this subsection, we will show how to use unlabeled samples to improve the accuracy in the Normalized LDA as well as the reason why it can learn a more discriminative subspace.

As described in the above subsection, the objective of Normalized LDA is to maximize the total diversity and minimize the intra-class diversity. Such an objective can easily make use of unlabeled samples to improve the discriminating capability of the learned subspace. Specifically speaking, since the diversity of total samples evaluates how data scatters in the sample space, it considers the character of data space. Thus it does not utilize any class information. In other words, it does not depend upon labels of samples. Therefore, all available samples including labeled and unlabeled ones can used to calculate the total diversity. This case S_t becomes:

$$S_t{}' = \frac{1}{N_L + N_U} \sum_{x_i \in \{\mathcal{X}^L, \mathcal{X}^U\}} (x_i - m')(x_i - m')^T \quad (10)$$

in which $\mathcal{X}^L$ and $\mathcal{X}^U$ denote labeled sample set and unlabeled sample set respectively, while N_L and N_U are the number of samples of these sets, and m' is the mean vector of both labeled and unlabeled samples. In the view of statistics, more number of samples leads to a more accurate estimation of data space. In this way, the discriminating capability of the learned subspace is improved by a more accurate estimation of total diversity.

3. Weighting labeled samples

Estimating the intra-class diversity as well as the total diversity plays a key role in the Normalized LDA, and it is vital for the performance of the learned subspace. The more accurate estimation is, the more discriminative the learned subspace is. By adding a large number of unlabeled samples to the training set, estimation of total diversity of the data set will be more accurate. However, obtaining a large number of samples in each class to guarantee an accurate estimation of the intra-class diversity is nearly impractical. Usually, we can only get a few samples in each class. Thus the estimation of the intra-class diversity is sensitive to noise, since that if one sample is noise then it will distort the estimation of the intra-class diversity largely due to the lack of enough samples. In order to reduce the influence of noise and further improve the semi-supervised learning of LDA, we introduce a weight for each labeled sample x_i instead of treating labeled samples equally, denoted by $w(x_i)$. If one sample is likely to be a noise, it should have small weight. By assigning weights to different labeled samples to reduce the influence of noise, it is expected that the estimation of the intra-class diversity will be more accurate. With a weight for each labeled sample, Equ. (7) is rewritten as:

$$tr(P^T S_w P) = \sum_{l=1}^{K} \sum_{x_i \in C_l} w(x_i) \left\| P^T x_i - P^T m_l \right\|^2 \quad (11)$$

For simplicity in presentation, we denote the sum of the weights of samples in class C_l by s_l. In other words, $s_l = \sum_{i \in C_l} w(x_i)$. Then, we denote W_l to be the diagonal matrix of the weights in class C_l. Therefore, the mean vector m_l of class C_l can be rewritten as $m_l = X_l W_l e_l / s_l$, where X_l is a data matrix containing samples in class C_l, each column is a sample in C_l, and e_l is the vector of all ones whose size is the number of samples in C_l.

We then rewrite the intra-class diversity as follows:

$$\begin{aligned} tr(P^T S_w P) &= \sum\nolimits_{l=1}^{K} \sum_{i \in C_l} w(x_i) \left\| P^T x_i - P^T m_l \right\|^2 \\ &= \sum\nolimits_{l=1}^{K} \sum_{i \in C_l} w(x_i) \left\| P^T x_i - P^T X_l \frac{W_l e_l}{s_l} \right\|^2 \\ &= \sum\nolimits_{l=1}^{K} \left\| P^T (X_l - X_l \frac{W_l e_l}{s_l} e_l^T) W_l^{1/2} \right\|_F^2 \\ &= \sum\nolimits_{l=1}^{K} \left\| P^T X_l W_l^{1/2} (I - \frac{W_l^{1/2} e_l e_l^T W_l^{1/2}}{s_l}) \right\|_F^2 \end{aligned}$$

Let $M_l = I - \frac{W_l^{1/2} e_l e_l^T W_l^{1/2}}{s_l}$ and note that $s_l = e_l^T W_l e_l$, we have $M_l^2 = M_l$. Based on the fact that $\|A\|_F^2 = tr(AA^T)$, we obtain:

$$tr(P^T S_w P) = tr(P^T \sum_{l=1}^{K} A_l P) \tag{12}$$

in which

$$A_l = X_l W_l^{1/2} M_l W_l^{1/2} X_l^T \tag{13}$$

Then, the task is how to set the weight for each labeled sample to improve the robustness of Normalized LDA. We believe that the weight of one sample should be related to its location in the data space. If one sample scatters in the margin, it's more likely to be a noisy, thus it should have small weight. On the contrary, one sample should have big weight if it scatters in the dense area. Here we set the weight of one sample inverse to its distance to the mean of the class that this sample belongs to, which is

$$w(x_i \in C_l) \propto \frac{1}{Dis(x_i, m_l)}$$

where $Dis()$ is a function of distance. This paper uses Euclidean distance.

4. The Algorithm

Given a set $\mathcal{X}^L = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ of samples belonging to K classes as labeled sample set in the training set, in which y_i is the label of the i-th sample. Without loss of generality, we assume that $\mathcal{X}^L$ is ordered according to sample's label, which is $\mathcal{X}^L = \{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_K\}$ where $\mathcal{X}_i$ is the set of samples in the i-th class. Besides, we have an unlabeled sample set $\mathcal{X}^U = \{x_{N+1}, x_{N+2}, \cdots, x_{N+M}\}$ in the training set. Assume the transform matrix which we want to find is P, the steps of the Normalized LDA can be summarized as follows:

1. **Calculate the total diversity**: According to Equ. (10), calculate the total scatter matrix $S_t{}'$ and $tr(P^T S_t' P)$ is the total diversity.

2. **Calculate the intra-class diversity**: Either assign a weight to each sample in $\mathcal{X}_l$ inverse to its distance to the mean of samples in $\mathcal{X}_l$ and then calculate A_l according to Equ. (13) or calculate the intra-class scatter matrix S_w according to Equ. (1). Thus the intra-class diversity is $tr(P^T S_w{}' P)$ in which $S_w{}' = \sum_{l=1}^{K} A_l$ or $S_w{}' = S_w$.

3. **Find optimal solution**: The objective of the Normalized LDA is to find a transform matrix P that minimize the normalized intra-class diversity as follows:

$$minimize \quad \frac{tr(P^T S_w{}' P)}{tr(P^T S_t{}' P)} \tag{14}$$

The solution can be obtained by solving the generalized eigen-problem shown in the following:

$$S_w{}' P = \lambda S_t{}' P \tag{15}$$

Then, computing the eigenvectors and eigenvalues in Equ. (15), and sorting these eigenvectors according to the value of their corresponding eigenvalues in an ascend manner. The sorted eigenvectors are denoted by $[V_1, V_2, \cdots, V_d]$, where d is the dimension of sample.

4. **Select projection directions**: Construct the transformation matrix due to the number of extracted features. If we want to extract $N_f (N_f < d)$ features, then the transformation matrix is $P = [V_1, V_2, \cdots, V_{N_f}]$.

5. Experiments

In this section, we conduct experiments on face recognition to test our algorithm, using FRGC version 2 database [9] and CMU PIE database [14]. In the experiments, each facial image in the databases is in grey scale and cropped to be 71×60 pixels by fixing the positions of two eyes. Fig. 1 shows some sample facial images in the databases. PCA is used to all of samples in the training set, the gallery set and the probe set to reduce the computational complexity. After extracting features, the nearest neighbor classifier with Euclidian distance metric is adopted to classify the samples. Experimental results on both of these two databases demonstrate that the proposed approach in this paper outperforms the SDA.

5.1. On FRGC Database

The query set of facial images in the FRGC version 2 is used for the experiment to evaluate our algorithm. There are 8104 facial images of 466 subjects containing the variations of illumination, expression, time and blurring. Since there are only a few images available for some persons in this set, a labeled subset is selected to ensure the number of images of each subject in the selected subset is no less than 10. Specifically, the first 10 facial images are taken into the subset if the number of images of this subject is no less than 10. Thus we get a subset of 3160 facial images of 316 subjects. Then, the selected labeled subset is further divided into three subsets: the training set, the probe set and the gallery set. First, 200 subjects' images are selected as labeled samples in the training set, and the remaining 116 subjects' images are exploited as the gallery set and the probe set. Second, five facial images of each subject in the remaining 116 subjects are selected as the gallery set and the other five images are used as the probe set. The images in the query set but not in the selected labeled subset are served as the unlabeled samples in the training set.

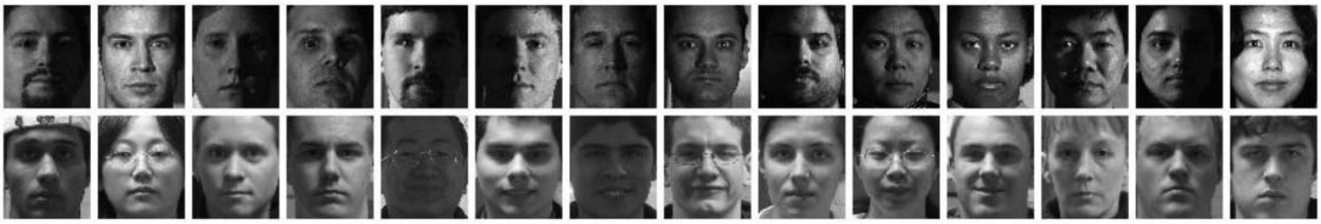

Figure 1. Sample facial images in the data set. First row shows samples in the CMU PIE database, and second row shows sample images in the FRGC database

To reduce variation, the experiment is repeated ten times through random selection. Thus the results shown in Fig. 2 are the average results. In Fig. 2, Norm-LDA indicates the proposed method of this paper, W-Norm-LDA indicates Norm-LDA with samples have different weights discussed in the Section 3 and SDA indicates Cai's method [3]. In Cai's method, there are two parameters needed to be set, which are number of nearest neighbors p that is used to calculate the graph matrix and the coefficient α controls balance between the model complexity and the empirical loss. Since there is no guide for setting these parameters in [3], we set the value of them to be 3 and 0.01 respectively which are the optimal values according to the results of many times of experiment. For comparison, we also demonstrate the result of LDA. We can find that by adding unlabeled samples to the training set, it does improve the discriminating capability of the learned subspace, either our method or Cai's method. However, compared with Cai's method, our method has a better performance on the improvement of classifying accuracy. Moreover, there is no parameter to be set in our method while Cai's method need to set several parameters which is depend on experiences. Furthermore, the higher performance of W-Norm-LDA compared with Norm-LDA validates that our strategy for introducing different weight to each sample is useful.

5.2. On CMU PIE Database

The CMU PIE database contains 68 subjects with 41,368 facial images under 13 different poses, 43 different illumination conditions, and with 4 different expressions. The frontal pose (C27) subset of CMU PIE database is chosen for our experiment. It contains facial images with varying lighting and illumination with fixed pose and expression. In the experiment, we randomly choose 30 subjects for training. For each subject, 10 images are randomly selected as the labeled samples in the training set. The probe set and the gallery set are selected from the images of 20 subjects within the remaining subjects in the data set. Specifically, we first select 10 images in each person of the selected 20 persons to be exploited as the probe set and the gallery set. Then, five images in one subject are chosen as the probe set and the other five images of this subject are chosen as the gallery set. The the unlabeled samples in the training set are selected from the remaining images of this subset (C27). Finally, we obtain a training set with 300 labeled images from 30 subjects and 907 unlabeled samples, a probe set with 100 images belonging to 20 subjects and a gallery set with 100 images belonging to the same subjects as the images in the probe set. The set of persons for training is disjoint with that of persons in the gallery as well as the probe. The same as the experiment performed on the FRGC data set, we average the results over 10 random splits to reduce variation.

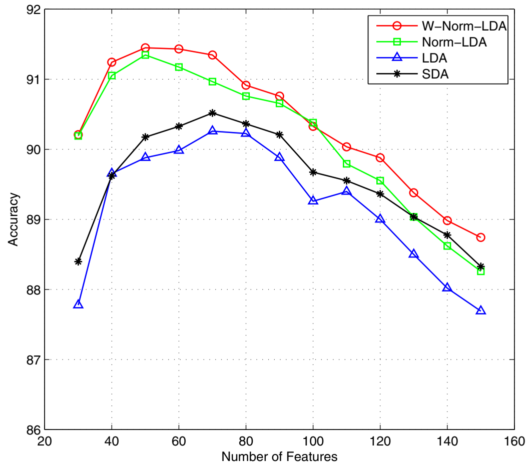

Figure 2. Experimental result on FRGC

Fig. 3 demonstrates the experimental results on CMU PIE database which are similar to the results on FRGC database. It can be found that the proposed method, either Norm-LDA or W-Norm-LDA, has better performance than the SDA. Meanwhile, the W-Norm-LDA outperforms Norm-LDA in Fig. 3, which further indicates that the strategy for introducing different weight to each sample is useful.

6. Conclusion

This paper proposed a new method called Normalized LDA for semi-supervised LDA learning, which is motivated

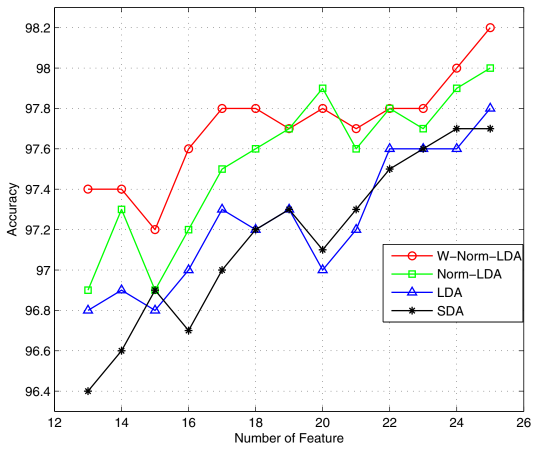

Figure 3. Experimental result on CMU PIE

by reformulating LDA in a normalized way. It can make efficient use of unlabeled samples to improve the discriminating capability of the learned subspace. Specifically, both unlabeled and labeled samples are used to maximize the diversity of the data space, while labeled samples in each class are used to minimize the intra-class diversity. By introducing weights for labeled samples, the performance of Normalized LDA is further improved. The experiments of face recognition on FRGC version 2 database and CMU PIE database demonstrate the effectiveness of the proposed approach.

7. Acknowledgement

This work was supported by the following funds: Chinese National Natural Science Foundation Project #60518002, Chinese National 863 Program Projects #2006AA01Z192, #2006AA01Z193, and #2006AA780201-4, Chinese National Science and Technology Support Platform Project #2006BAK08B06, Chinese Academy of Sciences 100 people project, and AuthenMetric R&D Funds.

References

[1] M.-F. Balcan, A. Blum, P. P. Choi, J. Lafferty, B. Pantano, M. R. Rwebangira, and X. Zhu. Person identification in webcam images: An application of semi-supervised learning. In *ICML Workshop on Learning with Partially Classified Training Data*, 2005.

[2] M. Belkin, P. Niyogi, and V. Sindhwani. On manifold regularization. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics(AISTAT 2005)*, 2005.

[3] D. Cai, X. He, and J. Han. Semi-supervised discriminant analysis. In *Proceedings of International Conference On Computer Vision*. IEEE, 2007.

[4] R. Chellappa, C. L.Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. In *Proceedings of the IEEE*, volume 83, pages 705 – 740. IEEE, 1995.

[5] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu. A new lda-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 33:1713 – 1726, 2000.

[6] J. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84:165 – 175, 1989.

[7] M.Turk and A.Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71 – 86, 1991.

[8] P.N.Belhumeur, J.P.Hepanha, and D.J.Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transaction on Pattern Analysis and Machine Learning*, 19(7):711 – 720, 1997.

[9] P.Phillips, P.Flynn, T.Scruggs, and K.Bowyer. Overview of the face recognition grand challenge. In *Proceedings of Computer Vision and Pattern Recognition*. IEEE, 2005.

[10] F. Roli and G. Marcialis. Semi-supervised pca-based face recognition using self-training. In *Int. Workshop. on Structural and Syntactical Pattern Recognition and Statistical Techniques in Pattern Recognition(S+SSPR06)*, volume Springer LNCS 4109, pages 560 – 568. IAPR, 2006.

[11] C. Rosenberg, M. Hebert, and H. Schneiderman. Semi-supervised self-training of object detection models. In *Seventh IEEE Workshop on Application of Computer Vision*, volume 1, pages 29 – 36. IEEE, 2005.

[12] A. Samal and P. A. Iyengar. Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern Recognition*, 25:65 – 77, 1992.

[13] M. Szummer and T. Jaakkola. Partially labeled classification with markov random walks. In *NIPS*, 2001.

[14] T.Sim, S.Baker, and M.Bsat. The cmu pose, illumination, and experssion database. *IEEE Transaction on Pattern Analysis and Machine Learning*, 25(12):1615 – 1618, 2003.

[15] W.Zhao, R.Chellappa, P.J.Phillips, and A.Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys (CSUR)*, 35:399 – 458, 2003.

[16] H. Yu and J. Yang. A direct lda algorithm for high-dimensional data – with application to face recognition. *Pattern Recognition*, 34:2067 – 2071, 2001.

[17] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In *NIPS*, 2003.

[18] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[19] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields. In *ICML*, 2003.