Side-Information based Linear Discriminant Analysis for Face Recognition

Meina Kan[1,2,3]
mnkan@jdl.ac.cn
Shiguang Shan[1,2]
sgshan@jdl.ac.cn
Dong Xu[3]
dongxu@ntu.edu.sg
Xilin Chen[1,2]
xlchen@jdl.ac.cn

[1] Digital Media Research Center, Institute of Computing Technology, CAS, Beijing, China
[2] Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences, Beijing, China
[3] School of Computer Engineering, Nanyang Technological University, Singapore

Abstract

In recent years, face recognition in the unconstrained environment has attracted increasing attentions, and a few methods have been evaluated on the Labeled Faces in the Wild (LFW) database. In the unconstrained conditions, sometimes we cannot obtain the full class label information of all the subjects. Instead we can only get the weak label information, such as the side-information, *i.e.*, the image pairs from the same or different subjects. In this scenario, many multi-class methods (*e.g.*, the well-known Fisher Linear Discriminant Analysis (FLDA)), fail to work due to the lack of full class label information. To effectively utilize the side-information in such case, we propose Side-Information based Linear Discriminant Analysis (SILD), in which the within-class and between-class scatter matrices are directly calculated by using the side-information. Moreover, we theoretically prove that our SILD method is equivalent to FLDA when the full class label information is available. Experiments on LFW and FRGC databases support our theoretical analysis, and SILD using multiple features also achieve promising performance when compared with the state-of-the-art methods.

1 Introduction

In the past few decades, face recognition has received increasing attentions due to its wide potential applications in various fields. As surveyed in [1, 2, 3, 4], numerous methods have been proposed, such as Eigenface [5], Fisherface [6], Bayesian face recognition [7], Elastic Bunch Graph Matching [8], Gabor Fisher Classifiers [9], Sparse representation [10], and so on. Most of them work well in constrained environments as evaluated on some public databases, such as ORL [11], AR [12], PIE [13], XM2VTS [14] and FERET [15]. However due to the large appearance variations in pose, aging, lighting, occlusion, expression and so on, many of them degenerate seriously when applied to the unconstrained environment [16].

In order to promote face recognition in the unconstrained environment, a large scale database, Labeled Faces in the Wild (LFW) [17] is released recently. LFW is collected with "natural" variability that may be encountered in our daily life including pose, lighting, expression, age, gender, race and so on which makes this database suitable for evaluating the face recognition technologies in unconstrained environment. LFW has two different training modes: image-restricted mode and image-unrestricted mode. In the former mode, only *side-information*, *i.e.*, whether a pair of images belongs to the same class (also referred as image pairs hereafter), is available, while in the latter mode, the full class label information is provided. Compared with the latter case, the former case is more common in real world and also more challenging since only partial information is provided.

After the release of LFW database, a few methods [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30] have been specifically designed for it, and evident progress can be observed from the reported results in [21, 22, 23, 29, 30]. These methods can be generally categorized into two categories: feature-oriented approaches and similarity-oriented approaches. The former category aims to extract effective features for face representation, while the latter focuses more on the face similarity computation.

Among the feature-oriented methods, local descriptor based methods are more popular. For instance, Wolf *et al*. [19] proposed three-patch Local Binary Pattern (LBP) and four-patch LBP to encode the similarities between neighbouring patches of pixels in order to capture the information complementary to the original LBP features. In [20], each face was described in terms of multi-region probabilistic histograms of visual words. In [21], Cao *et al*. encoded the micro-structures of face by using an unsupervised learning-based encoding method. In [25], a discriminative and robust feature descriptor called Patterns of Oriented Edge Magnitudes (POEM) was built by applying a self-similarity based structures on the oriented magnitudes. In [30], N. Pinto *et al*. used the biologically-inspired visual representations selected by feature search. Moreover, the similarities among face images were also exploited as features. In [22], Kumar *et al*. proposed a simile classifier using the similarity of faces to some specific reference people as features. In [23], Wolf *et al*. used the ranking of the images most similar to a query image as the descriptor of this query image. Additionally, Pinto *et al*. [24] investigated the capability of face recognition system with a modern face recognition test set using only simple features.

Similarity-oriented methods aim at novel metric computation between two face images. Typically, in [19, 31], one-shot similarity was employed to measure the likelihood of each sample belonging to the same class as the other. It was further extended to two-shot similarity [23] and multiple one-shot similarity by utilizing class label information [28] respectively. In [26], Nowak *et al*. obtained the similarity using characteristic difference of local descriptors sampled from images that are quantized with an ensemble of extremely randomized binary trees. In [27], two distance measures were proposed, including a logistic discriminant based approach and a nearest neighbour based approach which computed the probability of two images belonging to the same class. In [29], an metric learning method designed for cosine similarity was proposed and it achieved promising performance.

Many of the above methods deal with the side-information scenario by employing the typical two-class SVM classifier. However the multi-class methods including Fisherface [6] and its numerous extensions cannot be used in this scenario because the crucial class label information are not provided in the image-restricted evaluation mode.

In this work we propose a Side-Information based Linear Discriminant Analysis (SILD) method that can work well only with side-information, in which the within-class

and between-class scatter matrices are computed by directly using the side-information. It is worth mentioning that our method is different from the two-class FLDA, specifically only one projection direction can be obtained by using two-class FLDA while much more projection directions can be obtained by using our method. Moreover, we theoretically prove that, our SILD method is equivalent to multi-class FLDA when class labels are provided.

The remainder of this paper is organized as follows. Section 2 describes the side-information based linear discriminant analysis. Section 3 details the experimental evaluations of SILD on LFW database. Finally, conclusions are given in section 4.

2 Side-Information based Linear Discriminant Analysis

In this section, we first give a brief description of FLDA, and then present the definition of our SILD method that is applicable in scenario of side-information. Finally we prove that the new definition is equivalent to the traditional definition when class label is provided.

2.1 Fisher Linear Discriminant Analysis (FLDA)

Fisher Linear Discriminant Analysis aims to find a set of most discriminative linear projections by maximizing the ratio of the determinant of the between-class scatter matrix to that of the within-class scatter matrix:

$$W_{opt} = \arg\max_{W} \frac{\left|W^T S_B W\right|}{\left|W^T S_W W\right|} \tag{1}$$

The within-class scatter matrix S_W and between-class scatter matrix S_B are defined as:

$$S_W = \sum_{i=1}^{c}\sum_{j=1}^{n_i}\left(x_{ij} - m_i\right)\left(x_{ij} - m_i\right)^T \tag{2}$$

$$S_B = \sum_{i=1}^{c}\left(m_i - m\right)\left(m_i - m\right)^T \tag{3}$$

where c is the number of classes in the training set, x_{ij} is the j^{th} sample from i^{th} class, n_i is the number of samples from the i^{th} class, m_i is the mean of the i^{th} class, and m is the mean of all samples in the training set. The problem in (1) can be solved by a two-step method [32].

Fisrtly, S_W is diagonalized as follows:

$$S_W = H\Lambda H^T \tag{4}$$

$$\left(H\Lambda^{-1/2}\right)^T S_W \left(H\Lambda^{-1/2}\right) = I \tag{5}$$

Secondly, S_B is also diagonalized:

$$\left(H\Lambda^{-1/2}\right)^T S_B \left(H\Lambda^{-1/2}\right) = U\Sigma U^T \tag{6}$$

Finally, the projection matrix can be computed as:

$$W_{opt} = H\Lambda^{-1/2}U \tag{7}$$

where H and U are orthogonal matrices and Λ and Σ are diagonal matrices. As shown in [1, 6, 9], FLDA is a simple but effective method for face recognition.

2.2 Side-Information based Linear Discriminant Analysis (SILD)

However (see (2) and (3)), the class label of each sample need to be known in FLDA, so it cannot work in case that only side-information is available. The same as in [33], side-information, one type of weak label information, depicts whether a pair of images belong to the same class. In this case, FLDA fails to work because the S_W and S_B cannot be computed without the full class label information.

To address this problem, we propose a new definition for S_W and S_B that directly exploits the side-information. Specifically, the same-class image pairs are directly used to calculate the within-class scatter matrix and the different-class image pairs are employed to calculate the between-class scatter matrix.

Let us denote $S = \{(x_i, x_j) : l(x_i) = l(x_j)\}$ as the set of same-class image pairs and $D = \{(x_m, x_n) : l(x_m) \neq l(x_n)\}$ as the set of different-class image pairs, with $l(x)$ representing the class label of image x. Then, the within-class and between-class scatter matrices can be respectively defined as follows:

$$S_W^{sild} = \sum_{(x_i, x_j)\in S}\left(x_i - x_j\right)\left(x_i - x_j\right)^T \tag{8}$$

$$S_B^{sild} = \sum_{(x_m, x_n)\in D}\left(x_m - x_n\right)\left(x_m - x_n\right)^T \tag{9}$$

Compared with (2) and (3) in FLDA, the new definition do not need know the identity of each sample and only use the weakly-supervised side-information to directly calculate the total within-class and between-class scatter matrices.

Similarly to FLDA, the projection matrix in SILD can be obtained by solving the following optimization problem:

$$W_{opt}^{sild} = \arg\max_{W} \frac{\left|W^T S_B^{sild} W\right|}{\left|W^T S_W^{sild} W\right|} \tag{10}$$

Similarly as FLDA, SILD in (10) can also be solved by (4)-(7). Obviously, the size of set S and D can influence the stability of the new definition for S_W and S_B. Generally only a small fraction of image pairs in SILD can be generated from the class label information. In this case, the new within-class scatter matrix may have a large number of very small eigenvalues. In order to suppress the instability caused by the small eigenvalues, in our implementation, we only use the eigenvectors corresponding to the largest eigenvalues when diagonalizing the within-class scatter matrix. As the first step, S_W^{sild} is diagonalized as follows:

$$S_W^{sild} = H\Lambda H^T \tag{11}$$

Let us define Λ' as a small fraction of the columns of Λ with the eigenvalues corresponding to the top part of overall energy. In this work Λ' is used for the consequent computations in (5)-(7) instead of Λ to cope with the instability. As observed in our experiment, the less the side information, the larger the number of very small eigenvalues. So the Λ' corresponding to a smaller fraction can achieve a better performance given the less side-information. Usually Λ' corresponding to the top 80%~90% of overall energy can performs well which is about 30% of the columns of Λ.

2.3 Equivalence of FLDA and SILD in case of knowing class label

In this section, we prove that, the proposed SILD is equivalent to FLDA when the class label information is provided. Specifically, if all the classes have the same number of samples, SILD is identical to FLDA. Otherwise our SILD can be seen as an interesting variant of FLDA.

Here we assume that there are r samples in c classes and n_i samples in the i^{th} class. In case that the class label is provided, the set of the same-class image pairs S should consist of all the possible image pairs belonging to the same class, while the set of different-class image pairs D should be formed by all the image pairs whose class labels are different. Then the within-class scatter matrix of SILD can be rewritten as follows (more details can be found in appendix A):

$$S_W^{sild} = \sum_{i=1}^{c}\sum_{k=1}^{n_i}\sum_{l=1}^{n_i}(x_{ik}-x_{il})(x_{ik}-x_{il})^T = 2\sum_{i=1}^{c} n_i \sum_{k=1}^{n_i}(x_{ik}-m_i)(x_{ik}-m_i)^T \tag{12}$$

If all classes have the same number of samples, denoted as n, the within-class scatter matrix in SILD can be further formed as:

$$S_W^{sild} = 2\sum_{i=1}^{c} n \sum_{k=1}^{n_i}(x_{ik}-m_i)(x_{ik}-m_i)^T = 2nS_W \tag{13}$$

It means the newly defined within-class scatter matrix is equal to that of FLDA only up to a scale parameter.

Similarly, the between-class scatter matrix can be reformulated as (see appendix A for more details):

$$S_B^{sild} = 2rS_B - S_W^{sild} + 2rS_W \tag{14}$$

Given the new definition of within-class and between-class scatter matrices, the projection matrix of SILD can be solved as follows (see appendix A for more details):

$$W_{opt}^{sild} = \arg\max_W \frac{\left|W^T S_B^{sild} W\right|}{\left|W^T S_W^{sild} W\right|} = \arg\max_W trace\left(\frac{W^T S_T W}{W^T S_W^{sild} W}\right) \tag{15}$$

If all classes have the same number of samples, we further have:

$$W_{opt}^{sild} = \arg\max_W trace\left(\frac{W^T S_T W}{2nW^T S_W W}\right) = \arg\max_W \frac{\left|W^T S_B W\right|}{\left|W^T S_W W\right|} = W_{opt} \tag{16}$$

From the above equations, it is obvious that the projection matrix of SILD is identical to that of FLDA if the class label information is provided and all classes have the same number of samples. If each class has different number of samples, SILD is a variant of FLDA by focusing more on the classes with more samples (see the weight n_i in (12)). This leads to more robust calculation of the within-class scatter matrix by suppressing the unreliable classes with fewer samples. When the class label information is unavailable, SILD can be considered as an approximation of FLDA by exploiting a small fraction of full class label information only.

2.4 Boundary weighted SILD

Inspired by [34], a more discriminative model can be learnt if the samples near the boundary are emphasized. However, the method in [34] cannot be directly used without the class label information. As in (8), if a pair of samples (x_i, x_j) from the same person are far

away from each other, then $(x_i - x_j)$ will have large values and so it plays more contribution for S_W^{sild}, otherwise it plays less contribution for S_W^{sild}. So the samples that are hard to be classified are emphasized in the definition of S_W^{sild}. On the other hand, for S_B^{sild}, $(x_i - x_j)$ will be small in this case which means less attention is paid on if the pair of samples (x_i, x_j) from different persons are very similar, *i.e.*, a difficult pair that we should pay more attention to. So, we reweight the pairs in S_B^{sild} to emphasize the samples that are hard for classification as follows:

$$S_B^{sild} = \sum_{(x_m,x_n)\in D} w(x_m,x_n)(x_m-x_n)(x_m-x_n)^T$$
$$w(x_m,x_n) = cosine(x_m,x_n) \tag{17}$$

When the S_B^{sild} is calculated with (17), we refer to our method as 'Weighted SILD'.

3 Experiments

In this section, we first use LFW [17] and FRGC [35] databases to verify the equivalence of SILD and FLDA when the class label information is available. Then we compare SILD with the state-of-the-art methods on the unconstrained LFW database. The task on both databases is face verification.

LFW database has 13,233 images from 5,749 individuals with the resolution of 250 by 250. It is divided into two views. View 1 is employed for model selection, and view 2 is used for performance evaluation. In view 2, two training modes are designed including image-restricted training mode where only pair-wise samples are available and image-unrestricted training mode where the class label information for each sample is provided.

In our experiments, all face images are simply cropped to 80x150 pixels by just cutting out the centre region of the images provided by Wolf *et al.* [23]. In order to reduce the high dimensionality and suppress noise, PCA is employed as a pre-processing method. The dimension after PCA is determined by preserving about 95% energy. The similarity of two feature vectors is measured by cosine similarity.

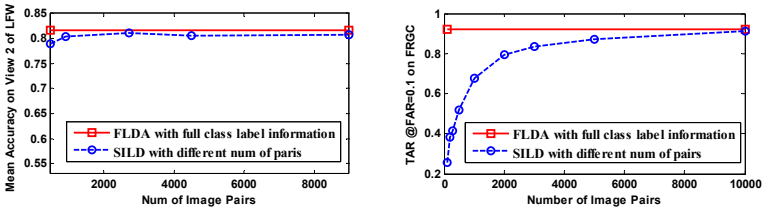
SILD also is tested on the experiment 4 of FRGC database, which has 12766 training images from 222 persons, 16028 target images, and 8014 query images from 466 persons. For FRGC database, the images are cropped to a smaller resolution with 40x50 pixels. Histogram equalization is used as pre-processing and the grey intensity is exploited as features. The PCA is also applied for reducing dimensions with 95% energy preserved.

3.1 Equivalence of FLDA and SILD under label information

As proved in previous, SILD can obtain the exact model as FLDA if class label information is provided and all classes have the same number of samples. And SILD should obtain almost the same model as FLDA if have different number of samples. In this section, the class label information is provided to verify this equivalence. Given full class label information, FLDA can be directly computed according to (1)-(3). While for SILD, about millions of pairs for S and D are formed using the class label information. Considering the limited computing resources, only a small part of possible pairs are randomly sampled.

We employ the LFW and FRGC databases for this evaluation. Specifically, 450, 900, 2700, 4500, 9000 (*resp.* 100, 300, 500, 1000, 2000, 3000, 5000, 10000) same-class image

pairs and the same number of different-class image pairs are randomly selected respectively for LFW (*resp*. FRGC).



(a) LFW (b) FRGC

Figure 1: Performance of SILD with different number of pairs.

As displayed in Fig.1, the result of FLDA is displayed as solid red line, while the results of SILD with different number of image pairs are plotted as dashed blue line. From Fig. 1, we observe that SILD can achieve comparable performance compared with FLDA (only with a small gap less than 1%) when using about 10000x2 sampled image pairs (*i.e.*, less than 1% of the total pairs). In indicates that SILD can obtain a model equivalent to FLDA if the full class label information is provided and SILD can achieve a comparable performance to FLDA if only a small fraction of side-information is provided.

3.2 Comparison with the state-of-the-art methods

In this section, the proposed SILD is compared to the state-of-the-art methods on LFW database according to the image-restricted protocol, including background samples based method [23], attribute and simile classifiers [22], multiple LE [21], cosine similarity metric learning [29], biologically-inspired feature based method [30] and other methods listed in [36].

The proposed SILD is tested using several well-known features. In the default setting, only the original intensity feature is used for SILD. To further improve the accuracy of SILD, Local Binary Patterns [28, 37] and Gabor wavelet feature [9] are also employed. Besides, similar to most of the above state-of-the-art methods on the LFW database, we also report the best result of SILD by combining different types of features.

Feature Name	Feature Type	SILD	Weighted SILD
Intensity	Original feature	0.8070 ± 0.0219	0.8020 ± 0.0213
	Square root feature	0.8026 ± 0.0212	0.8010 ± 0.0176
LBP	Original feature	0.8007 ± 0.0135	0.8412 ± 0.0108
	Square root feature	0.7958 ± 0.0132	0.8485 ± 0.0112
Gabor	Original feature	0.7898 ± 0.0184	0.7902 ± 0.0186
	Square root feature	0.8043 ± 0.0208	0.8102 ± 0.0201
Block Gabor	Original feature	0.8221 ± 0.0133	0.8233 ± 0.0164
	Square root feature	0.8443 ± 0.0151	0.8452 ± 0.0139
After Combination	**8 similarities combined**	**0.8578 ± 0.0205**	**0.8768 ± 0.0159**

Table 1: Mean accuracy of SILD with different types of feature on the LFW database.

The intensity feature is directly extracted by vectoring each grey-scale image to a 12,000-D feature vector. For LBP features, a histogram of 59 bins is extracted for each non-overlap block with the size of 10x10, and then all histograms are concatenated into one single 7,080-D vector. The Gabor features are extracted with 5 scales and 8 orientations, which leads to a quite high dimension. Therefore we adopt a 10x10 scaling factor to down-sample them. However, much structure information is lost after such a large scale down-sampling process. So Gabor images are also divided into 12 non-overlapping blocks as an alternative complement, and in each block a 2x2 down-sampling is employed to obtain a lower dimensional feature.

In this work, 'Intensity-SILD', 'LBP-SILD', 'Gabor-SILD', 'Block Gabor-SILD' means that SILD is combined with Intensity feature, LBP feature, 10x10 down-sampled Gabor feature and block based Gabor features respectively

In addition, the square root of the original features are also used as suggested in [23, 29]. Finally, the similarity scores of all the 8 types of features, including 4 types of original features and 4 square root features, are combined to further boost the accuracy by using SVM with RBF kernel which is denoted as 'Combined SILD'.

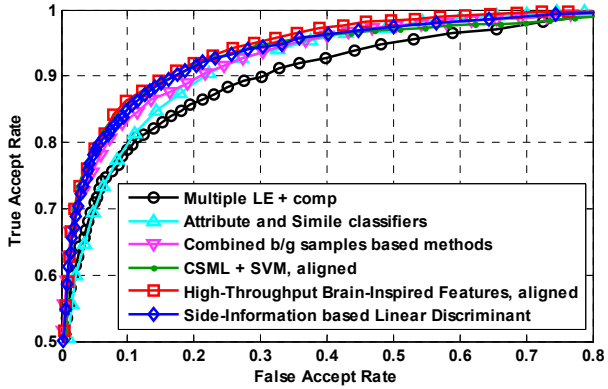


Figure 2: Performance of combined SILD and other state-of-the-art methods on the LFW database under image-restricted protocol.

The mean accuracies of SILD with different features are listed in Table 1. From it we can observe that the best accuracy of SILD with Intensity, LBP, Gabor and Block-Gabor features are 80.7%, 84.85%, 81.02%, 84.52% respectively. Compared with the single feature based method 'Single LE', SILD works better when with LBP and Block-Gabor feature and performs almost comparable when with Intensity and Gabor feature. It is indicates that the proposed SILD can perform well just using single type of low-level features. We also observe that the weighted SILD is better than SILD in most cases.

Table 2 compare SILD with multiple features with the state-of-the-art methods on the LFW database, and Fig. 2 shows the corresponding ROC curve. After fusing multiple features, SILD can achieve a higher result 87.68%. This result is also comparable to the state-of-the-art result 88% from Nguyen *et al*. [29] and 88.13% from Pinto *et al*.[30]. However, SILD only exploits four types of features without using the complex learning and searching process. It demonstrates that SILD achieves the state-of-the-art result by effectively exploiting the side-information.

Methods	Feature Type and similarities combined	Mean Accuracy [1]
Combined b/g samples based methods [23]	10 feature types, 60 similarities	0.8683 ± 0.0034
Attribute and Simile classifiers [22]	65~3000 similarities as feature	0.8529 ± 0.0123
Single LE + holistic [21]	1 feature type, 1 similarity	0.8122 ± 0.0053
Multiple LE + comp [21]	4 feature types, 9component, 36 similarities	0.8445 ± 0.0046
CSML + SVM [29]	6 feature types, 6 similarities	0.8800 ± 0.0037
High-Throughput Brain-Inspired Features [30]	11 feature types by feature selection, 3 rescaled crops, 33 similarities	0.8813 ± 0.0058
Weighted SILD after feature combination	**8 feature types, 8 similarities**	**0.8768 ± 0.0159**

Table 2: Mean accuracy of SILD with different types of feature on the LFW database.

4 Conclusion

By redefining the within-class and between-class scatter matrices based on the same-class and different-class sample pairs, Side-information based Linear Discriminant Analysis is proposed and applied in the scenario when only side-information is available. We have theoretically proved the equivalence of SILD to Fisher linear discriminant analysis. The comprehensive experiments demonstrate that SILD can achieve comparable results when compared with the state-of-the-art results which are obtained by using more type of features or learning process.

5 Acknowledge

This paper is partially supported by the Singapore National Research Foundation & Interactive Digital Media R&D Program Office, MDA under research grant NRF2008IDM-IDM004-018, and also partially supported by Natural Science Foundation of China under contracts No. 60803084; National Basic Research Program of China (973 Program) under contract 2009CB320902; and Beijing Natural Science Foundation (New Technologies and Methods in Intelligent Video Surveillance for Public Security) under contract No.4111003.

Appendix A

1. Inference for within-class scatter matrix of SILD in (12)

$$S_W^{sild} = \sum_{i=1}^{c}\sum_{k=1}^{n_i}\sum_{l=1}^{n_i}(x_{ik}-x_{il})(x_{ik}-x_{il})^T = \sum_{i=1}^{c}\left(\sum_{k=1}^{n_i}\sum_{l=1}^{n_i}x_{ik}x_{ik}^T - \sum_{k=1}^{n_i}\sum_{l=1}^{n_i}x_{ik}x_{il}^T - \sum_{k=1}^{n_i}\sum_{l=1}^{n_i}x_{il}x_{ik}^T + \sum_{k=1}^{n_i}\sum_{l=1}^{n_i}x_{il}x_{il}^T\right)$$
$$= \sum_{i=1}^{c}\left(n_i\sum_{k=1}^{n_i}x_{ik}x_{ik}^T - n_i\sum_{k=1}^{n_i}x_{ik}m_i^T - n_i\sum_{k=1}^{n_i}m_ix_{ik}^T + n_i\sum_{k=1}^{n_i}x_{ik}x_{ik}^T\right)$$
$$= 2\sum_{i=1}^{c}n_i\left(\sum_{k=1}^{n_i}x_{ik}x_{ik}^T - \sum_{k=1}^{n_i}x_{ik}m_i^T - \sum_{k=1}^{n_i}x_{ik}m_i^T + n_im_im_i^T\right) = 2\sum_{i=1}^{c}n_i\sum_{k=1}^{n_i}(x_{ik}-m_i)(x_{ik}-m_i)^T$$

[1] The mean accuracy is same as in [17] and results of other methods are from their papers shown in [36].

2. Inference for between-class scatter matrix of SILD in (14)

$$S_B = S_T - S_W = \frac{1}{2r}\sum_{i=1}^{c}\sum_{j=1}^{c}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j}(x_{ik}-x_{jl})(x_{ik}-x_{jl})^T - \sum_{i=1}^{c}\sum_{k=1}^{n_i}(x_{ik}-m_i)(x_{ik}-m_i)^T$$
$$= \frac{1}{2r}\sum_{i=1}^{c}\sum_{j\neq i}^{c}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j}(x_{ik}-x_{jl})(x_{ik}-x_{jl})^T + \frac{1}{2r}\sum_{i=1}^{c}\sum_{j=i}^{c}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j}(x_{ik}-x_{jl})(x_{ik}-x_{jl})^T - \sum_{i=1}^{c}\sum_{k=1}^{n_i}(x_{ik}-m_i)(x_{ik}-m_i)^T$$
$$= \frac{1}{2r}S_B^{sild} + \frac{1}{2r}S_W^{sild} - S_W$$

$$\Rightarrow S_B^{sild} = 2rS_B - S_W^{sild} + 2rS_W$$

3. Inference for model of SILD in (15)
(refer to [38] for the equivalence of determinant ratio and ratio trace)

$$W_{opt}^{sild} = \arg\max_W \frac{\left|W^TS_B^{sild}W\right|}{\left|W^TS_W^{sild}W\right|}$$
$$= \arg\max_W trace\left(\frac{W^TS_B^{sild}W}{W^TS_W^{sild}W}\right) = \arg\max_W trace\left(\frac{W^T\left(2rS_B - S_W^{sild} + 2rS_W\right)W}{W^TS_W^{sild}W}\right)$$
$$= \arg\max_W trace\left(\frac{2rW^TS_BW}{W^TS_W^{sild}W} - I + \frac{2rW^TS_WW}{W^TS_W^{sild}W}\right) = \arg\max_W trace\left(\frac{W^TS_BW}{W^TS_W^{sild}W} + \frac{W^TS_WW}{W^TS_W^{sild}W}\right)$$
$$= \arg\max_W trace\left(\frac{W^T(S_B+S_W)W}{W^TS_W^{sild}W}\right) = \arg\max_W trace\left(\frac{W^TS_TW}{W^TS_W^{sild}W}\right)$$

References

[1]. W.Y. Zhao, R. Chellappa, P.J. Phillips, and A.P. Rosenfeld. Face recognition: A literature survey. ACM Computing Surveys, 35(4):399-458, 2003.

[2]. X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang. Face recognition from a single image per person: A survey. Pattern Recognition, 39(9):1725-1745, 2006.

[3]. Á. Serrano, I.M.d. Diego, C. Conde, and E. Cabello. Recent advances in face biometrics with Gabor wavelets: A review. Pattern Recognition Letters, 31(5):372-381, 2010.

[4]. R. Brunelli and T. Poggio. Face recognition: features versus templates. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(10):1042-1052, 1993.

[5]. M. Turk and A. Pentland. Eigenfaces for Recognition. Journal of Cognitive Neuroscience, 3:71-86, 1991.

[6]. P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7):711-720, 1997.

[7]. B. Moghaddam, T. Jebara, and A. Pentland. Bayesian face recognition. Pattern Recognition, 33(11):1771-1782, 2000.

[8]. L. Wiskott, J.-M. Fellous, N. Krüger, and C.v.d. Malsburg. Face recognition by elastic bunch graph matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7):775-779, 1997.

[9]. C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Transactions on Image Processing, 11(4):467-476, 2002.

[10]. J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma. Robust Face Recognition via Sparse Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2):210-227, 2009.

[11]. F.S. Samaria and A.C. Harter. Parameterisation of a stochastic model for human face identification. in Proceedings of IEEE Workshop on Applications of Computer Vision, 1994.
[12]. A.M. Martinez and R. Benavente. The AR Face Database. in CVC Technical Report #24, 1998.
[13]. T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(12):1615-1618, 2003.
[14]. K. Messer, J. Matas, J. Kittler, J. Lüttin, and G. Maitre. XM2VTSDB: The Extended M2VTS Database. in Second International Conference on Audio and Video-based Biometric Person Authentication, 1999.
[15]. P.J. Phillips, H. Moon, S.A. Rizvi, and P.J. Rauss. The FERET evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(10):1090-1104, 2000.
[16]. P.J. Phillips, et al. Overview of the Multiple Biometrics Grand Challenge. in Proceedings of the Third International Conference on Advances in Biometrics, 2009.
[17]. G.B. Huang, Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. in Technical Report, 2007.
[18]. J. Ruiz-del-Solar, R. Verschae, and M. Correa. Recognition of Faces in Unconstrained Environments: A Comparative Study. EURASIP Journal on Advances in Signal Processing, 2009.
[19]. L. Wolf, T. Hassner, and Y. Taigman. Descriptor Based Methods in the Wild. in workshop of European Conference on Computer Vision, 2008.
[20]. C. Sanderson and B.C. Lovell. Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference. in International Conference on Biometrics (ICB), 2009.
[21]. Z. Cao, Q. Yin, X. Tang, and J. Sun. Face Recognition with Learning-based Descriptor. in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.
[22]. N. Kumar, A.C. Berg, P.N. Belhumeur, and S.K. Nayar. Attribute and Simile Classifiers for Face Verification. in IEEE International Conference on Computer Vision, 2009.
[23]. L. Wolf, T. Hassner, and Y. Taigman. Similarity Scores based on Background Samples. in Asian Conference on Computer Vision (ACCV), 2009.
[24]. N. Pinto, J.J. DiCarlo, and D.D. Cox. How far can you get with a modern face recognition test set using only simple features? in IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[25]. N.-S. Vu and A. Caplier. Face Recognition with Patterns of Oriented Edge Magnitudes. in European Conference on Computer Vision, 2010.
[26]. E. Nowak and F. Jurie. Learning Visual Similarity Measures for Comparing Never Seen Objects. in IEEE Conference on Computer Vision and Pattern Recognition, 2007.
[27]. M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? Metric Learning Approaches for Face Identification. in IEEE International Conference on Computer Vision, 2009.
[28]. Y. Taigman, L. Wolf, and T. Hassner. Multiple One-Shots for Utilizing Class Label Information. in The British Machine Vision Conference (BMVC), 2009.
[29]. H.V. Nguyen and L. Bai. Cosine Similarity Metric Learning for Face Verification. in Asian Conference on Computer Vision (ACCV), 2010.
[30]. N. Pinto and D. Cox. Beyond Simple Features: A Large-Scale Feature Search Approach to Unconstrained Face Recognition. in International Conference on Automatic Face and Gesture Recognition (FG), 2011.
[31]. L. Wolf, T. Hassner, and Y. Taigman. The One-Shot similarity kernel. in International Conference on Computer Vision, 2009.

[32]. D.L. Swets and J.J. Weng. Using discriminant eigenfeatures for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(8):831-836, 1996.
[33]. E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell. Distance Metric Learning, with Application to Clustering with Side-information. in Advances in Neural Information Processing Systems 15, 2002.
[34]. Z. Li, D. Lin, and X. Tang. Nonparametric Discriminant Analysis for Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(4):755-761, 2009.
[35]. P.J. Phillips, et al. Overview of the face recognition grand challenge. in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.
[36]. LFW : Results. 2010; Available from: http://vis-www.cs.umass.edu/lfw/results.html.
[37]. T. Ahonen, A. Hadid, and M. Pietikäinen. Face Recognition with Local Binary Patterns. in European Conference on Computer Vision, 2004.
[38]. H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang. Trace Ratio vs. Ratio Trace for Dimensionality Reduction. in IEEE Conference on Computer Vision and Pattern Recognition. pp. 1-8, 2007.