Optik 125 (2014) 2170–2174

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

Semi-supervised Discriminant Analysis and Sparse Representation-based self-training for Face Recognition

Jiang Jiang, Haitao Gan, Liangwei Jiang, Changxin Gao, Nong Sang*

National key laboratory of science and technology on multi-spectral information processing, School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China

A R T I C L E I N F O

Article history:
Received 12 May 2013
Accepted 11 October 2013

Keywords:
Face recognition
Semi-supervised learning
Semi-supervised discriminant analysis
Sparse representation
Self-training

A B S T R A C T

In this paper, we consider the problem of automatic face recognition with limited manually labeled training data. We propose a new semi-supervised self-training approach which is used to automatically augment the manually labeled training set with new unlabeled data. Semi-supervised Discriminant Analysis is used in each iteration of self-training for discriminative dimensionality reduction by making use of both labeled and unlabeled training data. Sparse representation is applied for classification. Experimental results on four independent databases show that our algorithm outperforms other face recognition methods under 3 different configurations, namely transductive, semi-supervised and single training image.

1. Introduction

For the remarkable face recognition ability of human visual system and practical application needs, automatic face recognition has been one of the most active research topics in computer vision, machine learning and biometrics in decades. The aim of automatic face recognition is to verify the identity of a person from digital images or a video source automatically.

High dimensional data is always a difficulty when we deal with face recognition problems. It has been suspected that since human faces have similar overall configuration, naturally generated high dimensional face data probably resides on a lower dimensional manifold. This leads to dimensionality reduction methods such as Principle Component Analysis (PCA) [1], Linear Discriminant Analysis (LDA) [2] which are known in face recognition literatures as Eigenface and Fisherface. Eigenface is an unsupervised method that performs dimensionality reduction by projecting the original n-dimensional data onto the $d(\ll n)$-dimensional subspace spanned by the leading eigenvectors of the data's covariance matrix. Fisherface is a supervised method that searches for the project directions on which the data points of different classes are far from each other while on the same time requiring data points of the same class to be close to each other.

In recent years, theories of Sparse Representation (SR) and Compressed Sensing (CS) have emerged as powerful tools for classification problems. Wright et al. [3] proposed a Sparse Representation-based Classification (SRC) algorithm, in which the training face images are the dictionary and an incoming test image is classified by finding its sparse representation with respect to this dictionary. This work was extended later to handle misalignment variation [4]. Nagesh and Li [5] have presented an expression-invariant face recognition method based on distributed compressed sensing and joint sparsity models.

The above automatic face recognition methods are suffered in practical applications because they use only a small number of manually labeled training data to learn their template or model. In these cases, inter-session variation, such as illumination, occlusion, pose and expression, which exists commonly in real face images is not well represented in the template or model. There are two ways to solve this difficulty. One is to synthesizing virtual sample to enlarge labeled data. Although much success has been achieved by synthetic sample techniques, such artificial process always has trouble in capturing the real face data distribution due to the variations of pose, illumination and facial expression [6]. The other more natural way to deal with this problem is semi-supervised learning (SSL). A large amount of unlabeled data is generally easily obtained since its collection does not demand costly manual labeling. Those unlabeled images acquired during general operation or testing could then be used to enhance the template or model via self-training approaches. By iteratively enlarging the training set with more and more images, inter-session variation is more likely to be incorporated into the template or model and thus better performance can be expected. Researchers have developed several SSL methods, such as SDA [7], SSLDA [8], lapLDA [9], SELF [10], etc.,

* Corresponding author. Tel. +86 13886014687.
E-mail address: nsang@hust.edu.cn (N. Sang).

0030-4026/$ – see front matter © 2013 Elsevier GmbH. All rights reserved.
http://dx.doi.org/10.1016/j.ijleo.2013.10.043

and reported that the semi-supervised extensions can generally outperform their supervised counterparts like linear discriminant analysis (LDA).

The most relevant work to our proposed approach is by Roli and Marcialis [11] and Zhao et al. [12]. Roli and Marcialis [11] proposed a semi-supervised face recognition algorithm based on PCA. First, a weakly trained PCA-based classifier is initialized with a small number of manually labeled examples, and then it is used to iteratively classify unlabeled training data to augment the labeled training set. However, PCA is an unsupervised dimensionality reduction approach. The self-training approach is only beneficial for the updated templates of each subject rather than the projection itself. In order to embrace the discriminant power of LDA, Zhao et al. [12] proposed an LDA-based self-training algorithm for semi-supervised face recognition. However, over-fitting can occur when the training data is limited and in this case performance can be drastically reduced [13].

In this paper, we propose a new semi-supervised face recognition approach based on SDA, self-training and sparse coding. Compared to the work in [11] and [12], we use SDA in each iteration of self-training to make use of the discriminant structure inferred from the labeled data, as well as the intrinsic geometric structure provided by both labeled and unlabeled data [7]. In this way, the over-fitting problem of [12] is avoided. The ability of sparse representation for face recognition has already been addressed in [3], here we combine it into our scheme so that the labeled data could be used as natural dictionary for classification.

The rest of the paper is organized as follows. In Section 2, we briefly introduce SDA and sparse representation for face recognition and propose our algorithm. The experimental results are presented in Section 3. Finally, we conclude the paper in Section 4.

2. Sparse representation-based SDA self-training algorithm

In this section, we introduce our SDA and sparse representation-based self-training algorithm.

2.1. Semi-supervised discriminant analysis

Linear Discriminant Analysis (LDA) [14] searches for subspaces where the data points from the same class are close and those from different classes are far. However, LDA only uses the training labeled data to decide the optimal projections that best preserve discriminative information. In order to harness the large amount of unlabeled data, Semi-supervised Discriminant Analysis (SDA) [7] extends LDA to incorporate the manifold structure instanced by both labeled and unlabeled data. Therefore, the subspace constructed by SDA preserves not only the discriminative information but also the geometric structure introduced by all the data available. Given a set of data $X = [x_1, x_2, \ldots, x_p] \in \mathbb{R}^{m\times p}$ including both labeled and unlabeled samples, the objective function of SDA is defined as follows [7]:

$$v_{opt} = \max_{v} \frac{v^T S_B v}{v^T (S_T + \alpha X L X^T + \beta I) v}, \tag{1}$$

$$S_B = \sum_{i=1}^{c} l_i (\mu_i - \mu)(\mu_i - \mu)^T, \tag{2}$$

$$S_T = \sum_{j=1}^{l} (x_j - \mu)(x_j - \mu)^T, \tag{3}$$

where S_B and S_T are the inter-class and total scatter matrix respectively and are calculated with labeled training data samples, c is the number of classes, l_i is the number of labeled samples in class i, μ_i is the mean of class i, μ is the global mean of all the labeled samples of all classes, I is the identity matrix, α and β are parameters controlling the balance among three terms in the denominator, and L is the Laplacian matrix defined as $L = D - W$. W is the weight matrix and D is a diagonal matrix which are defined by:

$$W_{ij} = \begin{cases} 1 & \text{if } x_i \in N_k(x_j) \text{ or } x_j \in N_k(x_i) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$D_{ii} = \sum_j W_{ij}, \tag{5}$$

where $N_k(x_i)$ denotes the set of k nearest neighbors of x_i. The solution to the problem in Eq.(1) can be given by the solution to the generalized eigenvalue solution problem:

$$S_B \mathbf{v} = \lambda (S_T + \alpha X L X^T + \beta I)\mathbf{v}. \tag{6}$$

2.2. Sparse coding representation for face recognition

Sparse representation (SR) considers parsimony as its guiding principle for coding high dimensional data [3]. A commonly used measure of parsimony is the number of selected atoms to represent the signal. Specifically, given a dictionary D composed of a set of atoms $D = [d_1, d_2, \ldots, d_n] \in \mathbb{R}^{m\times n}$ and any inputting data sample $\mathbf{y} \in \mathbb{R}^m$, SR tries to find a representation $\alpha \in \mathbb{R}^n$ of y by optimizing the following problem:

$$(l^0): \quad \hat{\alpha} = arg \min_{\alpha} \|\alpha\|_0 \quad \text{subject to} \quad D\alpha = y, \tag{7}$$

where $\|\cdot\|_0$ denotes l^0 norm, which counts the number of nonzero entries in a vector. However, the l^0 problem is NP-hard and the computational complexity of finding the sparsest solution is the same as exhausting all subsets of all D. Fortunately, under the condition that the solution of α is sparse enough, the solution of l^0-minimization problem (7) is equal to the following l^1-minimization problem [15]:

$$(l^1): \quad \hat{\alpha} = arg \min_{\alpha} \|\alpha\|_1 \quad \text{subject to} \quad D\alpha = y, \tag{8}$$

where $\|\cdot\|_1$ denotes the l^1 norm. In real world, since data samples are always noisy, it may not be possible to exactly express y by the superposition of atoms. The model (8) can be further relaxed to a noise allowance form as following:

$$\hat{\alpha} = arg \min_{\alpha} \|\alpha\|_1 \quad \text{subjectto} \quad \|D\alpha - y\|_2 \le \epsilon, \tag{9}$$

where ϵ is the bounded energy of noise. This convex optimization problem is known as Lasso [16] and Basis Pursuit [17] in literature and can be solved efficiently. Sparse representation-based classification (SRC) has been successfully applied in face recognition tasks [3]. For a new test sample y, SRC first computes the representation α for y with dictionary formed by all the labeled training data samples $X = [x_1, x_2, \ldots, x_n]$ according to Eq. (9), and then classify y based on how well the training samples of each class along with their coefficients reproduce y. Algorithm 1 below describes the recognition details of SRC.

Algorithm 1. Sparse Representation-based Classification (SRC)

Input: a matrix of labeled training samples $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{m\times n}$ for c classes, a test sample $y \in \mathbb{R}^m$, and an error tolerance $\epsilon > 0$.

1: Normalize the columns of A to have unit l^2-norm.
2: Solve the l^1-minimization problem:

$$\hat{\alpha}_1 = arg \min_{\alpha} \|\alpha\|_1 \quad \text{subject to} \quad \|X\alpha - y\|_2 \le \epsilon. \tag{10}$$