Pattern Recognition 47 (2014) 1835–1845

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Bilinear discriminative dictionary learning for face recognition

Hui-Dong Liu [a], Ming Yang [a,*], Yang Gao [b], Yilong Yin [c], Liang Chen [a]

[a] School of Computer Science and Technology, Nanjing Normal University, Nanjing 210023, PR China
[b] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, PR China
[c] School of Computer Science and Technology, Shandong University, Jinan 250101, PR China

ARTICLE INFO

Article history:
Received 30 March 2013
Accepted 3 November 2013
Available online 15 November 2013

Keywords:
Bilinear
Dictionary learning
Face recognition
l_2-Norm regularization

ABSTRACT

This work presents a novel dictionary learning method based on the l_2-norm regularization to learn a dictionary more suitable for face recognition. By optimizing the reconstruction error for each class using the dictionary atoms associated with that class, we learn a structured dictionary which is able to make the reconstruction error for each class more discriminative for classification. Moreover, to make the coding coefficients of samples coded over the learned dictionary discriminative, a discriminative term bilinear to the training samples and the coding coefficients is incorporated in our dictionary learning model. The bilinear discriminative term essentially resolves a linear regression problem for patterns concatenated by the training samples and the coding coefficients in the Reproducing Kernel Hilbert Space (RKHS). Consequently, a novel classifier based on the bilinear discriminative model is also proposed. Experimental results on the AR, CMU PIE, CAS-PEAL-R1, and the Sheffield (previously UMIST) face databases show that the proposed method is effective to expression, lighting, and pose variations in face recognition as well as gender classification, compared with the recently proposed face recognition methods and dictionary learning methods.

1. Introduction

Face recognition is an interesting and popular issue in the area of pattern recognition, computer vision, etc. In recent years, owing to the rise of Sparse Representation (Coding) [1,2] and the progress of optimizing the l_1-norm based minimization problems [3–8], there has been an abundant literature on sparse representation based image classification [1–14], and encouraging results have been achieved by sparse representation based classification methods.

Sparse representation assumes that a test sample could be the linear combination of a few atoms in a dictionary. Such a dictionary could be composed of either the original training samples or some redundant bases learned from the training samples. However, the original training samples could contain outliers or noises, and may only reveal the weak discriminative structure among classes. Therefore, dictionary learning techniques are proposed to learn a set of dictionary atoms from the given samples, so that the learned atoms could better reconstruct the test samples.

Some dictionary learning methods have been proposed to fulfill this task. In Ref. [15], Aharon et al. proposed the KSVD algorithm to learn an overcomplete dictionary from the training samples by imposing the l_0-norm constraint on the coding coefficients without any supervised information. Therefore, the dictionary learned by KSVD can not be used for classification. Based on Ref. [15], Aharon et al. [16] and Mairal et al. [17] added different supervised information to make the learned dictionary discriminative for different classes. Specifically, in Ref. [16], a linear regression term for the coding coefficients of the labeled data was incorporated in the dictionary learning model. Different from Ref. [16], Mairal et al. [17] added a term which aimed to make the reconstruction errors for different classes discriminative. Later, Mairal et al. [18] proposed the supervised dictionary learning (SDL), where the coefficients of the samples coded over the learned dictionary are discriminative for classification tasks. Zhang et al. [19] proposed the Discriminative KSVD (DKSVD) for dictionary learning. DKSVD was able to simultaneously learn the dictionary, coding coefficients and a linear classifier for the coding coefficients in one model applying the KSVD algorithm. In Ref. [20], Yang et al. proposed a method to learn a dictionary for each separate class. Ramirez et al. [21] imposed the incoherence on the dictionaries associated with different classes to make them as orthogonal as possible. Based on KSVD, Jiang et al. [22] proposed the LC-KSVD1 by adding a label consistent term to KSVD, to associate the label information with the dictionary items and subsequently the

* Correspondence to: No. 1, Wenyuan Road, Qixia District, Nanjing, Jiangsu 210023, China. Tel.: +86 25 85891987; fax: +86 25 85891685.
E-mail address: m.yang@njnu.edu.cn (M. Yang).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.11.007

LC-KSVD2, which further added a classification error term to LC-KSVD1. Yang et al. [23] applied the Fisher discrimination criterion to the dictionary atoms to learn a structured dictionary, and meanwhile, the Fisher discrimination criterion was also imposed to the coding coefficients to maximize the between-class scatter and minimize the within-class scatter. In Ref. [24], a dictionary was learned so that coding coefficients of the training samples could be fed to different tasks, such as semi-supervised learning, regression, binary classification, multiclass classification, etc. However, methods proposed in Ref. [16–19,22,24] only tried in different ways to exploit the discriminability of the coding coefficients to learn a dictionary, but neglected to make the learned dictionary discriminative for each class. Conversely, methods in Ref. [17,20,21] only focused on minimizing the reconstruction error for each separate class to make the dictionary discriminative while they did not explore the discriminative information hidden in the coding coefficients. Although the method proposed in Ref. [23] applied the Fisher discrimination criterion to make both the dictionary and the coding coefficients discriminative, the Fisher discrimination criterion itself could enforce strong sparsity on the coding coefficients even if no sparse regularization term is imposed. Besides, all the dictionary learning methods mentioned above impose either the l_0-norm or l_1-norm regularization or force strong sparsity on the coding coefficients to learn a dictionary.

Nevertheless, recent research suggests that face recognition is not a sparse representation or compressive sensing problem [25], but a collaborative representation problem [26,27]. Because the face recognition problem is the small sample size (SSS) problem [28], given a high dimensional query face image y, typically, the samples sharing the same label with y in the dictionary are inadequate to represent y, and samples from other classes could also have contributions to reconstruct y. Refs. [26,27] show that by replacing the l_1-norm regularization by the l_2-norm regularization, collaborative representation based classification (CRC) [26,27] achieves competitive results with sparse representation based classification (SRC) [1] but needs much less computational cost.

In this paper, we propose a novel dictionary learning method named bilinear discriminative dictionary learning (BDDL) for face recognition. In our dictionary learning model, we not only make the dictionary atoms associated with each class discriminative, but also exploit the discriminability hidden in the coding coefficients. Since it has been suggested that the l_2-norm regularization is good for coding face images, in our dictionary learning method, we impose the l_2-norm regularization on the coding coefficients. Next, we highlight some characteristics of our approach below:

(1) The l_2-norm regularization is adopted in our dictionary learning method, which not only makes the learned dictionary more suitable for face recognition, but also avoids the high computational cost compared to imposing l_1-norm regularization on the coding coefficients.
(2) To make the dictionary discriminative for each class, we minimize the reconstruction error for each class using the dictionary atoms associated with that class. In this way, a structured dictionary is learned, and the learned dictionary is able to make the reconstruction error for each class more discriminative for classification.
(3) To exploit the discriminative information hidden in the coding coefficients, we make use of both the original training samples and the coding coefficients in our discriminative term, because we hope that the discriminability of original training samples and the discriminability of coding coefficients could complement each other. Thus, we construct a new pattern by concatenating a training sample and a coding coefficient, and model a linear regression problem for the new patterns in the Reproducing Kernel Hilbert Space (RKHS) in our dictionary learning method. The final expression of this discriminative term is a bilinear regression term to the training samples and the coding coefficients.
(4) From the mechanism of the proposed BDDL, we define the bilinear discriminative features (BDF) of a sample. The BDF of a sample is obtained by the bilinear transformation of all the training samples and the coding coefficient of this sample. Along with the BDFs, we propose a novel classifier: the bilinear discriminative classifier (BDC) to classify the samples using the BDFs, and the confidence of the proposed BDC is evaluated based on the Kullback–Leibler divergence (K–L divergence) [29].

The rest of this paper is organized as follows: Section 2 introduces the proposed bilinear discriminative dictionary learning (BDDL) method. In Section 3, we provide the optimization strategy for BDDL. Section 4 introduces some classification approaches suitable for the dictionary learned by BDDL. Experimental results are presented in Section 5. Finally we provide some concluding remarks and future works in Section 6.

2. Bilinear discriminative dictionary learning (BDDL)

In this section, we propose a novel dictionary learning method: Bilinear Discriminative Dictionary Learning (BDDL). We first give some notations below:

Given the training samples of the i-th class $X_i=[x_1,\dots,x_{n_i}]\in R^{m\times n_i}$, where $x_j\in R^m$, $j=1,\dots,n_i$. Suppose we have c class, and let $X=[X_1,\dots,X_c]\in R^{m\times n}$ be the training data matrix, where $n=n_1+\cdots+n_c$. Dictionary learning aims to learn a dictionary $D=[D_1,\dots,D_c]\in R^{m\times k}$, where $D_i\in R^{m\times k_i}$ is the sub-dictionary associated with the i-th class X_i, and $k=k_1+\cdots+k_c$. Let $A_i=[A_i^1;\dots;A_i^j;\dots;A_i^c]\in R^{k\times n_i}$ denote the coefficients of X_i coded over the dictionary D, where $A_i^j\in R^{k_j\times n_i}$ is the j-th sub-coefficients of A_i, denoting the coding coefficients of X_i coded over the sub-dictionary D_j.

The framework of our BDDL model is expressed as follows:

$$\min_{(D,A,W)} f(X,D,A)+\alpha t(L,X,W,A)+\beta\Omega(A) \tag{1}$$

The model comprises three parts: the reconstruction error term $f(X,D,A)$, the bilinear discriminative term $t(L,X,W,A)$, and the regularization term $\Omega(A)$. Next, we shall present the details of the three terms.

2.1. The reconstruction error term $f(X,D,A)$ and the regularization term $\Omega(A)$

For $f(X,D,A)$, we adopt the fidelity term defined in Ref. [23]. $f(X,D,A)$ in our BDDL is designed as follows:

$$f(X,D,A)=||X-DA||_F^2+\sum_{i=1}^{c}||X_i-D_iA_i^i||_F^2 \tag{2}$$

where $||\cdot||_F$ denotes the Frobenious norm of a matrix. In $f(X,D,A)$, the first term $||X-DA||_F^2$ signifies that the training samples X can be well represented by the dictionary D with the coding coefficients A. The second term $\sum_{i=1}^{c}||X_i-D_iA_i^i||_F^2$ ensures that training samples X_i are represented by the i-th sub-dictionary D_i, which is composed of the dictionary atoms from class i. In $f(X,D,A)$, we drop the last term $\sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c}||D_jA_i^j||_F^2$ in the fidelity term used in Ref. [23], because it has been suggested that face recognition is not a compressive sensing problem [25], but a collaborative representation problem