Image and Vision Computing 51 (2016) 13–21

Contents lists available at ScienceDirect

Image and Vision Computing

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

Orthonormal dictionary learning and its application to face recognition☆

Zhen Dong, Mingtao Pei *, Yunde Jia

Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing, 100081, PR China

ARTICLE INFO

Article history:
Received 27 August 2015
Received in revised form 29 January 2016
Accepted 26 March 2016
Available online 8 April 2016

Keywords:
Orthonormal dictionary learning
Low-rank representation
Face recognition

ABSTRACT

This paper presents an orthonormal dictionary learning method for low-rank representation. The orthonormal property encourages the dictionary atoms to be as dissimilar as possible, which is beneficial for reducing the ambiguities of representations and computation cost. To make the dictionary more discriminative, we enhance the ability of the class-specific dictionary to well represent samples from the associated class and suppress the ability of representing samples from other classes, and also enforce the representations that have small within-class scatter and big between-class scatter. The learned orthonormal dictionary is used to obtain low-rank representations with fast computation. The performances of face recognition demonstrate the effectiveness and efficiency of the method.

1. Introduction

Dictionary learning has received attentions in recent years owning to its wide range of applications, such as face recognition [1,2,3], visual tracking [4,5], image classification [3,6,7], and person re-identification [8,9]. The existing methods can be roughly divided into two categories: unsupervised [10,11] and supervised [2,12,13,14]. Dictionaries learned by supervised methods are more discriminative and obtain better performances in classification tasks. But most supervised methods failed to take the relationship between dictionary atoms into account, which might yield similar atoms. A redundant dictionary with lots of similar atoms will result in the ambiguity of representations and high computation cost, and Fig. 1(a) shows the representation ambiguity problem brought by similar dictionary atoms.

In order to alleviate these problems, we present an orthonormal dictionary learning method. The dictionary is enforced to be orthonormal to eliminate similar atoms. Having a compact dictionary, both the ambiguities of representations and computation cost can be reduced. Our work is similar to [13] which achieved good performances on both classification and clustering tasks. Different from promoting incoherence between class-specific dictionaries in their work, our method endows the whole dictionary with the orthonormal property, which implies that all the atoms are as independent as possible whether they are in the same class or not. Fig. 1(b) shows the effect of the orthonormal property of the dictionary. Since the dimensionality of the feature vector is usually much larger than the number of dictionary atoms, forcing orthonormal constraint on the dictionary is feasible.

To get a more discriminative dictionary, we push the class-specific dictionary to have good ability to well represent samples from the associated class and also suppress the ability of representing samples from other classes. [13,14,15,16] explicitly modeled the idea and optimized the class-specific dictionaries one by one, which may neglect the relationship between class-specific dictionaries during the optimization procedure. Different from these work, we propose a concise discriminative regularization term only restricting on the representation. In our method, all the class-specific dictionaries are obtained simultaneously, and the relationship between them can well hold. Moreover, we encourage the representations to have small within-class scatter and big between-class scatter, which is implemented by using the Fisher discriminative criterion. The discriminative power of the representation is able to propagate to the dictionary since they are coordinately solved in a unified optimization framework.

Using the learned orthonormal dictionary, we are able to acquire low-rank representations with fast computation. Shu et al. [17] proved that the rank of the reconstruction data matrix is upper bounded by the number of non-zero rows of the representation matrix when the dictionary is orthonormal. According to this theorem, the nuclear norm in the low-rank representation can be replaced by its upper bound, and the representations can be obtained by the magnitude shrinkage function instead of the singular value shrinkage function which needs the SVD operation [18]. The solution procedure is much faster than solving the traditional low-rank representation problem [19,20].

As an application of the proposed method, face recognition from both images and videos are conducted. In still image datasets, faces from the same individual vary slightly and are highly correlated. They

☆ This paper has been recommended for acceptance by Xiaogang Wang.

* Corresponding author.
E-mail addresses: dongzhen@bit.edu.cn (Z. Dong), peimt@bit.edu.cn (M. Pei), jiayunde@bit.edu.cn (Y. Jia).

http://dx.doi.org/10.1016/j.imavis.2016.03.010
0262-8856/

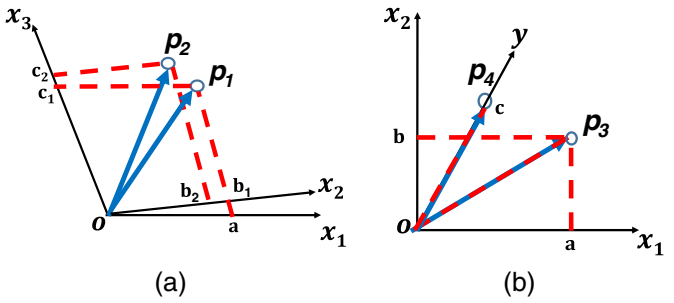

Fig. 1. The effect of the dictionary's orthonormal property. In (a), dictionary atoms $\mathbf{x}_1$ and $\mathbf{x}_2$ are similar, which brings the representation ambiguity problem. The sample $\mathbf{p}_1$ may have two representations, $(a,0,c_1)$ or $(0,b_1,c_1)$. More seriously, similar samples $\mathbf{p}_1$ and $\mathbf{p}_2$ may have quite different representations, $(a,0,c_1)$ and $(0,b_2,c_2)$. In (b), the orthonormal constraint is taken into account, $\mathbf{x}_1$ and $\mathbf{x}_2$ are dictionary atoms that belong to the same class, $\mathbf{y}$ is the atom of another class, and each pair of them are orthonormal. Samples $\mathbf{p}_3$ and $\mathbf{p}_4$ come from different classes, and their representations, $(a,b,0)$ and $(0,0,c)$, have strong discriminative power.

are assumed to lie in a linear subspace, so the reconstruction data of the faces are expected to be low-rank. What's more, the orthonormal dictionary is beneficial for reducing the ambiguities of the representations, and thus able to represent all the faces in the dataset faithfully and discriminatively. The results of various experiments on Extended Yale B and AR datasets demonstrate the effectiveness and efficiency of the proposed method in recognizing faces from still images. Compared with face image recognition, the task of face recognition from videos is more difficult since the videos are acquired under non-ideal conditions where illuminations, poses, or expressions vary large. The proposed dictionary learning method aims to learn a discriminative and compact dictionary which is quite robust to the variations in videos. We conduct experiments on Honda/UCSD and YouTube Celebrities datasets and achieve comparable results with the state-of-the-art methods.

The main contributions of this paper are as follows:

- We present an orthonormal dictionary learning method which can learn a discriminative and compact dictionary for low-rank representation.
- The learned orthonormal dictionary is used to obtain low-rank representations with fast computation and comparable performances to the state-of-the-art methods.
- The proposed method achieves excellent performance on the application of face recognition from both still image and video datasets.

The rest of this paper is organized as follows. We elaborate the formulation of the orthonormal dictionary learning method in Section 2, and detail the optimization and the initialization in Section 3. In Section 4, the fast low-rank representation method is presented. The experiments and discussions are in Section 5, and Section 6 concludes this paper.

2. Orthonormal dictionary learning

Given the observation sample matrix $\mathbf{Y}\in\mathbb{R}^{D\times N}$, our goal is to learn the dictionary $\mathbf{D}\in\mathbb{R}^{D\times K}$, the representation matrix $\mathbf{X}\in\mathbb{R}^{K\times N}$, and the noise matrix $\mathbf{E}\in\mathbb{R}^{D\times N}$ to satisfy $\mathbf{Y}=\mathbf{DX}+\mathbf{E}$. A redundant dictionary have lots of similar atoms, which results in high computation cost and ambiguity in corresponding representations. In order to alleviate these problems, we introduce the orthonormal constraint of the dictionary into our model. The orthonormal dictionary learning is given by

$$\begin{aligned}&\min_{\mathbf{D},\mathbf{X},\mathbf{E}} \mathcal{LR}(\mathbf{D},\mathbf{X})+\mathcal{SN}(\mathbf{E})+\mathcal{DC}(\mathbf{X},\mathbf{L})\\ &s.t.\mathbf{Y}=\mathbf{DX}+\mathbf{E},\mathbf{D}^T\mathbf{D}=\mathbf{I},\end{aligned}\tag{1}$$

where the function $\mathcal{LR}(\cdot)$ measures the low-rankness of the reconstruction data matrix **DX**, the $\mathcal{SN}(\cdot)$ measures the sparsity of the noise matrix **E**, and the $\mathcal{DC}(\cdot)$ is the discriminative term for improving the discriminative power of the learned dictionary. The **L** is the label matrix of the samples **Y**, and the **I** is the identity matrix.

2.1. Low-rank term $\mathcal{LR}(\boldsymbol{D},\boldsymbol{X})$

In classification task, training samples from the same class are highly correlated and expected to form a low-dimensionality subspace, so the training samples without noises, *i.e.* the reconstruction data matrix represented by **DX**, should be low-rank. We formulate the low-rank term as

$$rank(\mathbf{DX})=rank(\mathbf{Z}),\tag{2}$$

where $rank(\mathbf{Z})$ denotes the rank of the matrix **Z**. The minimization of Eq. (2) is an NP-hard problem and difficult to solve due to the discrete nature of the $rank(\cdot)$ function. Fortunately, Fazel [21] proved that the nuclear norm function $\|\mathbf{Z}\|_*$ (*i.e.* the sum of the singular values of **Z**) is the convex envelope of the rank function $rank(\mathbf{Z})$ on the set of $\{\mathbf{Z}|\|\mathbf{Z}\|_2<1\}$, and Candès et al. [22,23] proposed to minimize the nuclear norm function instead of the rank function. Accordingly, our low-rank term can be rewritten as $\|\mathbf{DX}\|_*$. As demonstrated in [17], $\|\mathbf{DX}\|_*$ is upper bounded by $\|\mathbf{X}^T\|_{2,1}=\sum_{i=1}^{K}\|\mathbf{x}_i\|_2=\sum_{i=1}^{K}\sqrt{\sum_{j=1}^{N}\mathbf{x}_{ij}^2}$ under the constraint of $\mathbf{D}^T\mathbf{D}=\mathbf{I}$, where $\mathbf{x}_i$ represents the i-th row of **X**, and $\mathbf{x}_{ij}$ denotes the element in the i-th row and j-th column of **X**. Our low-rank term is finally given by

$$\mathcal{LR}(\mathbf{D},\mathbf{X})=\|\mathbf{X}^T\|_{2,1}.\tag{3}$$

Furthermore, the 2,1-norm can be interpreted as the group sparsity of face representations, which is demonstrated to be quite effective on the face recognition with complex variances in [24]. Eq. (3) can be minimized efficiently as described in Sec.3.1.

2.2. Sparse noise term $\mathcal{SN}(\boldsymbol{E})$

Real-word data are often noisy or corrupted due to illumination variation, occlusion, and pixel corruption. The classifier trained with these data may overfit and the classification performance may degrade. Motivated by the low-rank recovery [17,22], the corrupted data matrix **Y** is decomposed into two parts: a low-rank component **DX** and a sparse noise component **E** to alleviate the problem. Here, we denote the noisy term as

$$\mathcal{SN}(\mathbf{E})=\|\mathbf{E}\|_{2,1}.\tag{4}$$

The $\|\mathbf{E}\|_{2,1}$ is used since it encourages the sum of $l2$-norm of all columns in **E** to be zero, which reflects the assumption that some training samples are corrupted and the others are not. In addition, $\mathbf{E}=\mathbf{Y}-\mathbf{DX}$ measures the reconstruction error so the minimization of $\mathcal{SN}(\mathbf{E})$ encourages the dictionary **D** to well represent the observation data.

2.3. Discriminative term $\mathcal{DC}(\boldsymbol{X},\boldsymbol{L})$

In order to enhance the discriminative power of the dictionary $\boldsymbol{D}$, we propose the discriminative regularization term $\|\mathbf{X}\odot\mathbf{S}\|_F^2$ where $\odot$ means the element-wise multiplication operator, $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. The $\mathbf{S}\in\mathbb{R}^{K\times N}$ is defined as

$$\mathbf{S}(i,j)=\begin{cases}0, & \text{if } \mathbf{d}_i \text{ and } \mathbf{y}_j \text{ belong to the same class}\\ 1, & \text{otherwise},\end{cases}\tag{5}$$