Discriminant Tensor Dictionary Learning with Neighbor Uncorrelation for Image Set Based Classification

Fei Wu[1,2,*], **Xiao-Yuan Jing**[1,2,*], **Wangmeng Zuo**[3], **Ruiping Wang**[1], **Xiaoke Zhu**[1]
[1] State Key Laboratory of Software Engineering, School of Computer, Wuhan University, China
[2] College of Automation, Nanjing University of Posts and Telecommunications, China
[3] School of Computer, Harbin Institute of Technology, China
{wufei_8888, jingxy_2000}@126.com

Abstract

Image set based classification (ISC) has attracted lots of research interest in recent years. Several ISC methods have been developed, and dictionary learning technique based methods obtain state-of-the-art performance. However, existing ISC methods usually transform the image sample of a set into a vector for processing, which breaks the inherent spatial structure of image sample and the set. In this paper, we utilize tensor to model an image set with two spatial modes and one set mode, which can fully explore the intrinsic structure of image set. We propose a novel ISC approach, named discriminant tensor dictionary learning with neighbor uncorrelation (DTDLNU), which jointly learns two spatial dictionaries and one set dictionary. The spatial and set dictionaries are composed by set-specific sub-dictionaries corresponding to the class labels, such that the reconstruction error is discriminative. To obtain dictionaries with favorable discriminative power, DTDLNU designs a neighbor-uncorrelated discriminant tensor dictionary term, which minimizes the within-class scatter of the training sets in the projected tensor space and reduces dictionary correlation among set-specific sub-dictionaries corresponding to neighbor sets from different classes. Experiments on three challenging datasets demonstrate the effectiveness of DTDLNU.

1 Introduction

In recent years, image set based classification (ISC) has attracted lots of research interest in computer vision and pattern classification communities [Zhang *et al.*, 2016]. An image set can convey rich within-class variations of an object, which is helpful for classification. ISC is also a challenging task, and how to effectively model a set and compute the similarity between two sets is a crucial research topic.

Over the past several years, we have witnessed a lot of methods developed for ISC. The subspace-based and manifold-based methods [Wang and Chen, 2009; Wang *et al.*, 2015] separately use subspace and manifold to model an image set, and the performances of them may degrade when the set has a small sample size but big data variations [Hu *et al.*, 2012]. In affine or convex hull based methods [Hu *et al.*, 2012; Cevikalp and Triggs, 2010], the between-set distance is defined as the distance between two closest points of two sets. This kind of methods relies highly on the location of each individual sample in the set, and the model fitting can be heavily deteriorated by outliers [Wang *et al.*, 2012]. Covariance matrix based methods [Wang *et al.*, 2012; Lu *et al.*, 2013] try to explore the second-order statistics of image set and represent each set with its covariance matrix, while they cannot explore intrinsic high-order structure of image set. Deep learning based methods [Hayat *et al.*, 2015; 2014; Shah *et al.*, 2016] introduce an adaptive multi-layer neural network structure and use it for learning class specific models. However, these deep learning based methods require a large amount of computation time.

Recently, a few sparse/collaborative representation and dictionary learning based methods have been developed for ISC and obtain state-of-the-art classification performance [Zhu *et al.*, 2014; Zheng *et al.*, 2017; Chen *et al.*, 2012]. This family of methods usually builds one dictionary for each image set or class, and uses the dictionaries to measure the similarity of image sets.

1.1 Motivation

Almost all the previous works transform the image sample in the set into a vector for subsequent processing, as shown in Fig. 1(a), which not only breaks the inherent spatial structure of image samples but also breaks the structure of the image set. In fact, an image is a data matrix with the size of $d_W \times d_H$. And an image set is a three-dimensional data array with the size of $d_W \times d_H \times d_N$, where d_N denotes the number of images in the set, as shown in Fig. 1(b). How to effectively model the image set without breaking its inherent structure and provide a corresponding similarity measure between sets is a crucial research topic.

Tensor is effective to model an image or image ensembles, and tensor-based learning methods can well retain the spatial structures of image and image ensembles [Li and Schonfeld, 2014]. On the other hand, existing dictionary learning based ISC methods own interesting classification effects [Zhu *et al.*, 2014; Lu *et al.*, 2014]. Inspired by these two aspect-

*Corresponding authors

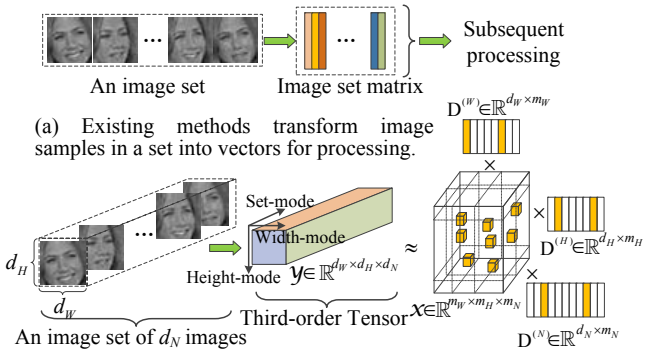

Figure 1: Illustration of the general difference between our work and existing ISC works. $\mathrm{D}^{(W)}$ and $\mathrm{D}^{(H)}$ are spatial dictionaries and $\mathrm{D}^{(N)}$ denotes the set dictionary. $\mathcal{X}$ is the sparse representation coefficient tensor. The constructed third-order tensor can be well represented by linear combination of columns in $\mathrm{D}^{(W)}$, $\mathrm{D}^{(H)}$ and $\mathrm{D}^{(N)}$.

s, we intend to employ the tensor dictionary learning (TDL) technique for ISC.

Nowadays, some TDL methods have been addressed [Peng *et al.*, 2014; Zubair *et al.*, 2014]. Compared with traditional dictionary learning technique [Zhu *et al.*, 2014], **the TDL technique can learn multiple dictionaries, with each corresponding to one mode of training tensors, which can fully exploit the information contained in training samples**. However, for existing TDL methods, there still exists much room for improvement:

(1) Existing TDL methods mainly focus on the reconstruction accuracy, whereas enhancing the total discriminability of tensor dictionaries has not been investigated comprehensively and thoroughly.

(2) Information redundancy among image sets will lead to redundancy in the learned dictionaries. How to effectively reduce the redundancy between tensor dictionaries corresponding to different classes has not yet been well studied.

1.2 Contribution

(1) We introduce the idea of tensor to model image set, which can preserve the structure of image set. And we propose a discriminant tensor dictionary learning with neighbor uncorrelation (DTDLNU) approach, which jointly learns two spatial dictionaries and one set dictionary. **These three dictionaries jointly reflect the spatial structure of image sets.** The spatial dictionaries and the set dictionary are composed by set-specific sub-dictionaries corresponding to the class labels, such that the obtained reconstruction error is discriminative. Fig. 1 illustrates the general difference between our work and existing ISC works.

(2) We design a neighbor-uncorrelated discriminant tensor dictionary term for TDL, which minimizes the within-class scatter of the training sets in the projected tensor space and reduces tensor dictionary correlation among set-specific sub-dictionaries corresponding to neighbor image sets from different classes. This designed term can make the learned dictionaries have favorable discriminative power and low between-class correlation.

2 Brief Review of Related Work

2.1 Image Set based Classification (ISC) Methods

Current ISC methods can be generally categorized into five kinds:

(1) Subspace and manifold based methods. Manifold discriminant analysis (MDA) [Wang and Chen, 2009] aims to learn an embedding space by maximizing manifold margin. The discriminant analysis on Riemannian manifold of Gaussian distributions (DARG) [Wang *et al.*, 2015] method represents image set as Gaussian mixture model comprising a number of Gaussian components and seeks to discriminate Gaussian components from different classes.

(2) Affine/convex hull based methods. Cevikalp and Triggs [2010] presented the affine hull based image set distance (AHISD) and convex hull based image set distance (CHISD) methods. Sparse approximated nearest points (SANP) method [Hu *et al.*, 2012] focuses on nearest points of two image sets, which can be sparsely approximated by the samples of its respective set.

(3) Covariance matrix based methods. Covariance discriminative learning (CDL) [Wang *et al.*, 2012] represents each image set with its covariance matrix and models the ISC problem as classifying points on the Riemannian manifold. Localized multi-kernel metric learning (LMKML) [Lu *et al.*, 2013] regards the out product between the covariance matrix and mean of image set as the third-order statistics, and combines the third-order statistics, second-order statistics (covariance matrix) and first-order statistics (mean vector) information for classification.

(4) Deep learning based methods. The deep reconstruction model with weighted voting (DRM-WV) [Hayat *et al.*, 2015] method designs a multi-layer neural network to learn class-specific deep reconstruction models. With the learned models, DRM-WV uses reconstruction error based weighted voting strategy for classification.

(5) Sparse/collaborative representation and dictionary learning based methods. The image set based collaborative representation and classification (ISCRC) method [Zhu *et al.*, 2014] models the query set as a convex or regularized hull, and represents the hull collaboratively over all the gallery sets for classification. The dictionary-based face recognition from video (DFRV) method [Chen *et al.*, 2012] builds one dictionary for each face image set and uses the learned dictionaries to measure the similarity of face image sets. The simultaneous feature and dictionary learning (SFDL) method [Lu *et al.*, 2014] jointly learns a feature projection matrix and structured dictionary for image set based face recognition.

As analyzed in Introduction, there exist respective shortcomings in these five categories of methods.

2.2 Tensor Dictionary Learning (TDL) Methods

Based on the theory of tensor, nowadays, some TDL methods have been developed [Quan *et al.*, 2015]. Considering the nonlocal similarity over space and the global correlation across spectrum, Peng et al. (2014) designed a decomposable nonlocal TDL method for multispectral image denois-

ing. Roemer et al. [Roemer *et al.*, 2014] presented tensor extensions of the popular MOD and K-SVD dictionary learning algorithms and obtained the tensor-MOD and higher-order K-SVD (K-HOSVD) algorithms. With the region covariance descriptor, Zhang et al. [Zhang *et al.*, 2013] introduced structural incoherence constraint between dictionary atoms from different classes to promote discriminating information into the dictionary.

As analyzed in Motivation, there exists much room for improvement in existing TDL methods. In addition, the TDL technique has not been used to solve the ISC problem.

3 Definitions and Notations

Let $\mathcal{A} \in \mathbb{R}^{I_1 \times \cdots \times I_n \times \cdots \times I_N}$ be a tensor of order N, whose elements are denoted as $a_{i_1 \cdots i_n \cdots i_N}$, $1 \leq i_n \leq I_n$. The Frobenius norm of the tensor $\mathcal{A}$ is defined as $\|\mathcal{A}\|_F = \left(\sum_{i_1,\cdots,i_N} |a_{i_1 \cdots i_N}|^2\right)^{\frac{1}{2}}$.

Definition 1 (Tensor Matricization, Mode-n Product, and Kronecker Product): The mode-n matricization of $\mathcal{A}$ is $\mathcal{A}_{(n)} \in \mathbb{R}^{I_n \times (I_1 \cdots I_{n-1} I_{n+1} \cdots I_N)}$. The mode-$n$ product of $\mathcal{A}$ by a matrix $\mathrm{B} \in \mathbb{R}^{J_n \times I_n}$, denoted by $\mathcal{A} \times_n \mathrm{B}$, is also an N^{th}-order tensor $\mathcal{C} \in \mathbb{R}^{I_1 \times \cdots \times J_n \times \cdots \times I_N}$. The mode-$n$ product $\mathcal{C} = \mathcal{A} \times_n \mathrm{B}$ can also be calculated by $\mathcal{C}_{(n)} = \mathrm{B}\mathcal{A}_{(n)}$. For $\mathcal{C} = \mathcal{A} \times_1 \mathrm{B}_1 \times_2 \mathrm{B}_2 \times_3 \cdots \times_N \mathrm{B}_N$, we can simplify the notation as $\mathcal{C} = \mathcal{A} \prod_{k=1}^{N} \times_k \mathrm{B}_k$. The Kronecker product of matrices $\mathrm{A} \in \mathbb{R}^{I \times J}$ and $\mathrm{B} \in \mathbb{R}^{L \times M}$, denoted by $\mathrm{A} \otimes B$, is a matrix of size $(IL) \times (JM)$. The detailed definitions can be found in [Li and Schonfeld, 2014]

Definition 2 (Block-sparsity): The concept of block sparsity for tensor is presented in [Peng *et al.*, 2014]. For the tensor $\mathcal{A}$, its block-sparsity with respect to N modes is $\|\mathcal{A}\|_B = (r_1, r_2, \cdots, r_N)$ if and only if the smallest index subsets $I_1, I_2, \cdots, I_N$ satisfying $a_{i_1 i_2 \cdots i_N} = 0$ for all $(i_1, i_2, \cdots, i_N) \notin I_1 \times I_2 \times \cdots I_N$ contain $r_1, r_2, \cdots, r_N$ elements, respectively. $Sub(\mathcal{A}) \in \mathbb{R}^{r_1 \times r_2 \times \cdots r_N}$ denotes the intrinsic sub-tensor of $\mathcal{A}$ extracted from the entries of the N dimensions of $\mathcal{A}$ specified by the index sets $I_1, I_2, \cdots, I_N$, respectively.

4 The Model of DTDLNU

4.1 Neighbor-uncorrelated Discriminant Tensor Dictionary Term

Let $\mathrm{Y} = \{\mathrm{Y}_c\}$, $c = 1, \cdots, C$ be the training set containing image sets from C different classes and $\mathrm{Y}_c = \left\{\mathrm{Y}_c^l\right\}_{l=1}^{L_c}$ be the collection of image sets of the c^{th} class, where L_c is the number of image sets in Y_c. Each Y_c^l contains a set of images with the size of $d_W \times d_H$, where d_W and d_H separately denote the spatial width and height of an image. The number of images in Y_c^l is denoted by d_N. Then, each Y_c^l can be expressed as a 3^{rd}-order tensor $\mathcal{Y}_c^l \in \mathbb{R}^{d_W \times d_H \times d_N}$ with two spatial modes and one set mode.

We aim to learn two structured spatial dictionaries $\mathrm{D}^{(W)} = \left[\mathrm{D}_1^{1(W)}, \cdots, \mathrm{D}_1^{L_1(W)}, \cdots, \mathrm{D}_C^{1(W)}, \cdots, \mathrm{D}_C^{L_C(W)}\right] \in \mathbb{R}^{d_W \times m_W}$ and $\mathrm{D}^{(H)} = \left[\mathrm{D}_1^{1(H)}, \cdots, \mathrm{D}_1^{L_1(H)}, \cdots, \mathrm{D}_C^{1(H)}, \cdots, \mathrm{D}_C^{L_C(H)}\right] \in \mathbb{R}^{d_H \times m_H}$ and one structured set dictionary $\mathrm{D}^{(N)} = \left[\mathrm{D}_1^{1(N)}, \cdots, \mathrm{D}_1^{L_1(N)}, \cdots, \mathrm{D}_C^{1(N)}, \cdots, \mathrm{D}_C^{L_C(N)}\right] \in \mathbb{R}^{d_N \times m_N}$ from the total training set, where $\mathrm{D}_c^{l(W)} \in \mathbb{R}^{d_W \times r_c^{l(W)}} (r_c^{l(W)} \leq d_W)$, $\mathrm{D}_c^{l(H)} \in \mathbb{R}^{d_H \times r_c^{l(H)}} (r_c^{l(H)} \leq d_H)$ and $\mathrm{D}_c^{l(N)} \in \mathbb{R}^{d_N \times r_c^{l(N)}} (r_c^{l(N)} \leq d_N)$ are set-specified sub-dictionaries associated with the l^{th} set of the c^{th} class, and $m_W > d_W$, $m_H > d_H$ and $m_N > d_N$. Here, W, H and N separately represent the width, height and set modes of $\mathcal{Y}_c^l$. $r_c^{l(W)}$, $r_c^{l(H)}$ and $r_c^{l(N)}$ denote the numbers of atoms in dictionaries $\mathrm{D}_c^{l(W)}$, $\mathrm{D}_c^{l(H)}$ and $\mathrm{D}_c^{l(N)}$, respectively. And $m_W = \sum_{c=1}^{C} \sum_{l=1}^{L_c} r_c^{l(W)}$, $m_H = \sum_{c=1}^{C} \sum_{l=1}^{L_c} r_c^{l(H)}$, and $m_N = \sum_{c=1}^{C} \sum_{l=1}^{L_c} r_c^{l(N)}$.

The quality of the learned dictionary influences the performance of subsequent tensor sparse representation based classification. **To make the learned dictionaries be discriminative for image sets in Y, we require that the within-class scatter in the projected tensor space should be minimized.** The mode-i within-class scatter matrix in the partially projected tensor subspace (by all tensor modes except for i) can be defined as:

$$\mathrm{S}_w^{(i)} = \sum_{c=1}^{C} \sum_{l=1}^{L_c} \left[\left(\mathcal{Y}_c^l - \mathcal{M}_c\right) \prod_{\substack{j=\{W,H,N\} \\ j \neq i}} \times_j \mathrm{D}^{(j)^T} \right]_{(i)} \left[\left(\mathcal{Y}_c^l - \mathcal{M}_c\right) \prod_{\substack{j=\{W,H,N\} \\ j \neq i}} \times_j \mathrm{D}^{(j)^T} \right]_{(i)}^T \tag{1}$$

where $i = \{W, H, N\}$, $(\cdot)^T$ denotes the transposition operation and $\mathcal{M}_c = \frac{1}{L_c} \sum_{l=1}^{L_c} \mathcal{Y}_c^l$ is the class mean tensor for class c. Therefore, we should minimize the value of

$$f\left(\mathrm{D}^{(W)}, \mathrm{D}^{(H)}, \mathrm{D}^{(N)}\right) = \sum_{i=\{W,H,N\}} tr\left(\mathrm{D}^{(i)^T} \mathrm{S}_w^{(i)} \mathrm{D}^{(i)}\right) \tag{2}$$

Due to the information redundancy between image sets, there exists much redundancy in dictionary of each mode. To reduce the redundancy and further enhance discriminabilities for dictionaries, we require that the set-specific tensor sub-dictionaries of different classes own low correlation. Reducing the correlation between sub-dictionaries of different classes can make that a set should be more likely to be represented by sub-dictionaries of its own class rather than those of the other classes, and thus generally results in improved discriminative power. In real-world ISC applications, the number of training image sets may be very large, and for the observed image set, only a few sets with different class labels are close to it. **Focusing on these inseparable sets, we require that the set-specific sub-dictionaries corresponding to the neighbor image sets that are from different classes should own low tensor dictionary correlation.** Thus, we should minimize

$$u\left(\mathrm{D}^{(W)}, \mathrm{D}^{(H)}, \mathrm{D}^{(N)}\right) = \sum_{i=\{W,H,N\}} \sum_{c=1}^{C} \sum_{l=1}^{L_c} \sum_{\substack{h=1 \\ h \neq c}}^{C} \sum_{q=1}^{L_h} \beta_{cl}^{hq} \left\| \mathrm{D}_c^{l(i)^T} \mathrm{D}_h^{q(i)} \right\|_F^2 \tag{3}$$

where $\beta_{cl}^{hq} = \begin{cases} 1, & if\ \mathcal{Y}_h^q\ is\ one\ of\ the\ k\ nearest\ neighbors\ of\ \mathcal{Y}_c^l \\ 0, & otherwise \end{cases}$.

By minimizing $u(\mathrm{D}^{(W)},\mathrm{D}^{(H)},\mathrm{D}^{(N)})$, we can reduce the local between-class tensor dictionary correlation.

4.2 The Objective Function of DTDLNU

Considering the dictionary reconstruction error and the neighbor-uncorrelated discriminant tensor dictionary term, we formulate the objective function of DTDLNU as follows:

$$\begin{aligned}&\left\langle \mathrm{D}^{(W)},\mathrm{D}^{(H)},\mathrm{D}^{(N)},\mathcal{X}_c^l(c=1,\cdots,C;l=1,\cdots,L_c)\right\rangle\\&=\arg\min_{\mathrm{D}^{(W)},\mathrm{D}^{(H)},\mathrm{D}^{(N)},\mathcal{X}_c^l}\sum_{c=1}^{C}\sum_{l=1}^{L_c}\left\|\mathcal{Y}_c^l-\mathcal{X}_c^l\times_1\mathrm{D}^{(W)}\times_2\mathrm{D}^{(H)}\times_3\mathrm{D}^{(N)}\right\|_F^2\\&+\lambda\{\sum_{i=\{W,H,N\}}tr(\mathrm{D}^{(i)T}\mathrm{S}_w^{(i)}\mathrm{D}^{(i)})+\sum_{i=\{W,H,N\}}\sum_{c=1}^{C}\sum_{l=1}^{L_c}\sum_{\substack{h=1\\h\neq c}}^{C}\sum_{q=1}^{L_h}\beta_{cl}^{hq}\|\mathrm{D}_c^{l(i)T}\mathrm{D}_h^{q(i)}\|_F^2\}\\&s.t.\ \left\|\mathcal{X}_c^l\right\|_B\preceq\left(r_c^{l(W)},r_c^{l(H)},r_c^{l(N)}\right)\end{aligned}\tag{4}$$

where $\mathcal{X}_c^l\in\mathbb{R}^{m_W\times m_H\times m_N}$ is the sparse coding coefficient tensor of $\mathcal{Y}_c^l$ over $\mathrm{D}^{(W)},\mathrm{D}^{(H)}$ and $\mathrm{D}^{(N)}$, $\mathrm{v}_1\preceq\mathrm{v}_2$ denotes that each entry of v_1 is no more than the corresponding entry of v_2. The block-sparsity of $\mathcal{X}_c^l$ guarantees that $\mathcal{X}_c^l$ is only associated with $r_c^{l(W)},r_c^{l(H)},r_c^{l(N)}$ atoms of dictionaries $\mathrm{D}^{(W)},\mathrm{D}^{(H)},\mathrm{D}^{(N)}$, respectively. λ is a balance factor.

Due to the redundancy setting of spatial and set dictionaries, we can assume that each training image set $\mathcal{Y}_c^l$ is only related to the sub-dictionaries: $\mathrm{D}_c^{l(W)}$, $\mathrm{D}_c^{l(H)}$ and $\mathrm{D}_c^{l(N)}$, which is similar to the idea in [Peng *et al.*, 2014]. Then, we can get

$$\begin{aligned}&\left\|\mathcal{Y}_c^l-\mathcal{X}_c^l\times_1\mathrm{D}^{(W)}\times_2\mathrm{D}^{(H)}\times_3\mathrm{D}^{(N)}\right\|_F^2\\&=\left\|\mathcal{Y}_c^l-Sub\left(\mathcal{X}_c^l\right)\times_1\mathrm{D}_c^{l(W)}\times_2\mathrm{D}_c^{l(H)}\times_3\mathrm{D}_c^{l(N)}\right\|_F^2\end{aligned}\tag{5}$$

where $Sub(\mathcal{X}_c^l)\in\mathbb{R}^{r_c^{l(W)}\times r_c^{l(H)}\times r_c^{l(N)}}$ is the intrinsic subtensor of $\mathcal{X}_c^l$.

When updating the variables corresponding to the l^{th} image set from the c^{th} class, the variables corresponding to other image sets are supposed to be fixed. To update $\mathcal{X}_c^l$, $\mathrm{D}_c^{l(W)}$, $\mathrm{D}_c^{l(H)}$ and $\mathrm{D}_c^{l(N)}$, (4) can be reduced to:

$$\begin{aligned}&\left\langle\mathcal{G}_c^l,\mathrm{D}_c^{l(W)},\mathrm{D}_c^{l(H)},\mathrm{D}_c^{l(N)}\right\rangle\\&=\arg\min_{\mathcal{G}_c^l,\mathrm{D}_c^{l(W)},\mathrm{D}_c^{l(H)},\mathrm{D}_c^{l(N)}}\left\|\mathcal{Y}_c^l-\mathcal{G}_c^l\times_1\mathrm{D}_c^{l(W)}\times_2\mathrm{D}_c^{l(H)}\times_3\mathrm{D}_c^{l(N)}\right\|_F^2\\&+\lambda\{\sum_{i=\{W,H,N\}}tr(\mathrm{D}^{l(i)T}\mathrm{S}_w^{(i)}\mathrm{D}^{l(i)})+\sum_{i=\{W,H,N\}}\sum_{\substack{h=1\\h\neq c}}^{C}\sum_{q=1}^{L_h}\beta_{cl}^{hq}\|\mathrm{D}_c^{l(i)T}\mathrm{D}_h^{q(i)}\|_F^2\}\end{aligned}\tag{6}$$

where $\mathcal{G}_c^l=Sub(\mathcal{X}_c^l)\in\mathbb{R}^{r_c^{l(W)}\times r_c^{l(H)}\times r_c^{l(N)}}$ is the representation coefficient tensor of $\mathcal{Y}_c^l$ over $\mathrm{D}_c^{l(W)}$, $\mathrm{D}_c^{l(H)}$, and $\mathrm{D}_c^{l(N)}$.

5 Optimization of DTDLNU

There is no theoretical guarantee that the objective function (6) is jointly convex to $(\mathcal{G}_c^l,\mathrm{D}_c^{l(W)},\mathrm{D}_c^{l(H)},\mathrm{D}_c^{l(N)})$; however, it is convex with respect to each of $\mathcal{G}_c^l$, $\mathrm{D}_c^{l(W)}$, $\mathrm{D}_c^{l(H)}$, and $\mathrm{D}_c^{l(N)}$ when the others are fixed. We develop an iterative algorithm to optimize the variables alternatively.

When updating $\mathcal{G}_c^l$, $\mathrm{D}_c^{l(i)}$ $(i=W,H,N)$ is supposed to be fixed. Then, (6) is reduced to

$$\left\langle\mathcal{G}_c^l\right\rangle=\arg\min_{\mathcal{G}_c^l}\left\|\mathcal{Y}_c^l-\mathcal{G}_c^l\times_1\mathrm{D}_c^{l(W)}\times_2\mathrm{D}_c^{l(H)}\times_3\mathrm{D}_c^{l(N)}\right\|_F^2\tag{7}$$

Algorithm 1 Optimization process of DTDLNU

1. Initialize $\mathrm{D}^{(W)}$, $\mathrm{D}^{(H)}$ and $\mathrm{D}^{(N)}$. Initialize all the atoms of $\mathrm{D}^{(i)}$ $(i=W,H,N)$ as random vectors, and orthonormalize each column of dictionaries.
2. Update $\mathcal{G}_c^l$ and $\mathrm{D}_c^{l(i)}$ $(i=W,H,N)$.
(1) Fix $\mathrm{D}_c^{l(i)}$ $(i=W,H,N)$, and calculate the coefficient tensor $\mathcal{G}_c^l$ with (9).
(2) Fix $\mathcal{G}_c^l$ and $\mathrm{D}_c^{l(j)}$ $(j\neq i)$, and update $\mathrm{D}_c^{l(i)}$ by solving (11). Orthonormalize each atom of $\mathrm{D}_c^{l(i)}$.
3. Output.
Return to **step 2** until the values of (4) in adjacent iterations are close enough. Output $\mathrm{D}^{(W)}$, $\mathrm{D}^{(H)}$ and $\mathrm{D}^{(N)}$.

According to [Lathauwer *et al.*, 2000], the solution can be obtained by solving a classical linear least-squares problem:

$$\mathcal{G}_c^l\times_1\mathrm{D}_c^{l(W)}\times_2\mathrm{D}_c^{l(H)}\times_3\mathrm{D}_c^{l(N)}=\mathcal{Y}_c^l\tag{8}$$

To avoid $\mathcal{G}_c^l$ being overdetermined, we orthonormalize each column of $\mathrm{D}_c^{l(i)}$ $(i=W,H,N)$. Then we can get

$$\mathcal{G}_c^l=\mathcal{Y}_c^l\times_1\mathrm{D}_c^{l(W)^T}\times_2\mathrm{D}_c^{l(H)^T}\times_3\mathrm{D}_c^{l(N)^T}\tag{9}$$

When updating $\mathrm{D}_c^{l(i)}$, $\mathcal{G}_c^l$ and $\mathrm{D}_c^{l(j)}$ $(j\neq i)$ are fixed. To calculate $\mathrm{D}_c^{l(i)}$, we can solve the following problem:

$$\begin{aligned}&\left\langle\mathrm{D}_c^{l(i)}\right\rangle=\arg\min_{\mathrm{D}_c^{l(i)}}\left\|\mathcal{Y}_{c(i)}^l-\mathrm{D}_c^{l(i)}\mathcal{G}_{c(i)}^l\overline{p}\left(\mathrm{D}_c^{l(i)}\right)\right\|_F^2\\&+\lambda\left(tr(\mathrm{D}_c^{l(i)T}\mathrm{S}_w^{(i)}\mathrm{D}_c^{l(i)})+\sum_{\substack{h=1\\h\neq c}}^{C}\sum_{q=1}^{L_h}\beta_{cl}^{hq}\left\|\mathrm{D}_c^{l(i)T}\mathrm{D}_h^{q(i)}\right\|_F^2\right)\end{aligned}\tag{10}$$

where $\mathcal{Y}_{c(i)}^l$ and $\mathcal{G}_{c(i)}^l$ are separately the mode-i unfolded forms of $\mathcal{Y}_c^l$ and $\mathcal{G}_c^l$, and

$$\overline{p}\left(D_c^{l(i)}\right)=\begin{cases}\left(\mathrm{D}_c^{l(N)}\otimes\mathrm{D}_c^{l(H)}\right)^T\ if\ i=W\\\left(\mathrm{D}_c^{l(N)}\otimes\mathrm{D}_c^{l(W)}\right)^T\ if\ i=H\\\left(\mathrm{D}_c^{l(H)}\otimes\mathrm{D}_c^{l(W)}\right)^T\ if\ i=N\end{cases}.$$

The solution of (10) can be easily derived by:

$$\begin{aligned}&\mathrm{D}_c^{l(i)}\left(\mathcal{G}_{c(i)}^l\overline{p}\left(\mathrm{D}_c^{l(i)}\right)\right)\left(\mathcal{G}_{c(i)}^l\overline{p}\left(\mathrm{D}_c^{l(i)}\right)\right)^T\\&+\lambda\left(\mathrm{S}_w^{(i)}+\sum_{\substack{h=1\\h\neq c}}^{C}\sum_{q=1}^{L_h}\beta_{cl}^{hq}\mathrm{D}_h^{q(i)}\mathrm{D}_h^{q(i)T}\right)\mathrm{D}_c^{l(i)}\\&=\mathcal{Y}_{c(i)}^l\left(\mathcal{G}_{c(i)}^l\overline{p}\left(\mathrm{D}_c^{l(i)}\right)\right)^T\end{aligned}\tag{11}$$

(11) is a standard Sylvester equation, which can be effectively solved using existing tools [Bartels and Stewart, 1972]. Algorithm 1 describes the optimization of DTDLNU. The optimization is an example of generalized block coordinate descent algorithm where its convergence has been theoretically analyzed for multiconvex optimization [Xu and Yin, 2013].

6 The Classification Scheme of DTDLNU

When $\left\{\mathrm{D}^{(W)},\mathrm{D}^{(H)},\mathrm{D}^{(N)}\right\}$ is available, a test image set can be classified via coding it over these dictionaries. For the given test image set Z^{test}, we can organize it as a 3^{rd}-order

tensor $\mathcal{Z}^{test} \in \mathbb{R}^{d_W \times d_H \times d_N}$. The sparse coding coefficient tensor $\mathcal{Q}$ can be obtained by solving:

$$\min \left\| \mathcal{Z}^{test} - \mathcal{Q} \times_1 \mathrm{D}^{(W)} \times_2 \mathrm{D}^{(H)} \times_3 \mathrm{D}^{(N)} \right\|_F^2 \quad s.t.\ \|\mathcal{Q}\|_0 \leq S_0 \tag{12}$$

where S_0 refers to the total sparsity (i.e. the number of non-zeros) of $\mathcal{Q}$. $\mathcal{Q}$ can be achieved with the Tensor-OMP algorithm [Caiafa and Cichocki, 2012]. The reconstruction error associated with the c^{th} $(c = 1, \cdots, C)$ class is computed by:

$$e_c = \min_l \left\| \mathcal{Z}^{test} - \mathcal{Q}_c^l \times_1 \mathrm{D}_c^{l(W)} \times_2 \mathrm{D}_c^{l(H)} \times_3 \mathrm{D}_c^{l(N)} \right\|_F^2 \tag{13}$$

where $\mathcal{Q}_c^l$ is the coding coefficient tensor corresponding to the l^{th} set of the c^{th} class. The classification can be done by assigning the test image set to the class with the smallest reconstruction error.

7 Experiments

7.1 Compared Methods

In experiments, we compare DTDLNU with five categories of state-of-the-art related methods including:
(1) Subspace and manifold based methods: **MDA** [Wang and Chen, 2009] and **DARG** [Wang *et al.*, 2015];
(2) Affine/convex hull based methods: **AHISD** [Cevikalp and Triggs, 2010], **CHISD** [Cevikalp and Triggs, 2010] and **SANP** [Hu *et al.*, 2012];
(3) Covariance matrix based methods: **CDL** [Wang *et al.*, 2012] and **LMKML** [Lu *et al.*, 2013];
(4) Deep learning method: **DRM-WV** [Hayat *et al.*, 2015];
(5) Dictionary learning based methods: **ISCRC** [Zhu *et al.*, 2014], **DFRV** [Chen *et al.*, 2012] and **SFDL** [Lu *et al.*, 2014].

7.2 Datasets

In experiments, we use three challenging and large datasets, i.e., YouTube Celebrities (YTC) [Kim *et al.*, 2008], COX [Huang *et al.*, 2015], and YouTube Faces (YTF) [Wolf *et al.*, 2011]. YTC contains 1,910 video sequences of 47 celebrities from YouTube. COX is a dataset involving 1,000 different subjects, each of which has 3 videos captured by different camcorders. YTF contains 3,425 videos of 1,596 subjects downloaded from YouTube. And there are large variations in pose, illumination, expression, and resolution in these videos.

We employ the Viola-Jones face detector [Viola and Jones, 2004] to detect the faces in each frame and resize the detected faces to gray-scale images of 30×30 for YTC, 32×40 for COX, and 30×30 for YTF. Histogram equalization is implemented to reduce the illumination variations.

7.3 Experimental Settings

To make a fair comparison with related ISC methods, we follow the protocol used in [Wang *et al.*, 2015; Hayat *et al.*, 2015; Zhu *et al.*, 2014; Lu *et al.*, 2014]. On YTC, ten random selections for training and testing videos are conducted for reporting average experimental results. The whole dataset is equally divided into ten folds with each containing 9 videos per subject. In each fold, 3 videos per subject are randomly selected for training, and the remaining 6 are selected for testing. For COX, we follow the same protocol as the prior work [Huang *et al.*, 2015; Wang *et al.*, 2015], which conducted ten-fold cross validation, i.e., 10 randomly selected gallery/probe combinations. Since there are 3 independent testing sets of videos in COX, each person has one video as the gallery and the remaining two videos for two different probes, thus in total 6 groups of testings need to be conducted. For YTF, we follow the standard evaluation protocol [Wolf *et al.*, 2011]. 5,000 video pairs are collected randomly and half of them are from the same subject, half from different subjects. These pairs are then divided into 10 splits and each split contains 250 intra-personal pairs and 250 inter-personal pairs. The evaluation protocol of YTF was originally developed for face verification. For verification, we compute the class label for each video in the given video pair with our classification scheme, and then make a decision whether the video pair is an intra-personal pair or not. We perform 10-fold cross validation. For these datasets, one video is regarded as an image set.

In experiments, the tuning parameters (the balance factor λ and the neighboring set number k) of DTDLNU are set by using 5-fold cross validation with training data. Concretely, they are set as $\lambda = 1.5$ and $k = 50$ on YTC; $\lambda = 0.8$ and $k = 90$ on COX; and $\lambda = 1.5$ and $k = 220$ on YTF. The default dictionary atoms number for each set in DTDLNU, which is associated with $r_c^{l(W)}$, $r_c^{l(H)}$ and $r_c^{l(N)}$, is set as $r_c^{l(W)} = d_W$, $r_c^{l(H)} = d_H$ and $r_c^{l(N)} = d_N$. S_0 can be automatically selected.

7.4 Results and Analysis

Comparison with the State-of-the-Arts: Table 1 shows average recognition/verification results of compared methods on three datasets. From Table 1, DTDLNU performs better than eleven compared ISC methods on the YTC and YTF datasets. On COX, DTDLNU also outperforms the recently presented ISC methods, like [Wang *et al.*, 2015; Hayat *et al.*, 2015; Lu *et al.*, 2014], in all testing cases. To observe the effect of using tensor for image set modeling intuitively, we also compare DTDLNU with the method that organizes image samples as vectors and learns ordinary dictionary (rather than tensor dictionary) by using the vector version of our objective function. We call this method as DTDLNU$_{vec}$. It can be seen that DTDLNU significantly outperforms DTDLNU$_{vec}$ on all datasets. All these results indicate the effectiveness of modeling image set with tensor and learning uncorrelated discriminant tensor dictionaries.

Evaluation of the Neighbor-uncorrelated Discriminant Tensor Dictionary (NDTD) Term: Fig. 2 shows tensor

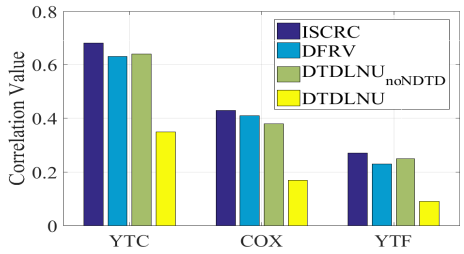

Figure 2: Between-class set-specific sub-dictionary correlation of ISCRC, DFRV, DTDLNU$_{noNDTD}$ and DTDLNU.

Table 1: Average recognition/verification rates (%) of all compared methods on three datasets. "COXij" represents the experiment using the i^{th} set of image sets as gallery and the j^{th} set of image sets as probe.

Datasets	MDA	DARG	AHISD	CHISD	SANP	CDL	LMKML	DRM-WV	ISCRC	DFRV	SFDL	DTDLNU$_{vec}$	DTDLNU
YTC	68.12	78.16	66.58	67.20	68.39	69.97	78.35	76.21	73.88	74.53	76.91	78.41	**80.16**
YTF	66.28	79.03	63.43	65.69	76.69	72.15	77.80	84.71	77.92	78.65	80.21	82.66	**85.16**
COX12	65.83	83.71	53.03	56.90	58.76	78.43	56.14	75.33	69.74	58.57	76.42	80.45	**84.26**
COX13	62.96	90.13	36.13	30.13	38.07	85.31	44.26	87.83	60.71	77.14	89.29	89.31	**90.55**
COX23	36.20	85.08	17.50	15.03	31.49	79.71	33.14	78.75	37.66	80.05	81.44	82.75	**85.12**
COX21	55.53	81.96	43.51	44.36	45.22	75.56	55.37	80.96	61.09	40.02	78.65	79.16	**83.10**
COX31	43.24	89.99	34.99	26.40	48.10	85.84	39.83	84.15	64.96	51.43	86.72	88.23	**90.57**
COX32	29.94	88.35	18.80	13.69	28.43	81.87	29.54	81.07	37.71	51.45	81.37	84.25	**89.49**

Table 3: Computation time (seconds) of all compared methods on YTC for training and testing (classification of one image set).

Methods	MDA	DARG	AHISD	CHISD	SANP	CDL	LMKML	DRM-WV	ISCRC	DFRV	SFDL	DTDLNU
Training	183.9	359.2	N/A	N/A	N/A	68.3	4225.9	3873.4	N/A	8637.1	7518.2	3605.9
Testing	3.3	8.7	8.4	6.8	47.2	12.4	204.2	4.8	42.5	5.4	6.5	157.1

Table 2: Classification accuracies (%) of the versions of DTDLNU without the NDTD term, only with part 1, only with part 2, or with the term.

Datasets	Without the term	With part 1	With part 2	With the term
YTC	76.38	78.47	78.26	80.16
COX	78.94	84.06	82.63	87.18
YTF	79.10	82.59	81.18	85.16

dictionary correlation[1] of set-specific sub-dictionaries corresponding to different classes learned by DTDLNU and DTDLNU$_{noNDTD}$ (the version of DTDLNU that does not include the NDTD term). In Fig. 2, we also report the inter-class set-specific dictionary correlation corresponding to ISCRC and DFRV. Table 2 reports the classification results of DTDLNU and DTDLNU$_{noNDTD}$. In the table, we also report the results of DTDLNU only with part 1 or part 2 of the term. Here, **part 1** means minimizing the within-class scatter in projected tensor space; and **part 2** means minimizing correlation of sub-dictionaries corresponding to neighbor sets from different classes. It is noted that for COX, we report the average results across 6 groups of testings.

According to Fig. 2, with the designed term, DTDLNU owns lower between-class dictionary correlation. From Table 2, we can see that the designed term can improve the classification results, which demonstrates the effectiveness of the term. In addition, the part 1 plays a relatively more important role than the part 2 in the term.

Parameter Analysis: To evaluate the influences of the balance factor λ and the neighboring set number k, we separately conduct experiments by changing the values of λ from 0 to 1.5 with step length 0.05, and of k from 5 to 60 with step length 5 on YTC (when $k > 60$, the performance is stable, and when $\lambda > 1.5$, the performance will experience a slight decrease). Fig. 3 shows the classification accuracy of our approach versus different values of λ or k. We can see that its performances are stable with respect to λ in the range of [1.1,1.5], and with respect to k in the range of [50,60]. For simplicity, we set λ as 1.5 and k as 50 on YTC. A similar phenomenon also exists on the other two datasets.

Computational Time: Lastly, we report the computational time of compared methods. Our hardware configuration comprises a 2.8-GHz CPU and a 24GB RAM. Table 3 tabulates the computational time of different methods on YTC. The reported testing time refers to the time of classifying one image set. We can see that our approach requires less training time than that of LMKML, DRM-WV, DFRV and SFDL. In addition, the testing time of DTDLNU is comparable to that of other methods.

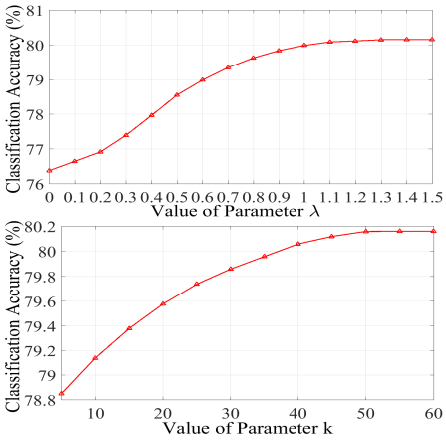

Figure 3: Classification accuracies versus λ and k on YTC.

8 Conclusion

In this paper, by modeling an image set as a third-order tensor, we can well preserve the inherent spatial structure of the set. We for the first time introduce TDL into ISC for learning two spatial dictionaries and one set dictionary. We thus propose a novel ISC approach DTDLNU. It can make the obtained dictionaries have favorable discriminability and reduce the between-class tensor dictionary correlation. We apply DTDLNU for ISC tasks on three challenging datasets. Experimental results demonstrate that DTDLNU achieves better classification results than several state-of-the-art methods.

Acknowledgments

The work described in this paper was supported by the Scientific Research Staring Foundation for Introduced Talents in NJUPT (NUPTSF, No. NY217009), and Province-School-Region Project of Henan University (No. 2016S11).

[1]Here, the between-class set-specific sub-dictionary correlation is calculated by $corr = \frac{1}{N_s} \sum_{i=\{W,H,N\}} \sum_{c=1}^{C} \sum_{l=1}^{L_c} \sum_{\substack{h=1 \\ h\neq c}}^{C} \sum_{q=1}^{L_h} \left\| \mathrm{D}_c^{l(i)T} \mathrm{D}_h^{q(i)} \right\|_F^2$. Here, N_s denotes the number of accumulating calculations.

References

[Bartels and Stewart, 1972] Richard H. Bartels and G. W. Stewart. Solution of the matrix equation ax+xb=c. *Communications of the ACM*, 15(9):820–826, 1972.

[Caiafa and Cichocki, 2012] Cesar F. Caiafa and Andrzej Cichocki. Block sparse representations of tensors using kronecker bases. In *ICASSP*, pages 2709–2712, 2012.

[Cevikalp and Triggs, 2010] Hakan Cevikalp and Bill Triggs. Face recognition based on image sets. In *CVPR*, pages 2567–2573, 2010.

[Chen *et al.*, 2012] Yi-Chen Chen, Vishal M. Patel, P. Jonathon Phillips, and Rama Chellappa. Dictionary-based face recognition from video. In *ECCV*, pages 766–779, 2012.

[Hayat *et al.*, 2014] Munawar Hayat, Mohammed Bennamoun, and Senjian An. Learning non-linear reconstruction models for image set classification. In *CVPR*, pages 1915–1922, 2014.

[Hayat *et al.*, 2015] Munawar Hayat, Mohammed Bennamoun, and Senjian An. Deep reconstruction models for image set classification. *TPAMI*, 37(4):713–727, 2015.

[Hu *et al.*, 2012] Yiqun Hu, Ajmal S. Mian, and Robyn Owens. Face recognition using sparse approximated nearest points between image sets. *TPAMI*, 34(10):1992–2004, 2012.

[Huang *et al.*, 2015] Zhiwu Huang, Shiguang Shan, Ruiping Wang, Haihong Zhang, Shihong Lao, Alifu Kuerban, and Xilin Chen. A benchmark and comparative study of video-based face recognition on cox face database. *TIP*, 24(12):5967–5981, 2015.

[Kim *et al.*, 2008] Minyoung Kim, Sanjiv Kumar, Vladimir Pavlovic, and Henry Rowley. Face tracking and recognition with visual constraints in real-world videos. In *CVPR*, pages 1–8, 2008.

[Lathauwer *et al.*, 2000] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. On the best rank-1 and rank-(r1, r2,..., rn) approximation of higher-order tensors. *SIMAX*, 21(4):1324–1342, 2000.

[Li and Schonfeld, 2014] Qun Li and Dan Schonfeld. Multilinear discriminant analysis for higher-order tensor data classification. *TPAMI*, 36(12):2524–2537, 2014.

[Lu *et al.*, 2013] Jiwen Lu, Gang Wang, and Pierre Moulin. Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In *ICCV*, pages 329–336, 2013.

[Lu *et al.*, 2014] Jiwen Lu, Gang Wang, Weihong Deng, and Pierre Moulin. Simultaneous feature and dictionary learning for image set based face recognition. In *ECCV*, pages 265–280, 2014.

[Peng *et al.*, 2014] Yi Peng, Deyu Meng, Zongben Xu, Chenqiang Gao, Yi Yang, and Biao Zhang. Decomposable nonlocal tensor dictionary learning for multispectral image denoising. In *CVPR*, pages 2949–2956, 2014.

[Quan *et al.*, 2015] Yuhui Quan, Yan Huang, and Hui Ji. Dynamic texture recognition via orthogonal tensor dictionary learning. In *ICCV*, pages 73–81, 2015.

[Roemer *et al.*, 2014] Florian Roemer, Giovanni Del Galdo, and Martin Haardt. Tensor-based algorithms for learning multidimensional separable dictionaries. In *ICASSP*, pages 3963–3967, 2014.

[Shah *et al.*, 2016] Syed Afaq Ali Shah, Mohammed Bennamoun, and Farid Boussaid. Iterative deep learning for image set based face and object recognition. *Neurocomputing*, 174:866–874, 2016.

[Viola and Jones, 2004] Paul Viola and Michael J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[Wang and Chen, 2009] Ruiping Wang and Xilin Chen. Manifold discriminant analysis. In *CVPR*, pages 429–436, 2009.

[Wang *et al.*, 2012] Ruiping Wang, Huimin Guo, Larry S. Davis, and Qionghai Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *CVPR*, pages 2496–2503, 2012.

[Wang *et al.*, 2015] Wen Wang, Ruiping Wang, Zhiwu Huang, Shiguang Shan, and Xilin Chen. Discriminant analysis on riemannian manifold of gaussian distributions for face recognition with image sets. In *CVPR*, pages 2048–2057, 2015.

[Wolf *et al.*, 2011] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, pages 529–534, 2011.

[Xu and Yin, 2013] Yangyang Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, 6(3):1758–1789, 2013.

[Zhang *et al.*, 2013] Yangmuzi Zhang, Zhuolin Jiang, and Larry S. Davis. Discriminative tensor sparse coding for image classification. In *BMVC*, 2013.

[Zhang *et al.*, 2016] Man Zhang, Ran He, Dong Cao, Zhenan Sun, and Tieniu Tan. Simultaneous feature and sample reduction for image-set classification. In *AAAI*, pages 1401–1407, 2016.

[Zheng *et al.*, 2017] Peng Zheng, Zhong-Qiu Zhao, Jun Gao, and Xindong Wu. Image set classification based on cooperative sparse representation. *PR*, 63:206–217, 2017.

[Zhu *et al.*, 2014] Pengfei Zhu, Wangmeng Zuo, Lei Zhang, Simon Chi-Keung Shiu, and David Zhang. Image set based collaborative representation for face recognition. *TIFS*, 9(7):1120–1132, 2014.

[Zubair *et al.*, 2014] Syed Zubair, Wenwu Wang, and Jonathon Chambers. Discriminative tensor dictionaries and sparsity for speaker identification. In *HSCMA*, pages 37–41, 2014.