Low-Rank Tensor Learning with Discriminant Analysis for Action Classification and Image Recovery*

Chengcheng Jia[1], **Guoqiang Zhong**[3] and **Yun Fu**[1,2]
[1] Electrical and Computer Engineering, [2]Computer and Information Science, Northeastern University, Boston, USA
[3]Department of Computer Science and Technology, Ocean University of China, Qingdao 266100, China
jia.ch@husky.neu.edu, gqzhong2012@gmail.com, yunfu@ece.neu.edu

Abstract

Tensor completion is an important topic in the area of image processing and computer vision research, which is generally built on extraction of the intrinsic structure of the tensor data. Drawing on this fact, action classification, relying heavily on the extracted features of high-dimensional tensors, may indeed benefit from tensor completion techniques. In this paper, we propose a low-rank tensor completion method for action classification, as well as image recovery. Since there may exist distortion and corruption in the tensor representations of video sequences, we project the tensors into a subspace, which contains the invariant structure of the tensors. In order to integrate useful supervisory information for classification, we adopt a discriminant analysis criterion to learn the projection matrices. The resulting multi-variate optimization problem can be effectively solved using the augmented Lagrange multiplier (ALM) algorithm. Experiments demonstrate that our method results with better accuracy compared with some other state-of-the-art low-rank tensor representation learning approaches on the MSR Hand Gesture 3D database and the MSR Action 3D database. By denoising the Multi-PIE face database, our experimental setup testifies the proposed method can also be employed to recover images.

Introduction

Images and video sequences can be naturally represented as high-dimensional tensors. However, the real tensor representations of images and videos are usually incomplete, due to missing elements or the presence of noise. This issue impels great research interest for recovering the original tensors these past recent years. Many tensor representation learning approaches have been proposed (Chen and Saad 2009; Koch and Lubich 2010; Haegeman et al. 2011; Holtz, Rohwedder, and Schneider 2012; Khoromskij, Oseledets, and Schneider 2012; Arnold and Jahnke 2012; Lubich et al. 2013; Uschmajew and Vandereycken 2013; Mu et al. 2013). Many of these previous approaches aim to learn the low-dimensional representations of tensors, while mainly using the high-order singular value decomposition (HOSVD). Regardless, some tensor approximation approaches have been proposed as well, which, in general, estimate a rank-one tensor via vector outer production (Espig 2007; Kazeev and Tyrtyshnikov 2010; Acar, Dunlavy, and Kolda 2011; Espig and Hackbusch 2012; Phan, Anh Huy and Tichavskỳ, Petr and Cichocki, Andrzej 2012; 2013; Shi et al. 2013).

As of recent, several **low-rank** tensor representation learning approaches have been proposed for computer vision applications, such as image reflection and alignment (Zhang et al. 2013), target tracking (Shi et al. 2013), face and object recognition (Ding, Huang, and Luo 2008; Zhong and Cheriet 2014). These methods aim to learn the invariant structure of the tensor data. However, the formulation and optimization of these approaches are quite different. For concreteness, Zhang et al. performed the low-rank tensor representation learning on the original images, in parallel to eliminate noise and recover missing pixels (Zhang et al. 2013); Shi et al. employed rank-one tensors for multi-target tracking (Shi et al. 2013); Ding et al. used rank-one tensors to reduce tensor dimensionality for applications such as video compression and face classification (Ding, Huang, and Luo 2008); Zhong and Cheriet proposed a manifold-based tensor representation learning model for face and object recognition (Zhong and Cheriet 2014). Note, although these low-rank tensor representation learning approaches have been successful when applied to different visual classification scenarios, they are rarely integrated in the supervisory information for maximizing class discrimination (Saghafi and Rajan 2012; Jia and Yeung 2008; Etemad and Chellappa 1997), which may dramatically improve the visual classification accuracy.

Some low-rank matrix learning approaches based on the discriminant analysis criterion have been addressed. For example, Zheng et al. used intra-class and inter-class information for face recognition (Zheng et al. 2013), also, Cai et al. employed the discriminant analysis criterion with low-rank matrix learning for face and digits recognition (Cai et al. 2013). These discriminative low-rank matrix learning approaches have shown that the label information of data is typically beneficial for visual classification. Rare previous work was integrated the discriminant analysis criterion into a low-rank tensor completion model, to the best of our knowl-

*This research is supported in part by the NSF CNS award 1314484, Office of Naval Research award N00014-12-1-1028, and U.S. Army Research Office grant W911NF-13-1-0160.

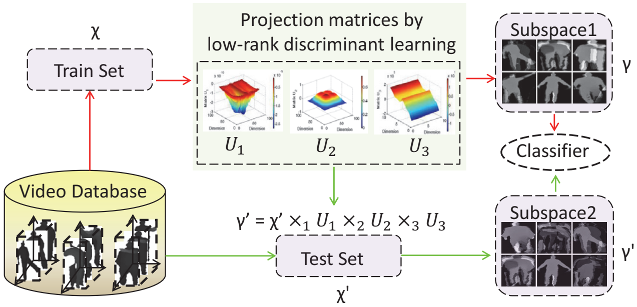


Figure 1: Framework of the proposed algorithm for action recognition. The tensor training set $\mathcal{X}$ is used for calculating the low-rank projection matrices, which are employed for subspace alignment of training and testing action videos $\mathcal{Y}$ and $\mathcal{Y}'$.

edge. This can be directly applied to visual classification applications, such as action classification.

In this paper, we present a supervised low-rank tensor completion method for dimensional reduction, to learn an optimal subspace for action video recognition. Our model automatically learns the low dimensionality of tensor, opposed to manually pre-defined, as other dimensional reduction methods. Considering the underlying structure information of the whole high-dimensional dataset, it can use the low-rank learning to extract the structure for image recovery, while integrating with the discriminant analysis criterion. Figure 1 shows the framework of our method applied to the video-based action classification. We first select a training set from an action video database to learn the low-rank projection matrices, which are then used to calculate a tensor subspace for the action classification. When calculating the low-rank projection matrices, we adopt a discriminant analysis criterion as a regularizer to avoid over-fitting. Meanwhile, with this discriminant analysis criterion, supervisory information is seamlessly integrated in the low-rank tensor completion model. After projecting the original training and testing sets to the learned tensor subspace, we predict the labels of the test video sequences with a K-nearest neighbor (KNN) classifier. We add the sample information to recovery some face images by removing different illuminations.

The contributions of this paper are as follows:

1. We proposed a new discriminative method for low-rank tensor completion, which automatically learns the low dimensionality of the tensor subspace for feature extraction.

2. We integrated the discriminant analysis criterion in the low-rank tensor completion model based on the given supervisory information.

3. The proposed model extracts the underlying structure of the original tensor data by low-rank learning, which reconstructs the data from the learned tensor subspace, for high-dimensional image recovery.

Preliminary

A N-dimensional array is called a tensor, which is represented as $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_n \times \ldots \times I_N}$, where I_n is the *mode-n* dimension ($1 \leqslant n \leqslant N$). A metadata of $\mathcal{A}$ is presented as $\mathcal{A}_{i_1 i_2 \ldots i_n \ldots i_N}$, where i_n is the index of *mode-n* ($1 \leqslant i_n \leqslant I_n$). The *mode-n* vectors of $\mathcal{A}$ are the vectors in $\mathbb{R}^{I_n}$, by keeping the vectors of other modes fixed (Kolda and Bader 2009).

Definition 1: *(Mode-n unfolding) The mode-n unfolding of $\mathcal{A}$ is denoted by matrix $A_{(n)} \in \mathbb{R}^{I_n \times (I_1 \cdot I_2 \ldots I_{n-1} \cdot I_{n+1} \ldots I_N)}$, with the column vectors that are the mode-n vectors of $\mathcal{A}$.*

Definition 2: *(Core tensor) A tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times \ldots \times I_N}$ is decomposed by $U_n \in \mathbb{R}^{I_n \times J_n} (1 \leqslant n \leqslant N)$ as*

$$\mathcal{S} = \mathcal{A} \times_1 U_1 \times_2 U_2 \ldots \times_n U_n \ldots \times_N U_N, \tag{1}$$

where $\mathcal{A} = \mathcal{S} \times_1 U_1^T \times_2 U_2^T \ldots \times_n U_n^T \ldots \times_N U_N^T$, $\times_n$ indicates mode-n product. The transformed tensor $\mathcal{S} \in \mathbb{R}^{J_1 \times J_2 \times \ldots \times J_N}$ is called the core tensor. Its mode-n unfolding matrix is represented as $S_{(n)} = (U_N \ldots U_n \ldots U_1 \mathcal{A})_{(n)}$.

Definition 3: *(Tensor Frobenius Norm) The tensor Frobenius Norm (F-norm) can be calculated by*

$$\|\mathcal{A}\|_F = \sqrt{\sum_{i_1} \cdots \sum_{i_N} \mathcal{A}^2_{i_1 i_2 \ldots i_n \ldots i_N}}\,. \tag{2}$$

Low-rank Tensor Completion

Here we introduce the proposed method, along with an in-depth lot at its formulation and optimization.

Given a set of N-order tensors $\mathcal{X} = \{\mathcal{X}_i \in \mathbb{R}^{I_1 \ldots I_N} \mid i = 1, \ldots, M\}$, the corresponding labels $\{l_1, \ldots, l_M\}$, and suppose the projection matrices are $U_n \in \mathbb{R}^{I_n \times J_n}$. Then tensors after projection can be calculated as

$$\mathcal{Y} = \mathcal{X} \times_1 U_1 \ldots \times_n U_n \ldots \times_N U_N. \tag{3}$$

Previous low-rank tensor completion and approximation methods (Romera-Paredes and Pontil 2013; Cai et al. 2013; Krishnamurthy and Singh 2013; Chen et al. 2013) are widely used for image denoising and recovering an alignment. The usual way to obtain the intrinsic structure of the tensors is to calculate the trace norm of the N-order tensor as following:

$$\min_{X_{(1)}, \ldots, X_{(N)}} \sum_{n=1}^{N} \|X_{(n)}\|_* + \lambda \|E_{(n)}\|_l, \tag{4}$$

where $X_{(n)}$ is the *mode-n* unfolding matrix, and $E_{(n)}$ is the *mode-n* error tensor, $l \in \{*, 1\}$. This means the error item can be calculated by trace norm or sparse learning.

To learn an effective subspace of the tensors for action classification, we alternatively optimize each projection matrices. We denote

$$X_{(n)} = U_n D_{(n)}, \tag{5}$$

where $X_{(n)}$ is the *mode-n* unfolding of tensor datum, U_n is the projection matrix, $D_{(n)} = (U_N \ldots U_{n+1}\, U_{n-1} \ldots U_1\, \mathcal{X})_{(n)}$. During learning the projection matrices U_n, $D_{(n)}$ is taken as a constant matrix. Hence, Problem (4) can be transformed to minimizing the trace norm of U_n according to

$$\min_{U_n} \sum_{n=1}^{N} \|U_n\|_*, \tag{6}$$

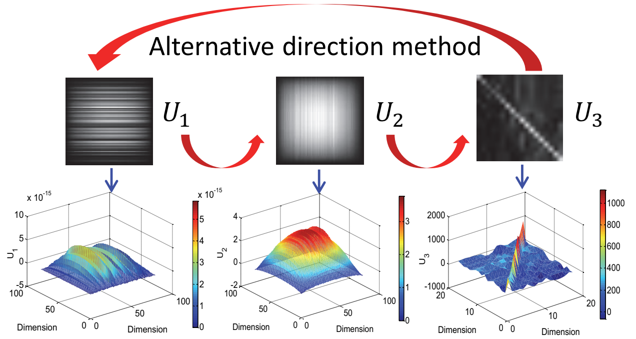


Figure 2: Illustration of 2, 3-dimensional projection matrices respectively. Form left to right: U_1, U_2 and U_3. During learning, each projection matrix is calculated by the alternative direction method.

with some conditions imposed. Meanwhile, the low dimensional structure of U_n can be automatically captured by the low-rank learning, which is useful for tensorial subspace learning and dimensional reduction.

The matrices U_n can indicate rotation properties of tensors in the subspace, such as row space, column space. It can also reflect the degree of movement in the frame space. Figure 2 shows how the U_n works. Here $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ is a 3-order tensor with 1≤ n ≤3. From left to right, it shows the learned projection matrices U_1, U_2 and U_3, which correspond to the transformation in the row, column, and frame space, respectively. Compared with the convectional vector-based method, the matrices can reflect different variances in the row, column of an image, and show the sample information as well. The first and second rows illustrate the 2, 3-dimensional projection matrices, respectively. U_1 and U_2 reflect the location of the movement in the row space and the column space of the database. This is different from the vector-based low-rank method (Liu, Lin, and Yu 2010), which cannot reflect the variance in the row and column of an image. As for U_3, each small block stands for a frame reflecting the significance of the frames – if it is a full-rank matrix, each frame plays an important role in the video sequence. Here, the color bar in the second row means different values of U_n (1≤ $n \leq 3$). During learning the projection matrices are calculated in an iterative process performed by the alternative direction method.

Discriminant analysis

In order to integrate supervisory information into the low-rank tensor completion model, the discriminant analysis criterion is adopted as a regularizer. For simplicity, let $A = X_{(n)}$. The inter-class and intra-class scatter matrices as follows:

$$B_n = \sum_{i=1}^{\mathrm{C}} m_i(\overline{A}_i - \overline{A})(\overline{A}_i - \overline{A})^{\mathrm{T}}, \tag{7}$$

$$W_n = \sum_{i=1}^{\mathrm{C}}\sum_{j=1}^{\mathrm{C}_i}(A_{ij} - \overline{A}_i)(A_{ij} - \overline{A}_i)^{\mathrm{T}}, \tag{8}$$

where B_n, W_n are the *mode-n* inter-class and intra-class matrices respectively. $\overline{A}_i$, $\overline{A}$ are the mean samples of the i-th class and the total number of samples, respectively. A_{ij} is the j-th sample of the i-th class. m_i denotes the number of i-th class.

The corresponding discriminant regularizer is given as

$$\lambda\|U_n^{\mathrm{T}}(W_n - \alpha B_n)U_n\|_{\mathrm{F}}^2, \tag{9}$$

where $\|\cdot\|_{\mathrm{F}}^2$ is the Frobenius Norm (Kolda and Bader 2009), α is the tuning parameter to control the value of the regularization, and λ is the parameter to balance the low-rank item and the discriminant item. According to the regularization constraint the low-rank tensor completion model can be expressed as follows:

$$\min_{U_1,\ldots,U_{\mathrm{N}}} \sum_{=1}^{\mathrm{N}} \|U_n\|_* + \lambda\|U_n^{\mathrm{T}}(W_n - \alpha B_n)U_n\|_{\mathrm{F}}^2. \tag{10}$$

With this, the discriminant regularizer not only avoids over-fitting, but also seamlessly integrates the intra-class and inter-class information into the proposed model.

Objective function

Provided an N-order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_{\mathrm{N}}}$. For its real data there always exists some noise or corruption $\mathcal{E}$, which satisfies the following conditions: (1) there is only small fragment or missing part; (2) the location of the error is unknown. Hence, the original tensor can be represented as

$$\mathcal{X} = \mathcal{Y} + \mathcal{E}, \tag{11}$$

where $\mathcal{Y} = U_{\mathrm{N}} \ldots U_1\mathcal{X}$ is the low-rank tensor, and $\mathcal{E}$ is the error. We next employed the error item as a constraint, defined by *mode-n* unfolding as follows:

$$\|Y_{(n)} - X_{(n)}\|_{\mathrm{F}}^2 \leq \epsilon, \tag{12}$$

where ϵ is the bias.

Considering the discriminant regularizer and the error item, The low-rank tensor completion model is rewritten as

$$\begin{aligned} &\min_{U_n} \sum_{n=1}^{\mathrm{N}} \|U_n\|_* + \lambda\|U_n^{\mathrm{T}}(W_n - \alpha B_n)U_n\|_{\mathrm{F}}^2 \\ &\text{s.t.} \quad \|Y_{(n)} - X_{(n)}\|_{\mathrm{F}}^2 \leq \epsilon, Y_{(n)} = (U_{\mathrm{N}} \ldots U_1\mathcal{X})_{(n)} \end{aligned} \tag{13}$$

This model is intractable, because the error item is not convex with respect to the variables. In order to solve this problem, we employ the augmented Lagrange multiplier (ALM) algorithm (Lin, Chen, and Ma 2010) to optimize Problem (13).

Optimization

Due to the difficulty of solving Eq. (13), we introduce two auxiliary matrices J_n and $M_{(n)}$ to the objection function. The regularization $\|M_{(n)} - X_{(n)}\|_{\mathrm{F}}^2 \leq \epsilon$ is set as an error term in Eq. (13), allowing the objective function to be integrated and rewritten as

$$\begin{aligned} &\min_{U_1,\ldots,U_N} \sum_{n=1}^{\mathrm{N}} \|J_n\|_* + \lambda\|{J_n}^{\mathrm{T}}(W_n - \alpha B_n)J_n\|_{\mathrm{F}}^2 \\ &\qquad\qquad + \beta\|M_{(n)} - X_{(n)}\|_{\mathrm{F}}^2 \\ &\text{s.t.} \quad U_n = J_n, Y_{(n)} = (U_{\mathrm{N}} \ldots U_1\mathcal{X})_{(n)}, Y_{(n)} = M_{(n)}, \end{aligned} \tag{14}$$

where β is the parameter of the error item. We use the ALM algorithm to solve the following unconstrained multi-variate optimization problem. The Lagrange function is defined as

$$
\begin{aligned}
L_n = & \underset{\substack{J_n, U_n, Y_{(n)}, \\ M_{(n)}, Y_1, Y_2, Y_3}}{\operatorname{argmin}} \sum_{n=1}^{\mathrm{N}} \|J_n\|_* + \lambda \|J_n^{\mathrm{T}}(W_n - \alpha B_n) J_n\|_{\mathrm{F}}^2 \\
& + \beta \|M_{(n)} - X_{(n)}\|_{\mathrm{F}}^2 + \mathrm{tr}\left[V_1^{\mathrm{T}}\left(Y_{(n)} - (U_{\mathrm{N}} \ldots U_1 \mathcal{X})_{(n)}\right)\right] \\
& + \mathrm{tr}\left[V_2^{\mathrm{T}}(U_n - J_n)\right] + \mathrm{tr}\left[V_3^{\mathrm{T}}\left(Y_{(n)} - M_{(n)}\right)\right] \\
& + \frac{\mu}{2}\left[\|Y_{(n)} - (U_{\mathrm{N}} \ldots U_1 \mathcal{X})_{(n)}\|_{\mathrm{F}}^2 + \|U_n - J_n\|_{\mathrm{F}}^2\right. \\
& \left. + \|Y_{(n)} - M_{(n)}\|_{\mathrm{F}}^2\right],
\end{aligned}
\tag{15}
$$

where V_1, V_2, V_3 are the Lagrange multipliers, $\mu > 0$ is the penalty operator, tr$(\cdot)$ is the trace of a matrix. All the variables in the Lagrange function are solvable as following:

$$
\begin{cases}
J_n = \underset{J_n}{\operatorname{argmin}} \frac{1}{\mu} \sum \|J_n\|_* + \frac{1}{2} \|J_n - (\mathbf{I} + \\
\qquad 2\frac{\lambda}{\mu}(W_n - \alpha B_n))^{-1}(U_n + \frac{V_2}{\mu})\|_{\mathrm{F}}^2, \\
U_n = \left(Y_{(n)} D_{(n)}^{\mathrm{T}} + J_n + \frac{1}{\mu}\left(V_1 D_{(n)}^{\mathrm{T}} - V_2\right)\right) \cdot \\
\qquad \left(D_{(n)} D_{(n)}^{\mathrm{T}} + \mathbf{I}\right)^{-1}, \\
Y_{(n)} = \frac{1}{2}(M_{(n)} + U_n D_{(n)} - \frac{1}{\mu}(V_1 + V_3)) = 0, \\
M_{(n)} = \frac{1}{2\beta + \mu}(2\beta X_{(n)} + \mu Y_{(n)} + V_3), \\
V_1 = V_1 + \mu\left(Y_{(n)} - (U_N \ldots U_1 \mathcal{X})_{(n)}\right), \\
V_2 = V_2 + \mu\left(U_n - J_n\right), \\
V_3 = V_3 + \mu\left(Y_{(n)} - M_{(n)}\right).
\end{cases}
\tag{16}
$$

The convergence conditions are $\|U_n - J_n\|_\infty < \varepsilon$, $\|Y_{(n)} - (U_N \ldots U_1 \mathcal{X})_{(n)}\|_\infty < \varepsilon$, and $\|Y_{(n)} - M_{(n)}\|_\infty < \varepsilon$. The whole iterative procedure is shown in Algorithm 1.

Algorithm 1 Low-rank tensor discriminant analysis (LRTD)

INPUT: M labeled N-order training tensors $\Gamma = \{\mathcal{X}_i\}$, associated labels $\{l_1, \ldots, l_{\mathrm{M}}\} \in \{1, \ldots, \mathrm{C}\}$, the tuning parameter α, λ, β, and the maximum number of training iterations $\mathrm{t}_{\mathrm{max}}$.

OUTPUT: Updated $U_n^{(t)}$ $(1 \leqslant n \leqslant \mathrm{N})$.

```
Initialize $U_n$ by eigen-decomposition of dataset $\Gamma$. $J_n = 0$, $V_1 = V_2 = V_3 = 0$, $\mu = 10^{-6}$, $\mu_{\mathrm{max}} = 10^6$, $\rho = 1.1$, and $\epsilon = 10^{-8}$.
for $t = 1$ to $t_{\mathrm{max}}$ do
    for $n = 1$ to N do
        $\mathcal{X}_i \leftarrow \mathcal{X}_i \times_1 (U_1^{(t-1)})^{\mathrm{T}} \cdots \times_{n-1} (U_{n-1}^{(t-1)})^{\mathrm{T}} \times_{n+1} (U_{n+1}^{(t-1)})^{\mathrm{T}} \cdots \times_{\mathrm{N}} (U_{\mathrm{N}}^{(t-1)})^{\mathrm{T}}$.
        while $t^{'} < t^{'}_{\mathrm{max}}$ do
            1) Update $B_n$, $W_n$ by Eqs. (7 ~8).
            2) Update $J_n$, $U_n$, $Y_{(n)}$, $M_{(n)}$ and multipliers $V_1$, $V_2$, $V_3$ via fixing others in equation set (16).
            3) Update $\mu$ by $\mu = \min_{\mu}(\rho\mu, \max_\mu)$.
        end while
        $U_n^{(t-1)} = U_n^{(t)}$.
    end for
end for
```

Improvement

In all actuality, there are many of high-dimensional images with noise or small corruption. Here, we improve the low-rank tensor model in order to complete such images. Motivated by the 2-dimensional image recovery method $X = AZ + E$ (Liu, Lin, and Yu 2010), where X is the original image with noise, A is the low-rank image and E is the error, we proposed a 3-dimensional image recovery method by learning the low-rank structure of the sample space Z. Given an image set with M 3-order samples $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3 \times M}$, which is transformed by $\mathcal{X} = U_3 U_2 U_1 \mathcal{X} Z + \mathcal{E}$, where U_1, U_2 and U_3 are the projection matrices of one sample, Z reflects the low-rank structure of all the samples, and $\mathcal{E}$ is the error. The original dataset $\mathcal{X}$ can be reconstructed from the discriminant subspace by low-rank learning of Z, therefore, the pure images without noise or illumination interference can be obtained. The model is given as follows:

$$
\begin{aligned}
& \min_{U_n} \sum_{n=1}^{3} \|U_n\|_* + \|Z\|_* + \lambda \|U_n^{\mathrm{T}}(W_n - \alpha B_n) U_n\|_{\mathrm{F}}^2 \\
& \text{s.t.} \quad \|Y_{(n)} - X_{(n)} Z\|_{\mathrm{F}}^2 \leq \epsilon, Y_{(n)} = (U_3 U_2 U_1 \mathcal{X} Z)_{(n)},
\end{aligned}
\tag{17}
$$

where $Z = \left(Y' D'^{\mathrm{T}} + J' + \frac{V_1 D'^{\mathrm{T}} - V_2}{\mu}\right)\left(D D'^{\mathrm{T}} + \mathbf{I}\right)^{-1}$, Y', D' and J' are the *mode-4* variates, which reflects the sample information.

Experiment Results

In this part, we use two databases to verify our algorithm and to compare it with other state-of-the-art low-rank tensor representation learning methods used for the action classification (see Figure 3).

On the MSR hand gesture 3D database

The MSR hand gesture 3D database (Oreifej, Liu, and Redmond 2013; Wang et al. 2012) contains 12 classes of hand gestures: letter "Z", "J", "Where", "Store", "Pig", "Past", "Hungary", "Green", "Finish", "Blue", "Bathroom", and "Milk". These are performed by 10 subjects, with each subject performs 2-3 times. There are total of 333 samples, each is an action video consisting of a depth image sequence. We use the same experimental set-up as (Oreifej, Liu, and Redmond 2013) (Wang et al. 2012) in this experiment. All the subjects are independent, and each video sequence is subsampled to be the size of $80 \times 80 \times 18$. The image dimension is sufficient to represent the gesture, and the third dimension is due to the least number of the video sequence.

The optimized low-rank projection matrices of each mode U_1, U_2, U_3 are illustrated in Figure 4. The x, y, z-axis indicate the dimension of row, column, and frame, respectively. The color bar represents the value of the matrices.

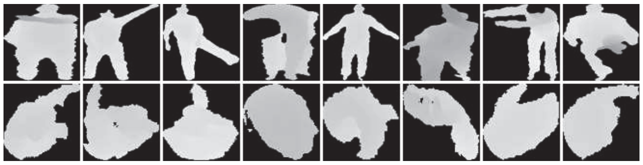


Figure 3: Key frames of different actions/gestures of (1) MSR action 3D database and (2) MSR hand gesture database on the first and second rows, respectively.

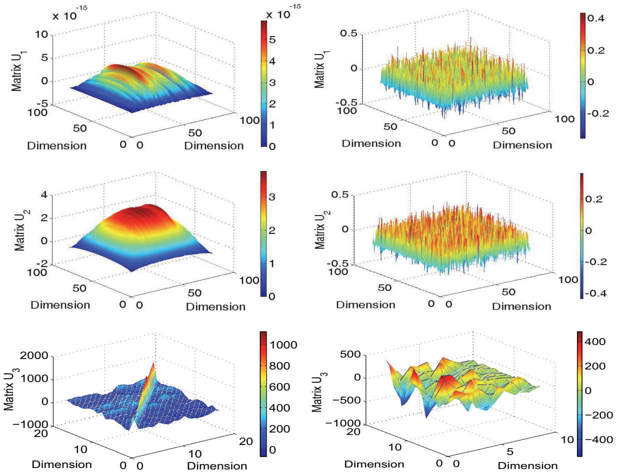


Figure 4: Illustration of the projection matrices on the MSR Hand Gesture database. The left column are our result, while the right column shows Zhong and Cheriet's. Form top to bottom: U_1, U_2, U_3.

The left column shows our matrices with regular color distribution, specifically, U_1, U_2 indicates the number of variations of the row space and column space, respectively. U_3 indicates the significant frames in the video sequence. It is similar with the full-rank matrix, that is, each frame in the sequence plays an important role in the action. In Zhong and Cheriet's method (Zhong and Cheriet 2014), the matrices do not have the obvious structure in the row, column, and frame space. The differences between Zhong's method and ours are twofold: (1) Zhong used the k-neighbors to construct local graph, while our method considers the global discriminant information, and it is sufficient for describing the whole dataset; (2) Zhong used gradient decrease method to update only one variable $W_n = U_n^{\mathrm{T}} \times U_n$ to solve their problem, while we use augmented Lagrange method (ALM) to update all the variables iteratively. In conjunction with this, the matrix U_n has two properties: (1) it is a low-rank structure; (2) contains the structure of the action videos in row, column, and frame space, respectively. The corresponding subspace obtained by the low-rank projection is shown in Figure 5. By reference of this, our method portrays the projected gestures of the action video with the details containing more energy (e.g moving fingers) compared with Zhong and Cheriet's method. This situation indicates that the projection matrix we obtained contains effective information that ensures a more reliable subspace for the classification task. Table 1 shows the accuracy of different methods. It should be evident that, the proposed method performs better than the state-of-the-art low-rank tensor representation learning methods. HON4D+D_{disc} (Oreifej, Liu, and Redmond 2013) is the latest work on the gesture database using normal orientation histogram. Zhang et al.'s work (Zhang et al. 2013) proposed to rectify align images with distortion and partial missing, which used image sequence after low-rank learning in this experiment. It had lower accuracy than our method, as it relies on the original images and can deal with the trivial changing, such as sparse noise, small fragment, and distortion; while it is not suitable for the large scale of movement, distortion or rotation in the gesture classification task. Zhong and Cheriet's method is less effective when compared with ours. Figure 6(a) shows the accuracy under different parameters β and λ. We can see the proposed method is robust across different parameter settings.

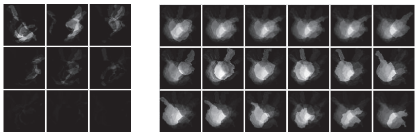


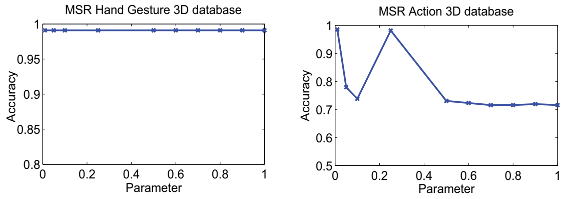
Figure 5: Left to right: Zhong and Cheriet's and our learned projected gestures of the MSR hand gesture database.

On the MSR action 3D database

The MSR action 3D database contains 20 classes of actions. This includes "arm waving", "horizontal waving", "hammer", "hand catching", "punching", "throwing", "drawing x", "drawing circle", "clapping", "two hands waving", "sideboxing", "bending", "forward kicking," "side kicking", "jogging", "tennis swing," "golf swing," "picking up and throwing". Each action is performed by 10 subjects, each performing 2-3 times. There are 567 samples in total. The action video is represented as a high-dimensional tensor in this experiment. In the following, we report two sets of results performed under different experimental settings.

Experiment setting 1 Here uses the same conditions as (Wang et al. 2012; Oreifej, Liu, and Redmond 2013). The first 5 subjects are chosen for training, while the rest are for testing. Considering the 0 value pixel as non-informative in the depth image, we first cropped the images using a bounding box to resize each image to 80 × 80. Next, we subsampled each tensor to 80 × 80 × 10. Figure 6(b) shows the accuracy with different value of parameters on the MSR action 3D database. This shows our method outperforms the state-of-the-art low-rank tensor representation learning methods. Table 2 shows the accuracy for different parameter settings of β and λ. The time for training of our method is approximately 190 seconds, while it takes 160 seconds in the recognition phase.

Experiment setting 2 Here we use the same conditions as Chen et al. (Chen, Liu, and Kehtarnavaz 2013). We split the MSR Action 3D database into 3 different sets. In the Test One (Two) set, we take the first (second) action video of each subject for training and the rest for testing. In the Cross Subject set we took the 1, 3, 5, 7, 9 subjects as training using the rest for testing. We performed three different tests on each action set. The results are shown in Figure 7. From top



(a) MSR Gesture database (b) MSR Action database

Figure 6: Accuracy of the proposed method under different parameter settings of λ and β on two used databases.

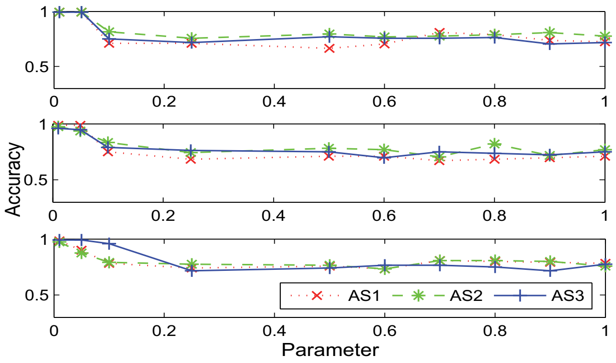


Figure 7: Accuracy with different parameters set for the MSR action 3D database. Top to bottom: Test One, Test Two, and Cross Subject Test. Each test contains three sets AS1, AS2, AS3, respectively.

to bottom is Test One, Test Two, and Cross Subject test, each with three training, and testing sets AS$_1$, AS$_2$, AS$_3$ with different parameters. The best result in each test experiment is obtained with the parameter set to 0.01 and scaled as $[0, 1]$. The results compared with the state-of-the-art methods are shown in Table 3. In the Test One and Cross Subject sets our method performs best. In the Test Two set, we have an accuracy just 2% lower than Chen et al.'s method. For Zhang et al's (Zhang et al. 2013) work, we used entire images in the database, *i.e.*, $10 * 567 = 5670$ images. Still, it was able to deal with the trivial sparse noise or distortion, such as the digit '3' in their test experiment (Zhang et al. 2013). However, the action video containing large scale movements in the arms or legs, making it not suitable for this application.

Method	Accuracy %
HON4D $+D_{disc}$	92.45
HON4D	87.29
Zhang et al.	89.93
Zhong et al.	69.44
LRTD	**99.09**

Table 1: Results for the MSR gesture database.

Method	Accuracy %
HON4D $+D_{disc}$	88.89
HON4D	85.85
Zhang et al.	95.96
Zhong et al.	92.88
LRTD	**98.50**

Table 2: Results for the MSR action database.

On CMU Multi-PIE face database

The CMU Multi-PIE face database (Gross et al. 2010) includes about 750,000 face images of 337 subjects, involving 15 various views, in 19 changes to illuminations, and 4 expressions. In this experiment, we use 67 subjects with total of 469 samples, half for training and half for testing. The discriminant information was used in this experiment. Here we selected 10 faces from one subject to show our method's performance when recovering images. The original face set $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3 \times M}$, the corresponding low-rank faces $U_3 U_2 U_1 \mathcal{X} Z$ and the errors $\mathcal{E}$ are shown in Figure 8, where M is the number of samples, $\mathcal{X} = U_3 U_2 U_1 \mathcal{X} Z + \mathcal{E}$. It shows that the illumination effect is well eliminated by the low-rank learning.

Table 3: Accuracy (%) of 3 sets on the MSR action database.

		Chen	Zhang	Zhong	Ours
Test One	AS1	97.3	46.67	92.76	**99.34**
	AS2	96.1	47.71	98.08	**99.36**
	AS3	98.7	11.33	80.26	**99.34**
	Average	97.4	35.24	90.37	**99.35**
Test Two	AS1	98.6	45.95	77.63	**98.68**
	AS2	**98.7**	47.24	91.03	97.44
	AS3	**100**	10.81	90.79	96.05
	Average	**99.1**	34.67	86.48	97.39
Cross Subject Test	AS1	96.2	44.35	91.67	**98.33**
	AS2	83.2	46.16	85.83	**97.50**
	AS3	92.0	10.81	85.83	**99.17**
	Average	90.5	33.78	87.78	**98.33**

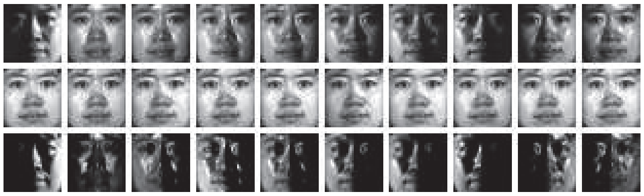


Figure 8: PIE face database. Top row: the original faces; second row: the low-rank faces; third row: the errors.

Conclusion

We proposed a low-rank tensor completion method with discriminant learning for action classification and image recovery. We employed the alternative direction method to calculate each projection matrix, by having the others fixed. In order to integrate the label information of the database, we use the discriminant analysis criterion in the low-rank tensor completion model as a regularizer. To obtain the optimized projection matrices, the augmented Lagrange method was used to solve the multi-variate optimization problem. The property of the projected matrices is explained in detail, *i.e.*, the matrices can reflect the low-rank structure in the row, column and frame space, respectively. In order to recover the high-dimensional images with noise or different illumination, we proposed an improved version that learns the low-rank structure of the sample space, and obtains good performance. Results on the MSR hand gesture 3D database and the MSR action 3D database have shown that our method performs better than the state-of-the-art low-rank tensor representation learning methods. Experiments on the Multi-PIE face database reveals the good recovery results of the faces under different illuminations.

Acknowledgements: We specially thank Xiaoqin Zhang for providing his code. Also, thanks to the anonymous reviewers for their valuable comments and suggestions.

References

Acar, E.; Dunlavy, D. M.; and Kolda, T. G. 2011. A scalable optimization approach for fitting canonical tensor decompositions. *Journal of Chemometrics* 25(2):67–86.

Arnold, A., and Jahnke, T. 2012. On the approximation of high-dimensional differential equations in the hierarchical tucker format. Technical report, Karlsruhe Institute of Technology, Department of Mathematics.

Cai, X.; Ding, C.; Nie, F.; and Huang., H. 2013. On the equivalent of low-rank linear regressions and linear discriminant analysis based regressions. *KDD* 1124–1132.

Chen, J., and Saad, Y. 2009. On the tensor svd and the optimal low rank orthogonal approximation of tensors. *SIAM Journal on Matrix Analysis and Applications* 30(4):1709–1734.

Chen, S.; Lyu, M. R.; King, I.; and Xu, Z. 2013. Exact and stable recovery of pairwise interaction tensors. In *NIPS*, 1691–1699.

Chen, C.; Liu, K.; and Kehtarnavaz, N. 2013. Real-time human action recognition based on depth motion maps. *Journal of Real-Time Image Processing* 1–9.

Ding, C.; Huang, H.; and Luo, D. 2008. Tensor reduction error analysis-applicationsl to video compression and classification. In *CVPR*, 1–8. IEEE.

Espig, M., and Hackbusch, W. 2012. A regularized newton method for the efficient approximation of tensors represented in the canonical tensor format. *Numerische Mathematik* 122(3):489–525.

Espig, M. 2007. *Effiziente Bestapproximation mittels Summen von Elementartensoren in hohen Dimensionen*. Ph.D. Dissertation, Ph. D. thesis.

Etemad, K., and Chellappa, R. 1997. Discriminant analysis for recognition of human face images. *JOSA A* 14(8):1724–1733.

Gross, R.; Matthews, I.; Cohn, J.; Kanade, T.; and Baker, S. 2010. Multi-pie. *Image and Vision Computing* 28(5):807–813.

Haegeman, J.; Cirac, J. I.; Osborne, T. J.; Pižorn, I.; Verschelde, H.; and Verstraete, F. 2011. Time-dependent variational principle for quantum lattices. *Physical Review Letters* 107(7):070601.

Holtz, S.; Rohwedder, T.; and Schneider, R. 2012. On manifolds of tensors of fixed tt-rank. *Numerische Mathematik* 120(4):701–731.

Jia, K., and Yeung, D.-Y. 2008. Human action recognition using local spatio-temporal discriminant embedding. In *CVPR*, 1–8. IEEE.

Kazeev, V., and Tyrtyshnikov, E. 2010. Structure of the hessian matrix and an economical implementation of newtons method in the problem of canonical approximation of tensors. *Computational Mathematics and Mathematical Physics* 50(6):927–945.

Khoromskij, B. N.; Oseledets, I. V.; and Schneider, R. 2012. Efficient time-stepping scheme for dynamics on tt-manifolds. Tech. Rep. 24, MPI MIS Leipzig.

Koch, O., and Lubich, C. 2010. Dynamical tensor approximation. *SIAM Journal on Matrix Analysis and Applications* 31(5):2360–2375.

Kolda, T. G., and Bader, B. W. 2009. Tensor decompositions and applications. *SIAM review* 51(3):455–500.

Krishnamurthy, A., and Singh, A. 2013. Low-rank matrix and tensor completion via adaptive sampling. In *NIPS*, 836–844.

Lin, Z.; Chen, M.; and Ma, Y. 2010. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*.

Liu, G.; Lin, Z.; and Yu, Y. 2010. Robust subspace segmentation by low-rank representation. In *ICML*, 663–670.

Lubich, C.; Rohwedder, T.; Schneider, R.; and Vandereycken, B. 2013. Dynamical approximation by hierarchical tucker and tensor-train tensors. *SIAM Journal on Matrix Analysis and Applications* 34(2):470–494.

Mu, C.; Huang, B.; Wright, J.; and Goldfarb, D. 2013. Square deal: Lower bounds and improved relaxations for tensor recovery. *arXiv preprint arXiv:1307.5870*.

Oreifej, O.; Liu, Z.; and Redmond, W. 2013. HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In *CVPR*.

Phan, Anh Huy and Tichavskỳ, Petr and Cichocki, Andrzej. 2012. On fast computation of gradients for candecomp/parafac algorithms. *arXiv preprint arXiv:1204.1586*.

Phan, Anh Huy and Tichavskỳ, Petr and Cichocki, Andrzej. 2013. Low Complexity Damped Gauss-Newton Algorithms for CANDECOMP/PARAFAC. *SIAM Journal on Matrix Analysis and Applications* 34(1):126–147.

Romera-Paredes, B., and Pontil, M. 2013. A new convex relaxation for tensor completion. *arXiv preprint arXiv:1307.4653*.

Saghafi, B., and Rajan, D. 2012. Human action recognition using pose-based discriminant embedding. *Signal Processing: Image Communication* 27(1):96–111.

Shi, X.; Ling, H.; Xing, J.; and Hu, W. 2013. Multi-target tracking by rank-1 tensor approximation. *CVPR* 2387–2394.

Uschmajew, A., and Vandereycken, B. 2013. The geometry of algorithms using hierarchical tensors. *Linear Algebra and its Applications*.

Wang, J.; Liu, Z.; Wu, Y.; and Yuan, J. 2012. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 1290–1297. IEEE.

Zhang, X.; Wang, D.; Zhou, Z.; and Ma, Y. 2013. Simultaneous rectification and alignment via robust recovery of low-rank tensors. In *NIPS*, 1637–1645.

Zheng, Z.; Zhang, H.; Jia, J.; Zhao, J.; Guo, L.; Fu, F.; and Yu, M. 2013. Low-rank matrix recovery with discriminant regularization. In *KDD*. Springer. 437–448.

Zhong, G., and Cheriet, M. 2014. Large margin low rank tensor analysis. *Neural Computation* 26(4):761–780.