Multi-view Discriminant Analysis with Tensor Representation and Its Application to Cross-view Gait Recognition

Yasushi Makihara, Al Mansur, Daigo Muramatsu, Zasim Uddin, Yasushi Yagi
The Institute of Scientific and Industrial Research, Osaka University, Osaka, Japan

Abstract— **This paper describes a method of discriminant analysis for cross-view recognition with a relatively small number of training samples. Since appearance of a recognition target (e.g., face, gait, gesture, and action) is in general drastically changes as an observation view changes, we introduce multiple view-specific projection matrices and consider to project a recognition target from a certain view by a corresponding view-specific projection matrix into a common discriminant subspace. Moreover, conventional vectorized representation of an originally higher-order tensor object (e.g., a spatio-temporal image in gait recognition) often suffers from the curse of dimensionality dilemma, we therefore encapsulate the multiple view-specific projection matrices in a framework of discriminant analysis with tensor representation, which enables us to overcome the curse of dimensionality dilemma. Experiments of cross-view gait recognition with two publicly available gait databases show the effectiveness of the proposed method in case where a training sample size is small.**

I. Introduction

Recognition from different views, namely, cross-view recognition, has been one of central topics in the computer vision and pattern recognition communities for a long time, since view changes are naturally observed in many applications (e.g., face, gait, gesture, and action recognition) and also induce drastic appearance changes of a target.

For this purpose, view-specific projections to a common subspace, are considered in many studies to cope with large appearance changes by view changes. The most popular way to obtain a common subspace for multiple views is canonical correlation analysis (CCA) [7], [6], which learns two projection matrices for a set of two variables so as to maximize a correlation between them in the common subspace. In additions, several variants of CCA have been also proposed, such as kernel CCA (KCCA) [2] and sparse CCA [5]. Whereas the above approach only consider to analyze pairwise variables, multi-view CCA (MCCA) [20] was proposed to obtain one common space for more than two views, where multiple view-specific transforms were obtained by maximizing the total correlation between any pairs of views.

Moreover, a family of regression is also regarded as one of view-specific approaches. Partial least squares (PLS) regression [24], [19] projects samples from two views to a common latent subspace, where samples from one view are regarded as regressor while those from the other view as regressand. For example, PLS is employed for face recognition with pose, low-resolution, and sketch in [21]. Support vector regression (SVR) [23] is an extension from support vector machine (SVM) to regression problem and it is employed in cross-view gait recognition [11] for example.

Although all the above methods could maximize correlation (or minimize differences) among two or more views, they do not take discrimination aspect into consideration. A straightforward solution is to employ discriminant analysis. A typical example is linear discriminant analysis (LDA) [18] which project an object with a single view-common matrix into a lower dimensional discriminant subspace in a supervised way, where a between-class variance is maximized and a within-class variance is minimized at the same time. It is, however, difficult in essence to efficiently mitigate the intra-class variance induced by view changes with a single view-common matrix.

On the other hand, discriminative approaches with multiple view-specific projections have been proposed. As extensions from CCA, correlation discriminant analysis (CDA) [15] and discriminative CCA (DCCA) [10] are proposed, where within-class correlation is maximized while between-class correlation is minimized. Moreover, as extensions from LDA, multi-view fisher discriminant analysis (MFDA) [3] for binary classification problem, and generalized multi-view analysis (GMLDA) [22] for multi-class classification from multiple views are proposed. While GMLDA requires hyper-parameter setting for regularization, multi-view discriminant analysis (MvDA) [9] provide more direct derivation from LDA for multiple view-specific projection matrices without any hyper parameters. In addition, MvDA simultaneous obtains a concatenation of multiple view-specific projection matrices by solving a single generalized eigenvalue problem in an analytical way.

On the other hand, despite that an object handled in computer vision and pattern recognition often has an higher-order tensor structure originally such as an image (a second-order tensor, namely, a matrix) and a voxel volume or spatio-temporal image (a third-order tensor), most of the above subspace learning approaches first vectorize the original tensor object into the first-order tensor, namely, a vector without keeping the original structure and thereafter project it into a lower-dimensional subspace. Such a first-order tensor of feature vector usually has a considerably high dimensionality (e.g., an image with 640 by 480 pixel size leads to a

This work was partly supported by JSPS Grants-in-Aid for Scientific Research (S) 21220003, Young Scientists (A) 23680017, "R&D Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Society", Strategic Funds for the Promotion of Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology, the Japanese Government, and the JST CREST "Behavior Understanding based on Intention-Gait Model" project.

978-1-4799-6026-2/15/$31.00 ©2015 IEEE

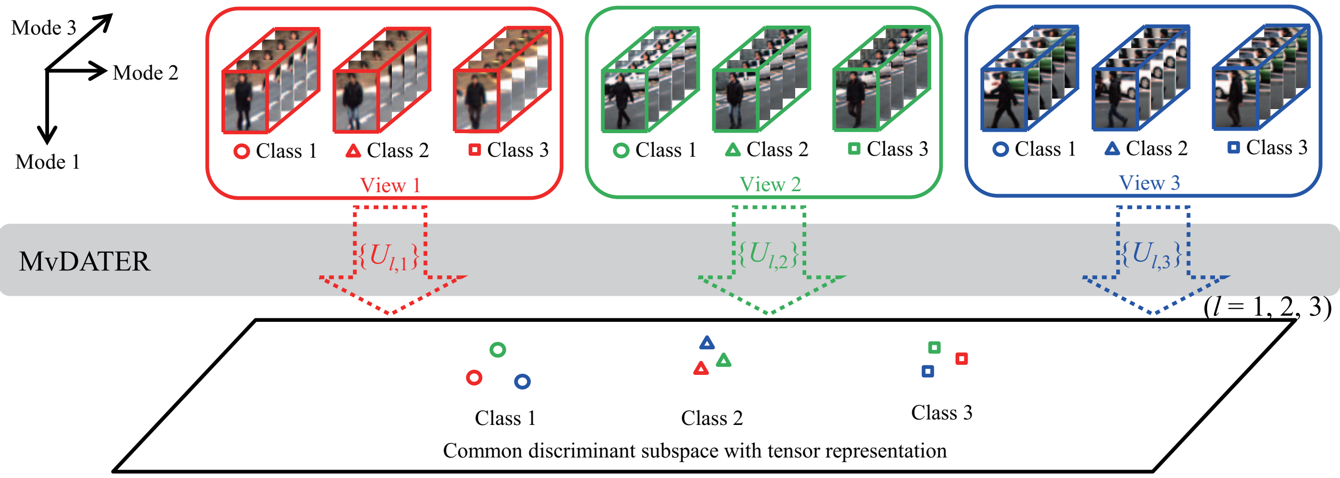


Fig. 1. Overview of MvDATER. In this example, a spatio-temporal image is treated as a third-order tensor and vertical, horizontal, and temporal axes are regarded as mode 1, 2, and 3, respectively. An objective of the MvDATER is to obtain a common discriminative tensor subspace using mode-specific and view-specific projection matrices $\{U_{l,j}\}$, where l and j are mode and view indices, respectively.

vector of 307,200 dimensions). Such a considerably higher dimensional vector often induces the curse of dimensionality dilemma or small sample size problem through subspace learning stage. More specifically, a within-class scatter matrix used in many discriminant analysis approaches, is in general singular (degenerated) in particular in case where the size of training samples is small.

In order to overcome the problem, Yan et al. [28] propose discriminant analysis with tensor representation (DATER) which treats an original tensor object as is rather than vectorizing it into a first-order tensor with high dimensionality. In DATER, multiple projection matrices are prepared for each mode, more specifically, L projection matrices for an L-order tensor object, and such mode-specific projection matrices are optimized in turn. Since the dimension considered in each optimization is at most the number of components in each mode (e.g., 640 and 480 for an image with 640 by 480 pixel size for the first and second modes, respectively) while the number of training samples is multiplied by the number of components in the other modes, DATER significantly mitigates the curse of dimensionality dilemma or small sample size problem. As an example of DATER application, Xu et al. [26] applies it to gait energy image (GEI) [4], which is a second-order tensor object, and show the effectiveness of DATER in human gait recognition problem. DATER may, however, still suffer from large within-class variations by view changes since it uses a single view-common projection matrix for each mode.

We therefore propose a method of multi-view discriminant analysis with tensor representation (MvDATER) by considering both advantages of enhanced discrimination capability by view-specific projections and tolerance to the small sample size problem by mode-specific projections. More specifically, we prepare multiple mode-specific and view-specific projection matrices (i.e., LN_V projection matrices for an L-order tensor from N_V views). We then encapsulate MvDA algorithm [9] in DATER algorithm [28] where multiple view-specific projection matrices for a certain mode is optimized through so-called l-mode discriminant analysis and where those for another mode is optimized in turn. Note that MvDATER is not a trivial combination of two existing approaches, i.e., MvDATER is beyond sequential application of MvDA and DATER, since it is a unified formulation obtained by extending the state-of-the-art MvDA into the tensor domain in a technically sound way.

Compared with previous works, the proposed method simultaneously satisfy the following properties: (1) A single common discriminative subspace is obtained for multiple views by jointly optimizing multiple view-specific projection matrices (see Fig. 1). (2) Optimization for each mode discriminant analysis is solved analytically through generalized eigenvalue problem. Since individual generalized eigenvalue problem is solved with much smaller dimensions, (3) computational cost for each mode discriminant analysis is reduced to a large extent and also (4) the curse of dimensionality dilemma is avoided. (5) The small sample size problem is overcome since the sample size is effectively multiplied by a large scale as described later.

II. Multi-view discriminant analysis with tensor representation

A. Tensor representation

Target objects in computer vision are often represented as a second or higher order tensor rather than a vector. For example, a single image and an image sequence (video) are represented as a second order tensor (matrix) and a third-order tensor, respectively. Most of the conventional approaches to subspace learning, such as principal component analysis (PCA), LDA, CCA, MvDA, firstly unfold the tensor object into a vector object (e.g., an image object $X \in \mathbb{R}^{H \times W}$ with the height H and the width W is unfolded into a vector $x \in \mathbb{R}^{HW}$ with the image size dimensionality HW), and then

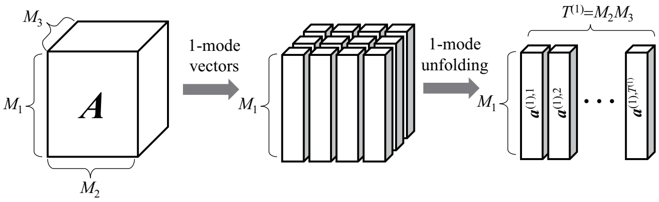


Fig. 2. Illustration of the 1-mode vectors and unfolding.

derive lower dimensional subspace from the vector object. As a result, they often suffer from the curse of dimension and/or the small sample size problem. We therefore derive a lower dimensional subspace directly from the tensor objects so as to keep the original data structure as well as avoid the curse of dimension and the small sample size problem.

More specifically, we consider an L-th order tensor object $A \in \mathbb{R}^{M_1 \times \ldots \times M_L}$, where l-mode dimensionality is represented by M_l and whose component is represented using L indices $\{m_l\}$ as $A(m_1, \ldots, m_L)$. Note that the total number of components in the tensor object A sums up to $M = \prod_{l=1}^{L} M_l$.

In the following, we further review a couple of basic techniques of tensor algebra [13].

Inner product, norm, and distance: The inner product (or scalar product) of two tensors $A, B \in \mathbb{R}^{M_1 \times \ldots \times M_L}$ with the same dimensionality is defined as

$$\langle A, B \rangle = \sum_{m_1=1}^{M_1} \ldots \sum_{m_L=1}^{M_L} A(m_1, \ldots, m_L) B(m_1, \ldots, m_L). \quad (1)$$

The Frobenius norm of a tensor A is defined as $\|A\|_F = \sqrt{\langle A, A \rangle}$, and subsequently a distance between two tensors A, B is defined as $d(A, B) = \|A - B\|_F$.

The l-mode product: The l-mode product of a tensor A by a matrix $U_l \in \mathbb{R}^{M_l \times M_l'}$, denoted as $A \times_l U_l$, is a tensor $B \in \mathbb{R}^{M_1 \times \ldots \times M_{l-1} \times M_l' \times M_{l+1} \times \ldots \times M_L}$ whose component is

$$\begin{aligned} & B(m_1, \ldots, m_{l-1}, m_l', m_{l+1}, \ldots, m_L) \\ = & \sum_{m_l=1}^{M_l} A(m_1, \ldots, m_l, \ldots, m_L) U_l(m_l, m_l'). \end{aligned} \quad (2)$$

Note that the l-mode product of tensor changes the l-mode dimensionality of the tensor from M_l to M_l' while keeping the dimensionalities of the other modes.

The l-mode vectors and unfolding: The l-mode vectors of a tensor A are defined as a set of M_l-dimensional vectors obtained from the tensor A by varying its index m_l while keeping the other indices $\{m_p\}(p \neq l)$ fixed as illustrated in Fig. 2. Since the total number of the l-mode vectors sums up to $T^{(l)} = \prod_{p \neq l} M_p$, the l-mode unfolding of the tensor A is represented a matrix $A^{(l)} = [a^{(l),1}, \ldots, a^{(l),T^{(l)}}] \in \mathbb{R}^{M_l \times T^{(l)}}$ whose column vectors $\{a^{(l),t}\}(t = 1, \ldots, T^{(l)})$ are the l-mode vectors. In this paper, we refer to the l-mode unfolding operation as $A^{(l)} \Leftarrow_l A$ and note that a bracketed superscript (l) indicates notation for the l-mode unfolding in this paper for the convenience. We also note that the norm of the l-mode product is rewritten using the l-mode unfolding by considering simple algebra as

$$\|A \times_l U_l\|_F = \left\| \sum_{t=1}^{T^{(l)}} \left(a^{(l),t}\right)^T U_l \right\|_F = \left\| \left(A^{(l)}\right)^T U_l \right\|_F. \quad (3)$$

B. Multi-view projections

Since the objective of the discriminant analysis is to find lower-dimensional discriminant subspace, we consider a multi-linear projection from an original tensor $X \in \mathbb{R}^{M_1 \times \ldots \times M_L}$ into a lower-dimensional but the same-order tensor $Y \in \mathbb{R}^{M_1' \times \ldots \times M_L'} (M_l' < M_l \; \forall l)$ as

$$Y = X \times_1 U_1 \ldots \times_L U_L, \quad (4)$$

where $U_l \in \mathbb{R}^{M_l \times M_l'}$ is a projection matrix for the l-mode product.

Although conventional approaches to discriminant analysis such as LDA consider a common projections regardless of the difference of data domains, it is in general difficult to find efficient common projections in case where tensor objects as features (e.g., face images or gait image sequences) are significantly different among the domains (e.g., different views in face or gait recognition).

We therefore introduce a multi-domain multi-mode projections to overcome such domain differences at the same time to keep tensor object structures, unlike the DATER [28] considers only the multi-mode aspect and the MvDA [9] does only the multi-domain aspect. Although we refer to the domain as *view* after this in accordance with the MvDA framework [9], note that the proposed framework is applicable to not only the view domain but also a variety of domains (e.g., illumination and expression in face recognition, walking speed and clothing in gait recognition).

Formally, we define the multi-view multi-mode projection matrices as $\mathbf{U} = \{U_{l,j} \in \mathbb{R}^{M_l \times M_l'}\}$ $(l = 1, \ldots, L, j = 1, \ldots, N_V)$, where subscripts l and j indicate mode and view, respectively, and N_V is the number of views. We then project a tensor object from any view into a common discriminant subspace by switching the projection matrix based on the domain where the tensor object X comes from accordingly, as shown in Fig. 1.

C. Discriminant tensor criterion

In this subsection, we introduce a discriminant tensor criterion with the multi-view multi-mode projection matrices. For this purpose, formally, let us define a set of training tensor objects $\mathbf{X} = \{X_{ijk} \in \mathbb{R}^{M_1 \times \ldots, \times M_L}\}(i = 1, \ldots, N_C, j = 1, \ldots, N_V, k = 1, \ldots, n_{ij})$, where X_{ijk} is the k-th training tensor object of the i-th class from the j-th view, and N_C and n_{ij} are the number of classes and the number of training samples of the i-th class from the j-th view, respectively. We subsequently define the number of training tensor objects of the i-th class as $n_i = \sum_{j=1}^{N_V} n_{ij}$ and also the total number of training tensor objects as $n = \sum_{i=1}^{N_C} n_i$.

Since the training tensor object X_{ijk} comes from the j-th view, the corresponding lower-dimensional tensor object

$Y_{ijk} \in \mathbb{R}^{M'_1 \times \ldots \times M'_L}$ in the common discriminant subspace is represented as

$$Y_{ijk} = X_{ijk} \times_1 U_{1,j} \ldots \times_L U_{L,j}. \tag{5}$$

Here, a set of multi-view multi-mode projection matrices are optimized by maximizing a between-class scatter while minimizing a within-class scatter, namely, by maximizing their ratio as

$$\mathbf{U}^* = \arg\max_{\mathbf{U}} \frac{\sum_{i=1}^{N_C} n_i \left\| \bar{Y}_i - \bar{Y} \right\|_F^2}{\sum_{i=1}^{N_C} \sum_{j=1}^{N_V} \sum_{k=1}^{n_{ij}} \left\| Y_{ijk} - \bar{Y}_i \right\|_F^2}, \tag{6}$$

where the denominator and the numerator are the within-class scatter and between-class scatter in the common discriminant subspace, respectively, and $\bar{Y}_i \in \mathbb{R}^{M'_1 \times \ldots \times M'_L}$ and $\bar{Y} \in \mathbb{R}^{M'_1 \times \ldots \times M'_L}$ are the i-th class mean and the total mean, respectively. The i-th class mean $\bar{Y}_i$ is derived as

$$\bar{Y}_i = \sum_{j=1}^{N_V} w_{ij} \left(\bar{X}_{ij} \times_1 U_{1,j} \ldots \times_L U_{L,j} \right), \tag{7}$$

where $w_{ij} = n_{ij}/n_i$ is a ratio of the number of the i-th class samples from the j-th view to the number of the i-th class samples from all the views, and $\bar{X}_{ij} \in \mathbb{R}^{M_1 \times \ldots \times M_L}$ is the i-th class mean from the j-th view in the original tensor space, which is defined as

$$\bar{X}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} X_{ijk}. \tag{8}$$

The total mean $\bar{Y}$ is similarly derived as

$$\bar{Y} = \sum_{i=1}^{N_C} w_i \bar{Y}_i = \sum_{i=1}^{N_C} w_i \sum_{j=1}^{N_V} w_{ij} \left(\bar{X}_{ij} \times_1 U_{1,j} \ldots \times_L U_{L,j} \right), \tag{9}$$

where $w_i = n_i/n$ is a ratio of the number of the i-th class samples to the total number of samples.

Now, by substituting Eqs. (5)(7)(9) into Eq. (6), we obtain

$$\mathbf{U}^* = \arg\max_{\mathbf{U}} \frac{\sum_{i=1}^{N_C} n_i \left\| \sum_{r=1}^{N_V} w_{ir} (\bar{X}_{ir} \times_1 U_{1,r} \ldots \times_L U_{L,r}) - \sum_{q=1}^{N_C} w_q \sum_{r=1}^{N_V} w_{qr} (\bar{X}_{qr} \times_1 U_{1,r} \ldots \times_L U_{L,r}) \right\|_F^2}{\sum_{i=1}^{N_C} \sum_{j=1}^{N_V} \sum_{k=1}^{n_{ij}} \left\| X_{ijk} \times_1 U_{1,j} \ldots \times_L U_{L,j} - \sum_{r=1}^{N_V} w_{ir} (\bar{X}_{ir} \times_1 U_{1,r} \ldots \times_L U_{L,r}) \right\|_F^2}. \tag{10}$$

Since there is in general no closed-form solution for Eq. (10) due to the higher-order tensor structure, we alternatively search for an iterative solution to derive the common discriminant subspace as described in subsection II-E.

D. *l-mode discriminant analysis*

In this subsection, we introduce an l-mode discriminant analysis, which is an essential technique for the iterative solution described in subsection II-E. More specifically, we consider another discriminant criterion focused only on the l-mode product with projection matrices $\mathbf{U}_l = \{U_{l,j} \in \mathbb{R}^{M_l \times M'_l}\} (j = 1, \ldots, N_V)$ as

$$\mathbf{U}_l^* = \arg\max_{\mathbf{U}_l} \frac{\sum_{i=1}^{N_C} n_i \left\| \sum_{r=1}^{N_V} w_{ir} (\bar{X}_{ir} \times_l U_{l,r}) - \sum_{q=1}^{N_C} w_q \sum_{r=1}^{N_V} w_{qr} (\bar{X}_{qr} \times_l U_{l,r}) \right\|_F^2}{\sum_{i=1}^{N_C} \sum_{j=1}^{N_V} \sum_{k=1}^{n_{ij}} \left\| X_{ijk} \times_l U_{l,j} - \sum_{r=1}^{N_V} w_{ir} (\bar{X}_{ir} \times_l U_{l,r}) \right\|_F^2}. \tag{11}$$

Recall that the norm of the l-mode product is represented by the l-mode unfolding as Eq. (3), we reformulate Eq. (11) as (please refer to supplementary material for the detailed derivation)

$$\mathbf{U}_l^* = \arg\max_{\mathbf{U}_l} \frac{Tr\left(\sum_{r=1}^{N_V} \sum_{s=1}^{N_V} U_{l,r}{}^T S_{B,rs}^{(l)} U_{l,s} \right)}{Tr\left(\sum_{r=1}^{N_V} \sum_{s=1}^{N_V} U_{l,r}{}^T S_{W,rs}^{(l)} U_{l,s} \right)}, \tag{12}$$

where $Tr(\cdot)$ means a trace of a matrix, and $S_{W,rs}^{(l)} \in \mathbb{R}^{M_l \times M_l}$ and $S_{W,rs}^{(l)} \in \mathbb{R}^{M_l \times M_l}$ are within-class and between-class scatter matrices, respectively, for l-mode unfolding from a pair of the r-th view and s-th view which are defined as

$$S_{W,rs}^{(l)} = \sum_{i=1}^{N_C} \left(\sum_{k=1}^{n_{ir}} \delta_{rs} X_{irk}^{(l)} \left(X_{isk}^{(l)} \right)^T - \frac{n_{is} n_{ir}}{n_i} \bar{X}_{ir}^{(l)} \left(\bar{X}_{is}^{(l)} \right)^T \right) \tag{13}$$

$$S_{B,rs}^{(l)} = \sum_{i=1}^{N_C} \frac{n_{ir} n_{is}}{n_i} \bar{X}_{ir}^{(l)} \left(\bar{X}_{is}^{(l)} \right)^T - \frac{1}{n} \left(\sum_{i=1}^{N_C} n_{ir} \bar{X}_{ir}^{(l)} \right) \left(\sum_{i=1}^{N_C} n_{is} \bar{X}_{is}^{(l)} \right)^T, \tag{14}$$

where δ_{rs} is Kronecker's delta, and $X_{ijk}^{(l)} \in \mathbb{R}^{M_l \times T^{(l)}}$ and $\bar{X}_{ij}^{(l)} \in \mathbb{R}^{M_l \times T^{(l)}}$ are the l-mode unfolding of X_{ijk} and $\bar{X}_{ij}$, i.e., $X_{ijk}^{(l)} \Leftarrow_l X_{ijk}$ and $\bar{X}_{ij}^{(l)} \Leftarrow_l \bar{X}_{ij}$, respectively.

Moreover, we can rewrite Eq. (12) by introducing view concatenated version of matrices as

$$U_l{}^* = \arg\max_{U_l} \frac{Tr\left(U_l{}^T S_B^{(l)} U_l \right)}{Tr\left(U_l{}^T S_W^{(l)} U_l \right)}, \tag{15}$$

where $U_l \in \mathbb{R}^{N_V M_l \times M'_l}$ is the l-mode view-concatenated projection matrix, and $S_W^{(l)} \in \mathbb{R}^{N_V M_l \times N_V M_l}$ and $S_B^{(l)} \in \mathbb{R}^{N_V M_l \times N_V M_l}$ are view-concatenated within-class and between-class scatter matrices for the l-mode unfolding, which are respectively defined as

$$U_l = \begin{bmatrix} U_{l,1} \\ \vdots \\ U_{l,N_V} \end{bmatrix}, \; S_W^{(l)} = \begin{bmatrix} S_{W,11}^{(l)} & \cdots & S_{W,1N_V}^{(l)} \\ \vdots & \ddots & \vdots \\ S_{W,N_V 1}^{(l)} & \cdots & S_{W,N_V N_V}^{(l)} \end{bmatrix},$$

$$S_B^{(l)} = \begin{bmatrix} S_{B,11}^{(l)} & \cdots & S_{B,1N_V}^{(l)} \\ \vdots & \ddots & \vdots \\ S_{B,N_V 1}^{(l)} & \cdots & S_{B,N_V N_V}^{(l)} \end{bmatrix}. \tag{16}$$

Since the closed form solution for the objective function in Eq. (17), which is in the form of trace ratio, does not exist according to [25], we relax the objective function into a more tractable one in the form of ratio trace as bellow:

$$U_l{}^* = \arg\max_{U_l} Tr\left(\frac{U_l{}^T S_B^{(l)} U_l}{U_l{}^T S_W^{(l)} U_l}\right), \quad (17)$$

which can be solved analytically through generalized eigenvalue problem:

$$S_B^{(l)} U_l = S_W^{(l)} U_l \Lambda, \quad (18)$$

where $\Lambda \in \mathbb{R}^{M_l \times M_l}$ is an orthogonal matrix whose diagonal components are eigenvalues. We then extract the first M_l' largest eigenvectors as a solution $U_l{}^*$. For more detailed discussion on the number of available projection directions, we refer the reader to the literature [28] due to page limitation.

E. Iterative solution

As described before, since the discriminant tensor criterion defined by Eq. (10) often has no closed-form solution, we introduce an iterative solution to this. More specifically, we firstly initialize the projection matrix $U_{j,l}$ as an identity matrix and then optimize the l-mode projection matrix U_l while keeping the projection matrices $\{U_p\}(p \neq l)$ for the other modes fixed and repeat this process by changing the mode for optimization target until satisfying a convergence condition or reaching the maximum iterations. To this end, we introduce a tensor object ${}^lY_{ijk} \in \mathbb{R}^{M_l' \times \ldots \times M_{l-1}' \times M_l \times M_{l+1}' \times \ldots \times M_L'}$ which is dimension reduced from an original training tensor object X_{ijk} except for the l-mode as

$${}^lY_{ijk} = X_{ijk} \times_1 U_{1,j} \ldots \times_{l-1} U_{l-1,j} \times_{l+1} U_{l+1,j} \ldots \times_L U_{L,j}. \quad (19)$$

Similarly, a tensor object ${}^l\bar{Y}_{ij} \in \mathbb{R}^{M_l' \times \ldots \times M_{l-1}' \times M_l \times M_{l+1}' \times \ldots \times M_L'}$ for the mean tensor object $\bar{X}_{ij}$ is introduced. Now, the l-mode discriminant analysis is applied for the tensor objects with reduced dimensions except for the l-mode as

$$\mathbf{U}_l^* = \arg\max_{\mathbf{U}_l} \frac{\sum_{i=1}^{N_C} n_i \left\| \sum_{r=1}^{N_V} w_{ir}\left({}^l\bar{Y}_{ir} \times_l U_{l,r}\right) - \sum_{q=1}^{N_C} w_q \sum_{r=1}^{N_V} w_{qr}\left({}^l\bar{Y}_{qr} \times_l U_{l,r}\right) \right\|_F^2}{\sum_{i=1}^{N_C} \sum_{j=1}^{N_V} \sum_{k=1}^{n_{ij}} \left\| {}^lY_{ijk} \times_l U_{l,j} - \sum_{r=1}^{N_V} w_{ir}\left({}^l\bar{Y}_{ir} \times_l U_{l,r}\right) \right\|_F^2}. \quad (20)$$

Since the only difference between this equation and Eq. (11) is replacement of the original tensor object X by the dimension-reduced tensor ${}^lY$, we can similarly solve this equation by replacing the original tensors by the dimension-reduced tensors in the following equations after Eq. (11).

An algorithm summary of the proposed method is shown in Algorithm 1.

F. Computational complexity

In this subsection, we discuss the properties of the proposed method in terms of computational complexity compared with closely related approaches such as LDA [18], MvDA [9], DATER [28] to the proposed method, MvDATER. For simplicity of the analysis, we consider a situation where each mode has the same dimensionality (i.e., $M_l = M \, \forall l$), and where the number of the training samples per class per view is the same (i.e., $n_{ij} = n_T \, \forall i, j$). In addition, since the generalized eigenvalue problem is the most important part w.r.t. the computational complexity, we focus on the generalized eigenvalue problem.

LDA: The L-order tensor is vectorized when computing the scatter matrices, and the dimensionality of the scatter matrix is then $\prod_{l=1}^{L} M = M^L$. Since the dimensionality M^L is considerably high, it is often the case that the number of training samples is much less than the dimensionality ($n \ll M^L$), which results in singularity of the within-class scatter matrix (small sample size problem, curse of the dimensionality). Moreover, the computational complexity for the generalized eigenvalue problem is cubic order of the dimensionality, LDA for the higher-order tensor objects requires high computational cost $O\left((M^L)^3\right)$.

MvDA: Since MvDA constructs a view-concatenated scatter matrix composed of $N_V \times N_V$ sub-matrices (see Eq. (16) for reference), the dimensionality is N_V-time larger than that for LDA (i.e., $N_V M^L$). In addition, the number of training samples needs to be considered at the submatrix level in essence, and it is hence N_V-time smaller than that for LDA (i.e., n/N_V). Although MvDA has a good discrimination capability for multiple views, it more suffers from the singularity problem than LDA and also higher computational cost $O\left((N_V M^L)^3\right)$.

DATER: Since DATER constructs a scatter matrix through unfolding operation for each mode, the mode-wise dimensionality is just M, which is much smaller than that of LDA M^L. In addition, since the unfolding operation also drastically increase the number of training samples as $N \prod_{p \neq l}^{L} M = NM^{L-1}$. Therefore, DATER mitigates the small sample size problem to large extent because the condition, $NM^{L-1} > M$, holds in most case. Moreover, the computational cost for each mode and loop is $O(M^3)$ and hence that for the whole process is at most $O\left(N_{iter} L M^3\right)$.

MvDATER: In analogous to relation between LDA and MvDA, the dimensionality is N_V-time larger than that for DATER (i.e., $N_V M$), while the number of training samples is N_V-time smaller than that for DATER (i.e., nM^{L-1}/N_V). As a result, the computational cost is $(N_V)^3$-time larger than that for DATER (i.e., $O\left(N_{iter} L (N_V M)^3\right)$. The number of views N_V is, however, much smaller than the date dimension M in general (e.g., $N_V = 2$ when handling pairwise view), we can say that the proposed MvDATER realize a reasonable tradeoff among discrimination capability, small sample size problem, and the computational cost.

III. Application to cross-view gait recognition

A. Setup

We evaluated the proposed MvDATER approach under cross-view gait recognition (i.e. gait-based person authentication) using the most prevailing gait feature, i.e., GEI [4], with two publicly available gait databases: (1) CASIA Gait Database B (call it CASIA later) [29] and (2) the OU-ISIR Gait Database, the Large Population Data set (call it OU-LP later) [8]. CASIA contains walking sequences from

Algorithm 1 MvDATER

Input: A set of L-order training tensor objects $\mathbf{X} = \{X_{ijk} \in \mathbb{R}^{M_1 \times \ldots \times M_L}\}$ $(i = 1, \ldots, N_C, j = 1, \ldots, N_V, k = 1, \ldots, n_{ij})$, a set of dimensions in the common discriminant subspace $\{M_l'\}$, convergence criteria ε, and the maximum iteration N_{iter}

Output: A set of projection matrices $\mathbf{U} = \{U_{l,j} \in \mathbb{R}^{M_l \times M_l'}\}$ $(l = 1, \ldots, L, j = 1, \ldots, N_V)$

$U^{prev}_{l,j}, U^{cur}_{l,j} \leftarrow I_{M_l}\ \forall l, j$ ▷ Initialization

for $iter = 1$ to N_{iter} **do**

  for $l = 1$ to L **do**

    ${}^lY_{ijk} \leftarrow X_{ijk} \times_1 U^{cur}_{1,j} \ldots \times_{l-1} U^{cur}_{l-1,j} \times_{l+1} U^{prev}_{l+1,j} \ldots \times_L U^{prev}_{L,j}\ \forall i,j,k$

    ${}^l\bar{Y}_{ij} \leftarrow \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} {}^lY_{ijk}\ \forall i,j$

    ${}^lY^{(l)}_{ijk} \Leftarrow_l {}^lY_{ijk}\ \forall i,j,k,\ {}^l\bar{Y}^{(l)}_{ij} \Leftarrow_l {}^l\bar{Y}_{ij}\ \forall i,j$ ▷ The l-mode unfolding

    $S^{(l)}_{W,rs} \leftarrow \sum_{i=1}^{N_C} \left\{ \sum_{k=1}^{n_{ir}} \delta_{rs} {}^lY^{(l)}_{irk} \left({}^lY^{(l)}_{isk}\right)^T - \frac{n_{is}n_{ir}}{n_i} {}^l\bar{Y}^{(l)}_{ir} \left({}^l\bar{Y}^{(l)}_{is}\right)^T \right\}\ \forall r,s$

    $S^{(l)}_{B,rs} \leftarrow \sum_{i=1}^{N_C} \frac{n_{ir}n_{is}}{n_i} {}^l\bar{Y}^{(l)}_{ir} \left({}^l\bar{Y}^{(l)}_{is}\right)^T - \frac{1}{n} \left(\sum_{i=1}^{N_C} n_{ir} {}^l\bar{Y}^{(l)}_{ir}\right) \left(\sum_{i=1}^{N_C} n_{is} {}^l\bar{Y}^{(l)}_{is}\right)^T\ \forall r,s$

    ▷ Within-class and between-class scatter matrices

    Update U^{cur}_l by solving $S^{(l)}_B U_l = S^{(l)}_W U_l \Lambda$ ▷ Generalized eigenvalue problem

  end for

  if $\|U^{cur}_l - U^{prev}_l\|_F < M_l M_l' \varepsilon\ \forall l$ **then** ▷ Convergence condition

    break

  end if

  $U^{prev}_l \leftarrow U^{cur}_l\ \forall l$ ▷ Update

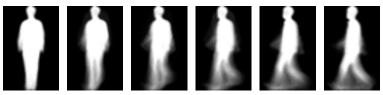
end for

Output $\{U^{cur}_l\}$

TABLE I

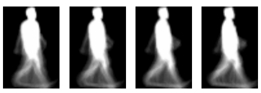
DIMENSIONALITY OF THE SCATTER MATRICES, THE NUMBER OF TRAINING SAMPLES, AND THE COMPUTATIONAL COMPLEXITY W.R.T. THE GENERALIZED EIGENVALUE PROBLEM

Approaches	LDA [18]	MvDA [9]	DATER [28]	MvDATER
Dimensionality	M^L	$N_V M^L$	M	$N_V M$
#Training samples	n	n/N_V	nM^{L-1}	nM^{L-1}/N_V
Complexity	$O\left((M^L)^3\right)$	$O\left((N_V M^L)^3\right)$	$O\left(N_{iter} L M^3\right)$	$O\left(N_{iter} L (N_V M)^3\right)$



0° 18° 36° 54° 72° 90°

Fig. 3. GEIs from CASIA.



55° 65° 75° 85°

Fig. 4. GEIs from OU-LP.

124 subjects captured from wide range of views (i.e., 18° intervals from 0° (frontal view) to 180° (rear view), and hence it is suitable to evaluate gait recognition under large view variations. We divided the subjects into disjoints sets of 62 test subjects and 62 training subjects. For the test subjects, we set the first normal walking sequences from a view 90° as a gallery, while we set the second to the sixth (five in total) normal walking sequences from views 72°, 54°, 36°, 18°, 0° as probes, respectively. For the training sets, we used normal walking sequences for pairwise views (e.g., 90° and 72°). Examples of the GEIs (40 by 60 pixel-size) at six different views can be seen in Fig. 3.

OU-LP includes walking image sequences from 1,912 subjects drawn from wide age generation of males and females at view angles 55°, 65°, 75°, 85°, and hence it is suitable for statistically reliable evaluation. The whole set was divided into disjoint sets of 956 test subjects and 956 training subjects. For the test subjects, we set the first sequence from a view 85° as a gallery, while we set the second sequence from views 55°, 65°, 75° as probes, respectively. Similar to CASIA, training was done for pairwise view. Examples of the GEIs (44 by 64 pixel-size) at four different views can be seen in Fig. 4.

As for performance measures, we picked up rank-1 identification rates (denoted as Rank-1 later) a.k.a. correct classification rate (CCR) in identification scenarios (i.e. one-to-many matching) as well as equal error rate (EER) of false acceptance rate of imposters (different persons) and false rejection rate of genuine (the same person) in verification scenarios (i.e. one-to-one matching). We compared the proposed MvDATER with three closely related approaches as benchmarks: LDA [18], MvDA [9], and DATER [28], and dissimilarity measures are computed by Euclidean distance in each discriminant space. Note that each benchmark is followed by the preprocessing dimension reduction approaches, more specifically, PCA for LDA and MvDA, and concurrent subspace analysis (CSA) [27] for DATER and MvDATER.

TABLE II

RESULTS FOR CASIA (#TRAINING SUBJECTS: 62). BOLD AND UNDERLINE MEAN THE BEST AND THE SECOND BEST PERFORMANCE, RESPECTIVELY. THE PROPOSED MVDATER ACHIEVED THE BEST OR THE SECOND BEST ACCURACIES IN MANY CASES.

Probe view	Rank-1					EER				
	72°	54°	36°	18°	0°	72°	54°	36°	18°	0°
LDA	**80.3%**	29.4%	6.5%	6.1%	**3.2%**	**13.5%**	27.3%	36.6%	43.2%	**42.9%**
MvDA	4.2%	4.2%	1.9%	1.9%	0.0%	43.6%	40.9%	45.8%	44.2%	45.5%
DATER	56.5%	9.0%	6.8%	1.6%	2.6%	23.2%	37.5%	44.9%	48.7%	49.6%
MvDATER	66.5%	**48.1%**	**16.8%**	**7.4%**	1.6%	14.2%	**20.2%**	**34.3%**	**41.2%**	44.4%

B. Results

CASIA: To show the robustness of the proposed method against small sample size problem, we picked up only one normal sequence per subject from the training set and trained the proposed MvDATER as well as the other benchmarks. As shown in Table II, we can see that the proposed MvDATER achieved the best or the second best for almost all the settings. Because within-class scatter matrices for LDA and MvDA suffers from singularity (in particular for MvDA) due to small sample size problem, trained projection matrices for LDA and MvDA did not perform well in low-dimensional discriminant subspaces. While the DATER overcome such a troublesome small sample size problem, it still suffers from insufficient discrimination capability because it only has a single view-common projection for each mode, which results in poor performance for large view variations. On the other hand, the proposed MvDATER has multiple view-specific projection matrices for each mode and at the same time avoids the small sample size problem, and it therefore outperforms the other benchmarks as a result. As an exception, LDA outperforms MvDATER for 72° probe. This is because gait features from 90° view and 72° view are relatively similar each other and hence even a single projection matrix can successfully absorb the intra-class variations among them.

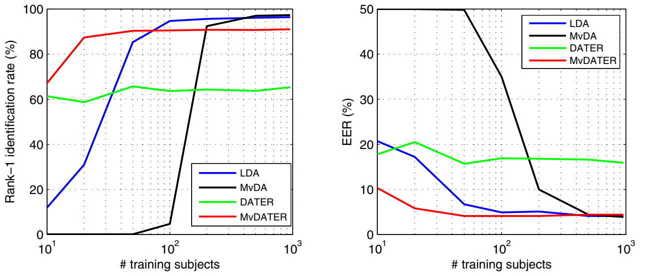
OU-LP: Whereas the samples per subject was limited in the previous experiment, we limited the number of training subjects to 10 in this experiment to check the robustness against small sample size problem. As shown in Table III, we can see that the proposed MvDATER performs well on average, although it does not work well for view 55°. In addition, DATER seems to be comparable to MvDATER. This is because the OU-LP contains much larger variation in test subjects than that of CASIA and hence DATER, which is the most robust to small training sample sizes, performs relatively well.

In order to further investigate the effect of the number of training subjects, we show the performance transition against the number of training subjects for view 75° from OU-LP in Fig. 5. From this graph, we observe the followings. (1) LDA and MvDA perform well for sufficient number of training subjects (e.g., more than 100 subjects), and their performance drastically drop as the number of training subjects decreases. In particular, MvDA, that is the most recent benchmark, performs quite poorly when training sample sizes are small. This reveals the limitation of MvDA and prompt us to more focus on the generalization capability aspect in future avenue of research. (2) DATER keeps its performance against the decrease of the number of training subjects, although its basic performance is lower than the other benchmarks. From another perspective, we can say that it does not increase the performance as the training sample sizes increase. (3) MvDATER exhibit higher performance than DATER thanks to multiple view-specific projections and keeps relatively good performance against the decrease of the number of training subjects. As a result, we can confirm the strength of the proposed method in case where the number of training samples is small.

TABLE III

RESULTS FOR OU-LP (#TRAINING SUBJECTS: 10). BOLD AND UNDERLINE MEAN THE BEST AND THE SECOND BEST PERFORMANCE, RESPECTIVELY. THE PROPOSED MVDATER IS CMOPARABLE TO DATER DUE TO EXTREMELY SMALL NUMBER OF TRAINING SUBJECTS.

Probe view	Rank-1			EER		
	75°	65°	55°	75°	65°	55°
LDA	12.0%	5.1%	2.8%	20.7%	28.0%	**30.6%**
MvDA	0.1%	0.1%	0.1%	50.0%	50.0%	50.0%
DATER	61.3%	**25.7%**	**9.1%**	17.8%	24.5%	31.4%
MvDATER	**67.2%**	12.6%	1.9%	**10.3%**	**22.0%**	37.5%



(a) Rank-1 (b) EER

Fig. 5. Performance transition against the number of training subjects for OU-LP, view 75°. The proposed MvDATER suppresses accuracy degradation against decrease of the number of training subjects compared with the other benchmarks.

IV. DISCUSSION

A. Connections to 2DLDA and its variant

As also discussed in the literature [28], while 2DLDA [12] only considers a 2-mode discriminant analysis, DATER considers discriminant analysis for all the modes in turn. More specifically, the 2DLDA is formulated as a special case of MvDATER with $N_V = 1$, $L = 2$, and $U_{1,1} = I$. In addition, a straightforward multi-view extension of the 2DLDA, that is, 2DMvDA, could be considered, and it is again formulated as a special case of MvDATER with $N_V = 2$, $L = 2$, and $U_{1,j} = I$ ($j = 1,2$). The proposed MvDATER is therefore regarded as a unified framework for these discriminant analyses.

B. Class masking problem

Since the proposed MvDATER is built upon the LDA which optimizes the Bayes error for the case of unimodal Gaussian classes with equal covariances, it might increase the overlap between the class conditional densities in the lower dimensional subspace in a heteroscedastics setting, which is so-called class masking problem. To cope with the class masking problem, Moustafa et al. [1] employed pareto discriminant analysis which simultaneously maximizes each class-pairwise distance and which thus encourages the case that all classes are equidistant from each other in the lower dimensional space. Since the pareto discriminant analysis can be encapsulate in each l-mode discriminant analysis of the proposed MvDATER, we will extend the MvDATER so as to mitigate the class masking problem in future.

V. CONCLUSION

This paper described a method of multi-view discriminant analysis with tensor representation (MvDATER) for cross-view recognition with a relatively small number of training samples. We introduce multiple view-specific and mode-specific projection matrices so as that high-order tensor objects from multiple views can be projected into a single common discriminant subspace. In the proposed algorithm, multiple view-specific projection matrices are jointly and analytically optimized via a single generalized eigenvalue problem with smaller dimension for each mode, which draws many of the advantages such as efficient cross-view handling and overcoming the curse of dimensionality dilemma and small sample size problem.

While we validated the effectiveness of the proposed method with cross-view gait recognition problems compared with the most relevant benchmarks, we will further compare it with more advanced approaches to cross-view gait recognition (e.g., [17], [16], [11], [14]). Moreover, we will further validate it with a variety of cross-view recognition such as action recognition and face recognition in future.

REFERENCES

[1] K. T. Abou-Moustafa, F. D. la Torre, and F. P. Ferrie. Pareto discriminant analysis. In *Proc. of IEEE computer society conference on Computer Vision and Pattern Recognition 2010*, pages 1–8, San Francisco, CA, USA, Jun. 2010.

[2] S. Akaho. A kernel method for canonical correlation analysis. *CoRR*, abs/cs/0609071, 2006.

[3] T. Diethe, D. Hardoon, and J. Shawe-Taylor. Constructing nonlinear discriminants from multiple data views. In *Machine Learning and Knowledge Discovery in Databases*, volume 6321 of *Lecture Notes in Computer Science*, pages 328–343. Springer Berlin Heidelberg, 2010.

[4] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316– 322, 2006.

[5] D. Hardoon and J. Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83(3):331–353, 2011.

[6] D. R. Hardoon, S. Szedmak, O. Szedmak, and J. Shawe-taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[7] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

[8] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi. The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE Transactions on Information Forensics and Security*, 7(5):1511–1521, Oct. 2012.

[9] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. In *Proc. of the 12th European Conf. on Computer Vision*, pages 808–821, Oct. 2012.

[10] T.-K. Kim, J. Kittler, and R. Cipolla. Learning discriminative canonical correlations for object recognition with image sets. In *Proc. of the 9th European Conference on Computer Vision*, pages 251–262, 2006.

[11] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li. Gait recognition under various viewing angles based on correlated motion regression. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(6):966–980, 2012.

[12] K. Liu, Y. Cheng, and J. Yang. Algebraic feature extraction. *IEEE Trans. Circuits Syst. Video Technol*, 26(6):903–911, 2006.

[13] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. A survey of multilinear subspace learning for tensor data. *Pattern Recognition*, 44(7):1540–1551, July 2011.

[14] J. Lu and Y.-P. Tan. Uncorrelated discriminant simplex analysis for view-invariant gait signal computing. *Pattern Recognition Letters*, 31(5):382–393, 2010.

[15] Y. Ma, S. Lao, E. Takikawa, and M. Kawade. Discriminant analysis in correlation similarity measure space. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 577–584, New York, NY, USA, 2007. ACM.

[16] R. Martín-Félez and T. Xiang. Gait recognition by ranking. In *Proceedings of the 12th European conference on Computer Vision - Volume Part I*, ECCV'12, pages 328–341, Berlin, Heidelberg, 2012. Springer-Verlag.

[17] D. Muramatsu, Y. Makihara, and Y. Yagi. Quality-dependent view transformation model for cross-view gait recognition. In *Proc. of the IEEE 6th International Conference on Biometrics: Theory, Applications and Systems*, number Paper ID: 12, pages 1–8, Washington D.C., USA, Sep. 2013.

[18] N. Otsu. Optimal linear and nonlinear solutions for least-square discriminant feature extraction. In *Proc. of the 6th Int. Conf. on Pattern Recognition*, pages 557–560, 1982.

[19] R. Rosipal and N. Kramer. Overview and recent advances in partial least squares. In *SLSFS*, volume 3940 of *Lecture Notes in Computer Science*, pages 34–51. Springer, 2005.

[20] J. Rupnik and J. Shawe-Taylor. Multi-view canonical correlation analysis. In *Proc. of Conference on Data Mining and Data Warehouses 2010*, pages 1–4, 2010.

[21] A. Sharma and D. W. Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *Proc of the 24th IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 2011.

[22] A. Sharma, A. Kumar, H. D. III, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *Proc of the 25th IEEE Conference on Computer Vision and Pattern Recognition*, pages 2160–2167, 2012.

[23] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, Aug. 2004.

[24] V. Vinzi, W. Chin, J. Henseler, and H. Wang, editors. *Handbook of Partial Least Squares*. Springer Handbooks of Computational Statistics. Springer, 2010.

[25] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang. Trace ratio vs. ratio trace for dimensionality reduction. In *Proc. of the 20th IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

[26] D. Xu, S. Yan, D. Tao, L. Zhang, X. Li, and H. jiang Zhang. Human gait recognition with matrix representation. *IEEE Trans. Circuits Syst. Video Technol*, 16(7):896–903, 2006.

[27] D. Xu, S. Yan, L. Zhang, H.-J. Z. andZhengkai Liu, and H.-Y. Shum. Concurrent subspaces analysis. In *Proc. of the IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, pages 203–208, Jun. 2005.

[28] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H.-J. Zhang. Discriminant analysis with tensor representation. In *Proc. of the IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, pages 526–532, Jun. 2005.

[29] S. Yu, D. Tan, and T. Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Proc. of the 18th Int. Conf. on Pattern Recognition*, volume 4, pages 441–444, Hong Kong, China, Aug. 2006.